CONTEXT-BASED HEALTH INFORMATION RETRIEVAL

Carla Teixeira Lopes

May 2013

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Informatics Engineering

to the

Departmento de Engenharia Informática
Faculdade de Engenharia
Universidade do Porto

This thesis was typeset using an adapted version of the LaTeX template made by Eivind Uggedal available at:
https://bitbucket.org/uggedal/thesis.

This work was financially supported by a scholarship
from the Fundação para a Ciência e a Tecnologia (FCT)
under grant SFRH/BD/40982/2007.

To my lovely children, Miguel and Alice.

ABSTRACT

The Web has profoundly changed the way we access information. The increase in the amount of information available and the easy access to the Web have contributed to the popularity of search in this medium. While this is true in several domains, it is particularly significant in the health domain. In fact, search for health information is the third most popular online activity and it is performed by almost 3 out of 4 American Internet users. Besides its popularity, health web search is particularly important to inform health consumers, encouraging them to become more participatory in their health management. In the health domain, the context surrounding the search is extremely rich and we believe it can contribute to improve system's performance. Although information retrieval has made significant progresses supporting its retrieval methods solely on the query and document collection, it has been recognized that context should be explored. Context can, for example, be used to disambiguate a query, to retrieve documents adjusted to the expertise of the searcher or to adjust documents to the patient's medical record.

In this dissertation we investigate the effects of context features in consumer health information retrieval. In addition, we propose and evaluate strategies to use context features in query formulation support. The popularity of web search on the health domain made us focus on Web retrieval. We opt to concentrate on health consumers due to the lack of research focusing on this specific public. The thesis behind our work is that consumer health information retrieval is affected by context features, which can be used in query formulation support to improve retrieval performance.

In three exploratory studies, we analyze how a large number of features affect different outcomes of the retrieval process. Based on findings from these studies and on the importance of query formulation support in a domain where terminology can be a barrier and translations to other languages are not always obvious, we explored the impact of query translations in users with different characteristics. Based on the assumption that a query using a language that is popular on the Web may easily reach high-quality contents, one study analyzes the effects of translating a query to the English language in users with different levels of English proficiency. Findings show that users having, at least, elementary English proficiency benefit from English query suggestions. The other studies focus on query terminology translation, that is, translation between lay and medico-scientific terminology, considering users' health literacy and topic familiarity. Although several strategies have been previously proposed to overcome the terminology gap between health consumers and web documents, none considers users' health literacy and topic familiarity. Findings suggest that users with inadequate health literacy and users who are unfamiliar with the topic should be provided with recommendations of lay queries. On

the other hand, users with higher health literacy or higher topic familiarity should be given alternative queries using medico-scientific terminology.

Based on the above findings, we developed a query suggestion system that, using domain information gathered from an existing consumer health vocabulary, identifies the medical concepts included in the query and returns four types of suggestions combining the Portuguese or English languages with the lay or medico-scientific terminologies. We found that suggestions offered by the system had a good acceptance, with English suggestions being preferred to Portuguese ones in *basic* and *proficient* users and medico-scientific suggestions being preferred to lay ones in higher levels of health literacy. We concluded that a retrieval system including the implemented suggestion strategy without any kind of personalization tends to be better than a system without suggestions with respect to precision, correctness of the resulting knowledge and also of its incorrectness. We also concluded that this system tends to be slightly worse in terms of motivational relevance. Of these, only the incorrectness difference is significant. This is extremely relevant in the health domain where incorrect information can have serious consequences. Moreover, we also found that the personalization of this system to users' English proficiency and health literacy, biasing users towards the suggestions more beneficial to them, outperforms the system without personalization, in terms of medical accuracy of the obtained knowledge.

RESUMO

A Web alterou profundamente a forma como acedemos à informação. O aumento da informação disponível e o fácil acesso à Web contribuíram para a popularidade das pesquisas neste meio de comunicação. Apesar disto acontecer em várias áreas, é particularmente relevante na área da saúde. De facto, pesquisar por informação de saúde é a terceira atividade mais popular na Internet, sendo feita por quase 3 em cada 4 utilizadores de Internet americanos. Para além da sua popularidade, as pesquisas de saúde na Web são particularmente importantes para informar os consumidores de saúde, encorajando-os a ser mais participativos na gestão da sua saúde. No domínio da saúde, o contexto que rodeia uma pesquisa é extremamente rico e, na nossa perspetiva, tem potencial para melhorar o desempenho dos sistemas de recuperação. Apesar da área de recuperação de informação ter evoluído significativamente com métodos de recuperação que se exploram apenas a coleção de interrogações e a coleção de documentos, é reconhecido que o contexto deve ser explorado. O contexto pode, por exemplo, ser utilizado para desambiguar uma interrogação, recuperar documentos ajustados ao conhecimento do utilizador ou para ajustar os documentos ao processo clínico do doente.

Nesta dissertação analisamos a influência de características contextuais na recuperação de informação de saúde por consumidores. Com base nesta análise, propomos e avaliamos estratégias de apoio à formulação de interrogações que usam características contextuais. A popularidade das pesquisas de informação de saúde na Web levou-nos a focar a investigação na recuperação web. Optámos por centrar a investigação nos consumidores de saúde devido à falta de investigação focada neste público específico. A tese subjacente a este trabalho refere que a recuperação de informação de saúde por consumidores é afetada por características contextuais, que podem ser usadas no apoio à formulação de interrogações para melhorar o desempenho da recuperação.

Em três estudos exploratórios, analisamos a forma como diversas características afetam várias vertentes do processo de recuperação. Com base nos resultados destes estudos e na importância do apoio à formulação de interrogações num domínio onde a terminologia pode ser um entrave e as traduções para diferentes idiomas nem sempre são óbvias, explorámos o impacto de traduzir interrogações em utilizadores com diferentes características. Com base no pressuposto de que uma interrogação formulada num idioma que seja popular na Web consegue, mais facilmente, recuperar conteúdos de qualidade, um dos estudos analisa o efeito de traduzir uma interrogação para Inglês em utilizadores com diferentes níveis de proficiência neste idioma. Os resultados mostram que os utilizadores que têm, pelo menos, uma proficiência elementar em Inglês beneficiam de sugestões de interrogações em Inglês. Os restantes estudos focam-se na tradução da terminologia da interrogação, ou seja,

na tradução entre terminologia leiga e terminologia médico-científica, considerando a literacia em saúde e a familiaridade do utilizador com o tópico. Apesar de já terem sido propostas diversas estratégias para ultrapassar as diferenças terminológicas entre consumidores de saúde e documentos web, nenhuma considera a literacia em saúde e a familiaridade dos utilizadores com o tópico. Os resultados sugerem que os utilizadores com literacia em saúde insuficiente e utilizadores que não estejam familiarizados com o tópico devem ser ajudados com recomendações de interrogações em terminologia leiga. Por outro lado, aos utilizadores com mais literacia em saúde ou familiaridade com o tópico devem ser apresentadas sugestões com terminologia médico-científica.

Com base nos resultados mencionados acima, foi desenvolvido um sistema de sugestão de interrogações que, usando informação do domínio obtida de um vocabulário com terminologia usada por consumidores de saúde, identifica os conceitos médicos existentes na interrogação e devolve quatro tipos de sugestões combinando os idiomas Português e Inglês com as terminologias leiga e médico-científica. As sugestões oferecidas pelo sistema tiveram uma boa aceitação, sendo as sugestões inglesas preferidas às portuguesas pelos utilizadores dos níveis superiores de proficiência em Inglês e as sugestões médico-científicas preferidas às sugestões leigas pelos níveis mais altos de literacia em saúde. Conclui-se que um sistema de recuperação que inclua a estratégia de sugestão implementada sem qualquer tipo de personalização tende a ser melhor do que um sistema sem sugestões em termos de precisão, correção do conhecimento adquirido e também da sua incorreção. Concluímos também que este sistema tende a estar associado a valores mais baixos de relevância motivacional. Nestas tendências, apenas a diferença relativa à incorreção é significativa. Este aspeto é extremamente relevante no domínio da saúde onde informação incorreta pode ter consequências graves. Descobriu-se também que a personalização deste sistema à proficiência em Inglês e à literacia em saúde do utilizador, apresentando apenas as sugestões mais indicadas para cada utilizador, supera o sistema sem personalização, em termos de precisão médica do conhecimento obtido.

CONTENTS

Appendices

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would especially like to thank my advisor, Cristina Ribeiro, for her willingness to support my exploration ideas, for her interesting insights and consistent encouragement. Her warm hospitality when, in 2009, I started working at the Faculty of Engineering, and her valuable counseling in my professional life cannot go unnoticed. I am extraordinarily lucky to have such a supportive mentor and role model. Thank you!

Two of the user experiments conducted in this PhD involved the collaboration of medical doctors who evaluated the medical accuracy of users' answers. For this reason, for her availability despite her busy schedule, and for her interesting perspectives on the user experiments, I would like to thank Dagmara Paiva. She assessed the medical accuracy of all the answers in the user experiment described in Chapter 8 and, in the experiment of Chapter 14, she managed and belonged to a committee that defined the list of correct answers and the list of items that should be ignored for each question included in the simulated situation. In the latter experiment, she also assessed the medical accuracy of 30% of the answers. I would also like to thank to Michael Luís for judging 30% of the answers of the experiment described in Chapter 8 and to Diana Mesquita for asking him to do so.

For providing me valuable contributions in the initial stages of this work, I would like to thank to Ian Ruthven and Altamiro da Costa Pereira, both part of my doctoral committee, and to Arjen de Vries and Padmini Srinivasan, mentors in the SIGIR'2009 doctoral consortium.

I am also grateful to Carlos Duarte, from the Faculty of Arts of the University of Porto, for giving me access to the instrument that was used in the experiment described in Chapter 14 to evaluate users' English proficiency. I also thank to the people who participated in the user experiments presented here.

This work was financially supported by a scholarship from the Fundação para a Ciência e a Tecnologia (FCT), under the grant SFRH/BD/40982/2007. I thank FCT for this important contribution. For conference travel support I also thank to the following organizations: ACM Special Interest Group on Information Retrieval, Fundação Calouste Gulbenkian, Microsoft Research and ACM-W.

Thanks to my friends for all the good moments throughout this period, allowing me to unwind and come back with more enthusiasm.

I thank to Manuel for his wise recommendations and Isabel for her permanent availability and babysitting favors. To my mother, Ana Bela, my father, Rui, and my sister, Cláudia, I am grateful for their love and moral support.

Finally, and foremost, I thank Sérgio for his everlasting optimism and confidence in me. Your wise insights in research issues, your technical suggestions

and your continuous incitement were extremely important. I feel very fortunate to have such a dedicated partner. This work is dedicated to my marvelous children, Miguel and Alice, both born during this journey.

PART I

INTRODUCTION

INTRODUCTION

Information retrieval (IR) is a broad and loosely defined term (Rijsbergen, 1979; Manning et al., 2008) that has become accepted with the work published by Cleverdon, Salton, Sparck Jones, Lancaster and others (Rijsbergen, 1979). One of the earliest definitions, dating from the 1960s, characterizes it as an area concerned with the structure, analysis, organization, storage, searching and retrieval of information (Salton, 1968). More recently, Baeza-Yates and Ribeiro-Neto (1999) presented a similar definition stating that IR deals with the representation, storage, organization and access to information items. An even more recent definition presents IR as "finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)" (Manning et al., 2008).

It is frequent to distinguish IR from *data retrieval*. Rijsbergen (1979) does it in eight different perspectives: matching, inference, model, classification, query language, query specification, items wanted and error response. Baeza-Yates and Ribeiro-Neto (1999) also distinguish these concepts defining *data retrieval* as retrieving all objects that satisfy clearly defined conditions where an erroneous object among the retrieved ones means a total failure. On the contrary, in IR it is not a major problem to identify non-relevant documents in the retrieved set of documents because this field of study deals with non-structured information that can be semantically ambiguous (Baeza-Yates and Ribeiro-Neto, 1999; Manning et al., 2008; Allan et al., 2003). Research in database systems is usually associated with the *data retrieval* field.

1.1 BRIEF HISTORY OF IR

The first manual IR systems appeared with the Sumerians, in the beginning of 3000 BC, when they constructed areas for storing and classifying written materials using cuneiform inscriptions, one of history's oldest writing system, to assist the operation of various social groups (Casson, 2002). As time went by, inventions like paper and the printed press raised the importance of systems to store, manage and retrieve information. After the invention of computers, in 1945, Bush (1945) wrote the article "As We May Think" in which he criticizes the artificiality of the epoch's indexing systems. He idealizes a system that operates through association like the human mind, and is mechanized so it may be consulted with speed and flexibility. This was the first conception of an automatic IR system. In 1950, Mooers (1950) coined the term *Information Retrieval*.

Since the late 1950s, the field of IR has evolved through several relevant works such as: the statistical approach proposed by Luhn (1957); the SMART system by Salton (1971) and his students where important IR concepts like the vector space model and relevance feedback were developed; Cleverdon (1967) Cranfield evaluation model; Jones (1972)' development of *idf* and the probabilistic models by Robertson and Jones (1976), Croft and Harper (1979) and Turtle and Croft (1991). The Text REtrieval Conference (TREC)[1] started in 1992, providing the necessary infrastructure for large-scale evaluation, allowing the modification of old models and techniques and the proposal of new ones.

With the development of the World Wide Web (Web), the interest in the IR field has spread from information specialists like librarians, the legal community and other to a broader audience (Allan et al., 2003; Callan et al., 2007; Manning et al., 2008). The freedom and low cost of publishing on the Web have contributed to a continuous increase of online information that, in turn, demands more from search systems. This contributed to a greater interest in the area, stimulating significant progresses on the field. The area has evolved from simple document retrieval to a broad range of related areas like question-answering, cross-lingual search, topic detection and tracking, summarization, multimedia retrieval and others (Allan et al., 2003; Callan et al., 2007)

Despite the numerous recent developments, IR is far from being a "solved problem" (Callan et al., 2007). In fact, information is being produced more than ever (Lyman and Varian, 2003) and the ways people produce, search, manage and use information is rapidly evolving.

1.2 MOTIVATION

In this section we present the motivation behind our work, describing the general importance of health IR and the specific importance of context in this retrieval domain.

1.2.1 The prevalence and importance of health IR

The Web and its advantages in terms of information access have a great potential in the health domain. In his book, Gigerenzer (2003) relates part of a discussion among 60 physicians, in which medicine is compared to the church in the sixteenth century, needing a reformation. According to the organizer of the meeting, the Internet is, nowadays, the equivalent to the printing press that boosted the Reformation, allowing the access to medical information that was difficult to obtain before. In his opinion, "the Internet can help us to level the disparity between the physician and the patient, the infallible and the uninformed". Moreover, his vision of reformation also includes having "doctors [...] use the best available evidence and consider the patients' goals".

Haux et al. (2002) agree with this vision in which the Internet has the potential to transform the health area. In a prognosis for health care in the year of 2013, Haux et al. (2002) say that "Patients and their families will be knowledgeable of the information resources available over the Internet and will make use of them. New services will arise", predicting the number of accesses to health

[1] http://trec.nist.gov/

web sites to increase by more than 30% and that over 20% of patients will inform themselves about health on the Internet. Moreover, "Knowledge about diseases will be current, comprehensive and internationally available via electronic media, including their availability to patients and their family members ('consumers'). This knowledge will be available in different qualities. Therefore, internationally accredited certification will be available for their contents (e.g. by specialty associations). Knowledge support will partially be integrated in clinical routines.". They predict that over 80% of polled medical knowledge will result from electronic media and that over 80% of the guidelines used routinely in clinical work, will be available electronically. These predictions depict well the present reality.

Patients, their family and friends, commonly designated by health consumers, are increasingly using the Web to search for health information. This is the third most popular online activity following email and using a search engine Fox (2011). A recent report on health information (Fox and Duggan, 2013) reveals that 59% of the American adults look online for health information. Among the Internet users, this proportion rises to 72%. The probability of conducting this type of searches is higher in women, younger people, white adults, those who live in households earning $75,000 or more, and those with a college degree or advanced degrees. This study also found that 77% of the health searches start at generalist search engines and 13% start at health-specific websites. A different study found that this activity is also popular on mobile devices, concluding that 52% of smartphone owners gather health information on their phones (Fox and Duggan, 2012).

In Portugal, the reality is a little less expressive but also depicts a growing and noteworthy tendency. In 2008, 19.6% of the population used the Internet to search for health information (Espanha and Lupiáñez Villanueva, 2008) and in 2012 this proportion increased to 25.7% (Espanha et al., 2012). The age group in which more health searches are conducted is the one of 25 to 44 years (Espanha and Lupiáñez Villanueva, 2008). Users mainly do it to search information for themselves (83.1%), once in a while (56.7%) and mostly using search engines (84.5%) (Espanha and Lupiáñez Villanueva, 2008). Users go online mostly to obtain specific information about health conditions (86.1%) and to increase their health knowledge (82.7%) (Espanha et al., 2012).

Consumers' use of the Web to search for health information is a controversial matter. There are clinicians who are against it claiming that consumers don't have the abilities to correctly interpret health information and to evaluate the quality of what they find online. On the other hand, other clinicians feel that, far from replacing the doctor, online health searches empowers consumers in the management of their health and encourages them to become more participatory.

Lately, the control and access to health information by consumers has been widely discussed. Plenty of government initiatives have appeared all over the world aiming to improve consumers' access to patient records. Generally, the movement for a consumer-driven healthcare is about giving consumer more power to manage their health. This can be done by giving them access to their health record anywhere/anytime, making it easy to find all needed information and providing online tools that enable personalized advice. According to Bosworth (2007), in a good health system, consumers should be able to discover the most relevant information, have access to health support person-

alized services and be able to learn from and educate those in similar health circumstances.

The importance of an easy access to online health information is recognized by the USA Department of Health and Human Services (2010) that set the following goal for 2020: "Increase the proportion of online health information seekers who report easily accessing health information".

1.2.2 The importance of context in (Health) IR

Typically, IR systems support their decisions solely on the query and document collection. Several implicit factors about the user and the search context (e.g. time, location, task, expertise, previous interaction) are ignored and have the potential to improve system's performance. All information activities take place within a context that affects the way people access information, interact with a retrieval system, make decisions about the documents retrieved and evaluate retrieval systems (Harper and Kelly, 2006; Ingwersen et al., 2005).

Context is gaining attention in the field of IR (Bierig and Göker, 2006) and its inclusion in IR has been pointed as one of the grand challenges for this area by several authors (Allan et al., 2003; Belkin, 2008; Hersh, 2008b). Yet, the difficulties of capturing and representing knowledge about context has been slowing the progress in this area (Allan et al., 2003). In Chapter 3 we describe, in more detail, the usefulness of context in IR and the increasing attention that context has been receiving from this research community.

In domains like health, the search context is extremely rich. Just like when someone goes to the doctor and explains the context of their condition, health searches could be improved if the IR process took into account this information. Patient characteristics like age, gender or health record and searcher characteristics like health literacy, knowledge on the topic being searched for or his proficiency in other languages are examples of context features that could be used. Moreover, the characterization of the scenario (e.g. treatment, diagnosis) in which the search occurs or information related to the topic being searched can also be used. The large quantity of contextual data that may be associated with an Health IR (HIR) system makes us believe that context can be useful to improve it.

1.3 PROPOSED THESIS

The relevance of context in the health field (Silva and Favela, 2006) and the contribution of HIR to well-informed health consumers (Ferguson, 2007) and professionals, motivate this research. The access advantages of the Web and the prevalence of online health searches make us focus on online health information retrieval, that is, retrieval done on the Web. Moreover, the lack of research focused on health consumers, as will be shown in Chapter 3, make us focus on this specific public instead of health professionals.

The goals of this research are to analyze what elements of context are potentially significant to HIR activities and to propose novel ways to improve HIR performance with the exploration of context features. More specifically, it intends to:

- Investigate the effects of context features on the retrieval process;

- Develop and evaluate strategies to support query formulation based on the studied context features;

and, to a lesser extent, to:

- Study how health information seeking behavior can be used to predict context features.

We decided to focus on query formulation support for its potential to incorporate context features and its smaller demands in terms of technical infrastructures and specific data sets. The two last goals are identified as characteristics of future search engines by a large number of renown figures in the field of Information Retrieval (Allan et al., 2003). These authors state that "Future search engines should be able to use context and query features to infer characteristics of the information need [...] and use these characteristics [...] to rank potential answers [...]". Our thesis can be stated as:

Consumer health information retrieval is affected by context features that can be used to support in query formulation support to improve retrieval performance.

1.4 RESEARCH OUTLINE

Considering the potential of context features surrounding the search to improve retrieval, this dissertation addresses the influence of context in IR performed by health consumers. This is a domain where searches have been growing, where terminology can be a barrier and where the quality of the retrieved information is crucial. This dissertation deals with this subject through several exploratory studies that analyze the influence of context features and through studies focused on query operations. Moreover, we propose a query suggestion system and evaluate it considering users' characteristics. In total, we conducted three user experiments that served as basis for our studies and for the evaluation of the query suggestion system. Moreover, we also did some preliminary work on the automatic detection of context features. In this section we describe the contents of this dissertation in four subsections: exploratory studies, use of context in query formulation support, context prediction and adopted methodology.

1.4.1 Exploratory studies

As will be seen in Chapter 3, context is a comprehensive concept, embracing several types of features. To embrace a large number of features and to analyze their impact on different outcomes of the retrieval process, we conducted three exploratory studies. In the first, described in Chapter 5, we concentrate on how search engines and task features influence precision. A summary of the analysis done in this study and the context features involved in it is depicted in Figure 1.1. Solid grey arrows point to the dependent variable from the independent one. Dashed vertical arrows represent variables considered during the analysis to which they point. For example, when we study the influence that the search engine has in precision, we do it in general and also in each type of clinical question (e.g.: diagnosis, treatment), medical specialty (e.g.: gastroenterology, cardiology) and condition severity.

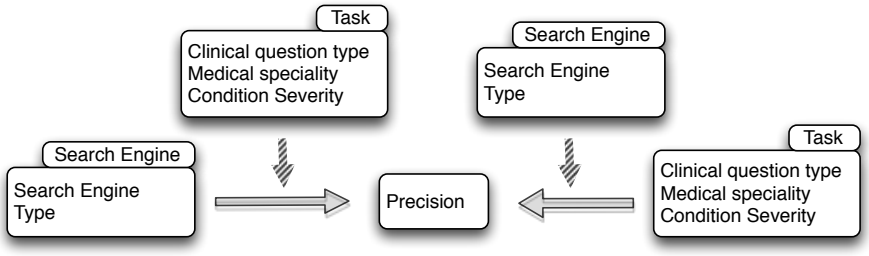

Figure 1.1: Comparative evaluation of search engines in health information retrieval - analysis and context features.

In the second exploratory study, described in Chapter 6, we focus on document features and we consider a wider range of dependent variables, as shown in Figure 1.2.

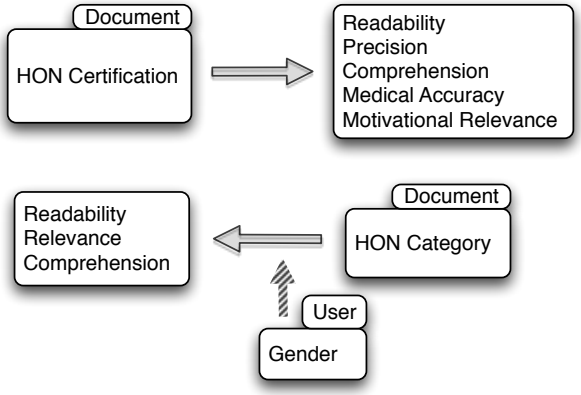

Figure 1.2: Data certification impact on health information retrieval - analysis and context features.

Chapter 7 describes the third exploratory study, the broadest in terms of context features. As shown in Figure 1.3, we consider user features like age, gender, web search skills and health search skills, task and topic features, documents features and also interaction features where the characteristics of the query are included. The influence of these features is analyzed in terms of query formulation and situational relevance assessment.

1.4.2 Use of context in query formulation support

In an IR system, a search is initiated by a user that specifies a query reflecting his information need. This query is then parsed and processed by the system that can also use it to suggest alternative queries or terms. The user can then accept or consider these suggestions to improve his query. This support is useful in many circumstances but it is particularly important in the health domain where users' terminology and knowledge may be scarce and where translations to other languages are not always obvious. These assumptions and some of the conclusions of the exploratory studies led us to analyze how the language and terminology of the query can influence IR in users with different characteristics. A first study focuses on the query language and is described in Chapter

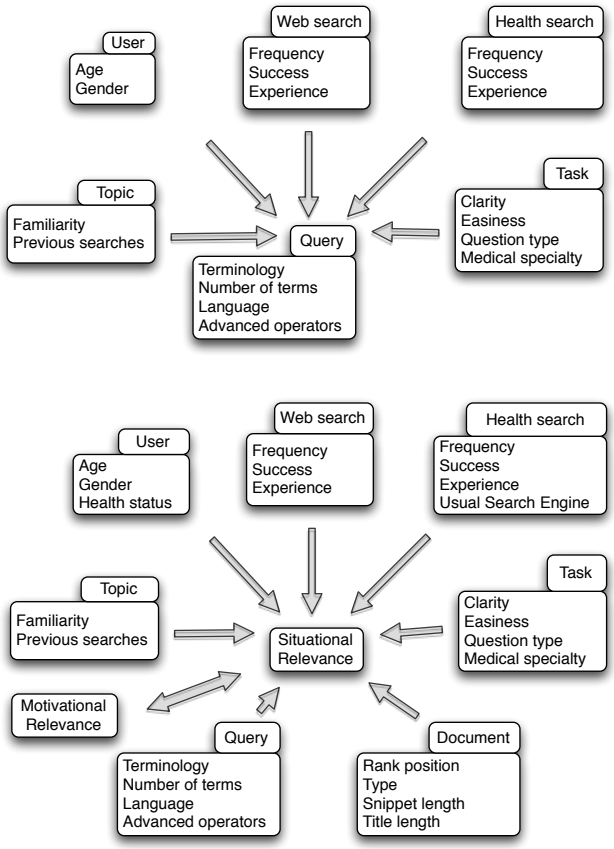

Figure 1.3: Context effect on query formulation ad subjective relevance in health searches - analysis and context features.

9. It is based on the assumption that languages with a stronger presence in the Web are associated with larger probabilities of retrieving quality health contents. As shown in Figure 1.4, it analyzes how translating a query to the English language affects documents' characteristics and, considering the user English proficiency, how it affects several outcomes of the retrieval process. In addition, it examines how documents' readability influences comprehension and relevance assessments considering the document's language and the English proficiency of the user. Finally, it investigates how previous interactions and users' English proficiency affect query formulation in terms of language.

Chapters 10, 11 and 12 focus on the terminology of the query considering users' health literacy and topic familiarity. Although, in the existing literature, several strategies have been proposed to overcome the terminology gap between health consumers and documents' language, none considers users' health literacy and topic familiarity. Moreover, in IR, the number of research works dealing with health literacy is scarce and, since it is "the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions" (USA Department of Health and Human Services, 2000), it deals with abilities that are critical in HIR. In the 2003 USA assessment of adult literacy (Kutner et al., 2006), it was found that 36% of adults in the United States have basic or below basic health literacy skills.

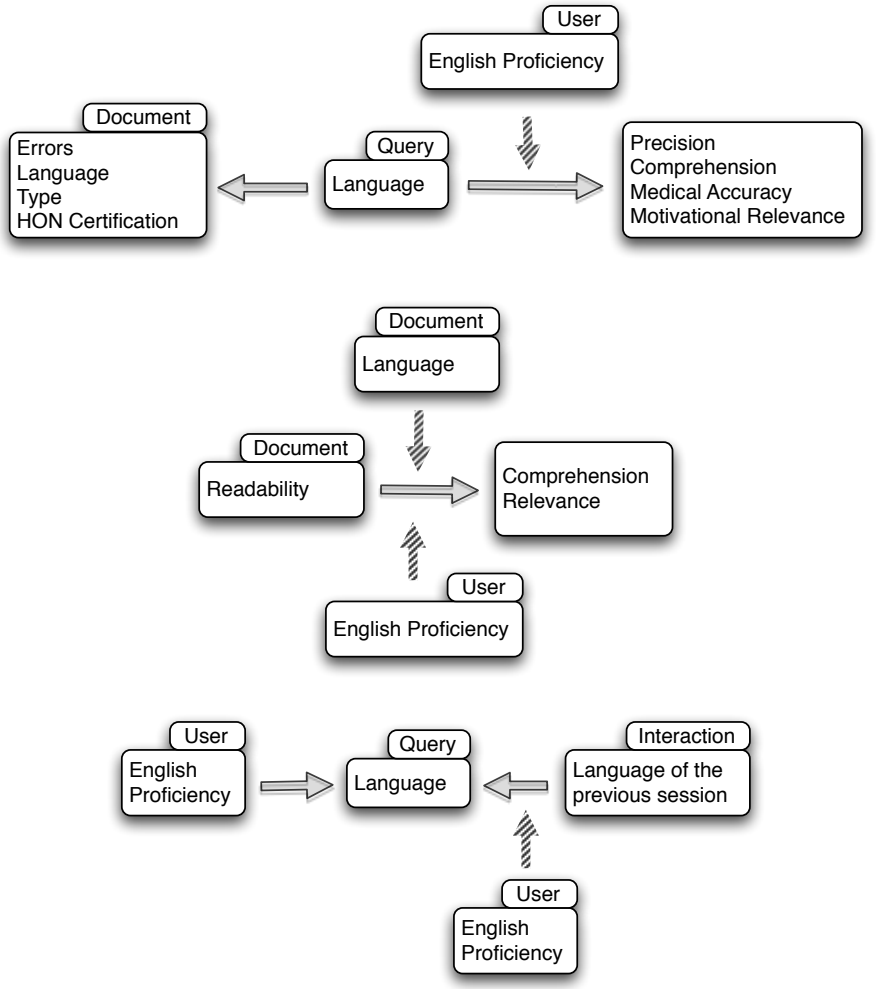

Figure 1.4: Measuring the value of health query translation: an analysis by user language proficiency - analysis and context features.

In Chapter 10, as depicted in Figure 1.5, the analysis is similar to the one shown in the top-up diagram of Figure 1.4, differing only in the query and user characteristics. In Chapter 11, we study how readability, comprehension, precision, medical accuracy and motivational relevance influence each other, considering the terminology of the query, as shown in Figure 1.6. In Chapter 12, as shown in Figure 1.7, we analyze how user characteristics and previous interaction affect user behavior in terms of query terminology.

Based on some of the results obtained in the studies described above, we decided to implement a query suggestion system that can be integrated in an IR system. The query suggestion system uses domain information gathered from an existing consumer health vocabulary, identifies the medical concepts included in the query and returns four types of suggestions combining the Portuguese or English language with the lay or medico-scientific terminology. More details regarding the query suggestion system are given in Chapter 13. This query suggestion was evaluated with a user experiment in which users performed half of the assigned tasks in an IR system with suggestions and the other half in a system without suggestions. The evaluation, as described in

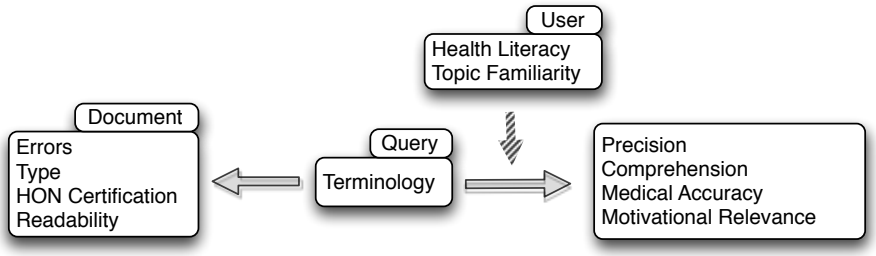

Figure 1.5: Effects of query terminology on health searches: an analysis by user's health literacy and topic familiarity - analysis and context features.

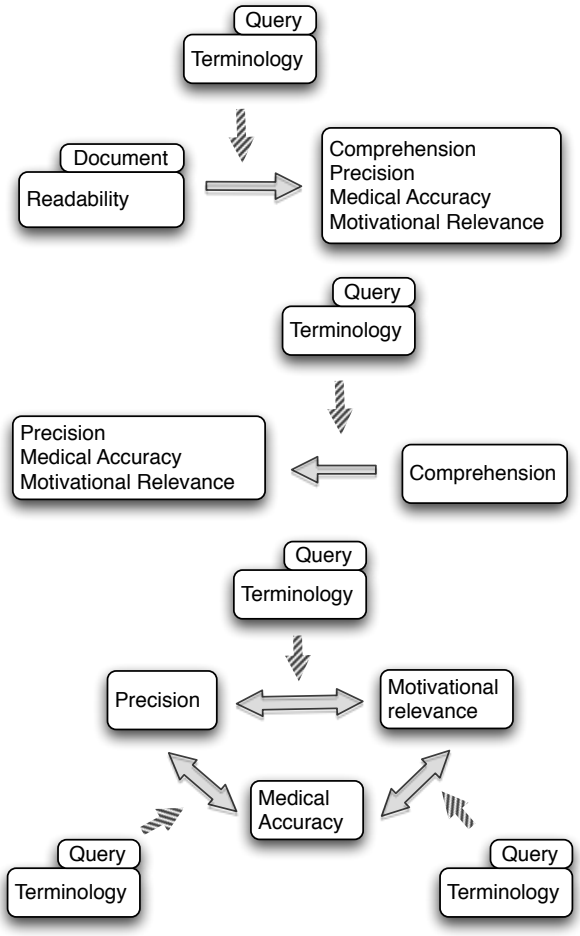

Figure 1.6: Interplay of context features considering the terminology of the query - analysis and context features.

Chapter 14 and summarized in Figure 1.8, focused on users' behavior in terms of suggestion use, compared the systems with and without suggestions (not depicted in Figure 1.8 because it does not involve context features) and analyzed how suggestions, in general and in particular, affected precision, medical accuracy and motivational relevance. In addition, we also analyzed if a personalized version of this suggestion system, that is, a system where each user only saw the suggestions more beneficial to him, would yield better results.

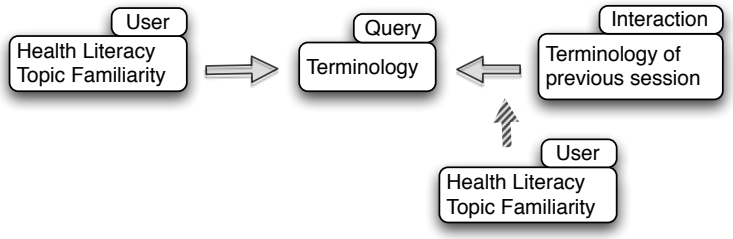

Figure 1.7: Query behavior: the impact of health literacy, topic familiarity and terminology - analysis and context features.

Chapter 15 discusses the evaluation results.

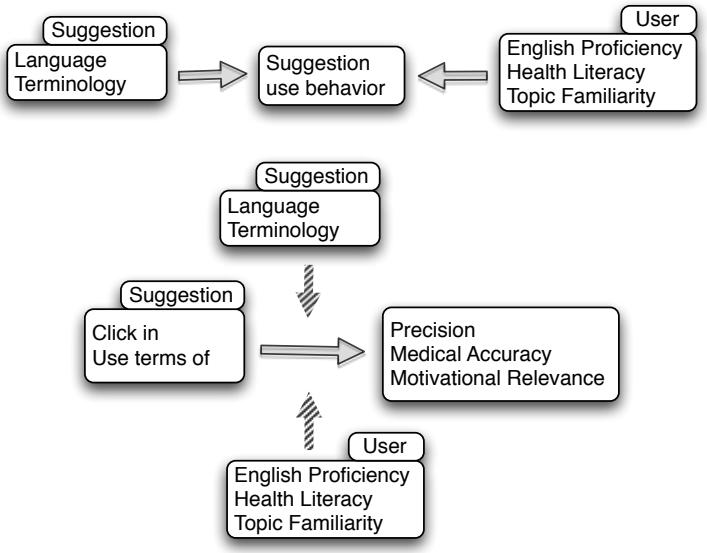

Figure 1.8: Evaluation of the query suggestion system - analysis and context features.

1.4.3 Context prediction

An IR system that incorporates context features should gather these features automatically, without interfering with the typical interaction between the user and the search engine. For a search engine, it is important to detect when the user is conducting a search in the health domain and to predict the context features that will be used to personalize the search experience, such as users' health literacy, topic familiarity and proficiency in languages other than the native one.

Although the prediction of context features was not the primary focus of this dissertation, we proposed 2 types of methods to identify health queries and compared them with a method proposed by other authors. This work is described in Chapter 16. Our methods are based on an existing health consumer vocabulary and the other method is based on the co-occurrence of the query terms with the word "health" in the Web. One of our proposed methods also has the ability to classify queries into categories like *Disease or Syndrome*

or *Anatomical Structure*. Note that, while the detection of health queries is useful for several types of personalization, it might not be essential. For example, a query suggestion system using a domain vocabulary, like the one we propose, might use information gathered from the intersection of the query terms and domain concepts to infer if the query is related to the health domain and decide if suggestions should or should not be presented.

In Chapters 9 and 12, we also superficially analyze if it is possible to raise hypothesis related to the prediction of English proficiency, health literacy and topic familiarity through users' search habits in terms of language and terminology used in past queries.

1.4.4 Methodology

The use of test collections is still the dominant approach to evaluate IR systems (Sanderson, 2010), being particularly well suited to system-oriented performance evaluations focusing on specific aspects of systems. Although popular, this type of evaluation is restrictive in terms of the cognitive and behavioral features of the IR system's environment (Borlund, 2003b). For these reasons, we use alternative evaluation methods. In total, we conducted three user experiments, described in Chapters 4, 8 and 14. Every experiment involved human subjects, was conducted in the laboratory and constitutes what Ingwersen (2009) calls an *interactive light* experiment. This type of experiment entails session-based multi-run interaction with monitoring like log analysis, interviews and observation. All the experiments included *simulated work task situations*, that is, short 'cover stories' describing the situation that leads to the use of the IR system, as defined by Borlund (2003b) in her Interactive IR evaluation model. Moreover, as suggested by Borlund (2003b), to achieve a higher degree of realism, users assessed relevance in a 3-graded scale. As a result, we used performance measures that take into account the non-binary nature of this scale.

Excluding the study described in Chapter 5, in every study, we evaluated the impact of the independent variable (e.g.: query language, query terminology, click in suggestion) in a holistic perspective, considering more than just the precision computed from users' relevance assessments. This way we assure we assess the global impact of the IR system on the user. In addition to precision, we consider users' comprehension of documents, the medical accuracy of the knowledge acquired during the session and the motivational relevance assessed through users' feeling of success and satisfaction (Saracevic, 1996).

It is interesting to note that the above experiments are aligned to what has been described as *Contextual Evaluation* by a brainstorming workshop held in 2012. This workshop intended to envisage future directions and perspectives for the evaluation of information access and retrieval systems (Agosti et al., 2012). *Contextual Evaluation* was identified as one of the six main challenges and described as an evaluation integrating users, tasks, search applications and underlying information retrieval systems in a holistic perspective assuring that global aspects like impact on users' workplace productivity or on the quality of life can be assessed. According to the report produced in this workshop, it is by understanding users and their context that systems can be improved. To provide a holistic approach to evaluation, the participants of this workshop suggested that both pre and post-retrieval contexts should be taken into account

in order to obtain a more realistic evaluation and one that is more related with the actual user satisfaction. The third experiment of this dissertation, although designed prior to these results, has taken into account the medical accuracy of users' knowledge in both stages, before and after the search.

1.5 PUBLISHED WORK

A part of the work presented in this dissertation has already been published as research papers and posters in peer-reviewed international journals, conferences and workshops.

Journals

1. Carla Teixeira Lopes, Cristina Ribeiro. **Measuring the value of health query translation: An analysis by user language proficiency**. Journal of the American Society for Information Science and Technology, 2013. ISSN 0002-8231. DOI: 10.1002/asi.22812.

2. Carla Teixeira Lopes, Cristina Ribeiro. **Comparative evaluation of web search engines in health information retrieval**. Online Information Review, 2011, vol. 35, iss. 6. ISSN 1468-4527. DOI: 10.1108/14684521111193175.

Conference Papers

3. Carla Teixeira Lopes and Cristina Ribeiro. **Query Behavior: The Impact of Health Literacy, Topic Familiarity and Terminology**. Accepted in SouthCHI 2013.

4. Carla Teixeira Lopes, Daniela Dias and Cristina Ribeiro. **Using Domain-specific Term Frequencies to Identify and Classify Health Queries**. In Álvaro Rocha, Ana Maria Correia, Tom Wilson and Karl A. Stroetmann, editors, Advances in Information Systems and Technologies, pages 221-226, Springer-Verlag. 2013.

5. Carla Teixeira Lopes and Cristina Ribeiro. **Data Certification Impact on Health Information Retrieval**. In Andreas Holzinger and Klaus-Martin Simonic, editors, Information Quality in e-Health, volume 7058 of Lecture Notes in Computer Science, pages 31-42, Springer Berlin / Heidelberg. 2011.

6. Carla Teixeira Lopes and Cristina Ribeiro. **Context effect on query formulation and subjective relevance in health searches**. In Proceeding of the Third Symposium on information interaction in Context (New Brunswick, New Jersey, USA, August 18-21, 2010). IIiX'10. ACM, New York, NY, 205-214. 2010.

7. Carla Teixeira Lopes. **Context Features and their use in Information Retrieval**. FDIA 2009: 3rd Symposium on Future Directions in Information Access. University of Pádua, Italy, 2009.

8. Carla Teixeira Lopes. **Context-Based Health Information Retrieval**. In Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in Information Retrieval (Boston, MA, USA, July 19-23, 2009). SIGIR '09. ACM, New York, NY, 845-845. 2009.

Workshops

9. Carla Teixeira Lopes and Cristina Ribeiro. **Context in Health Information Retrieval: What and Where**. In Proceedings of the Fourth Workshop on Human-Computer Interaction and Information Retrieval 2010 (New Brunswick, New Jersey, USA, August 22, 2010). HCIR 2010.

Conference Posters

10. Carla Teixeira Lopes and Cristina Ribeiro. **Using local precision to compare search engines in consumer health information retrieval**. In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (Geneva, Switzerland, July 19-23, 2010). SIGIR'10. ACM, New York, NY, 835-836. 2010.

1.6 DISSERTATION LAYOUT

This dissertation is composed by the 6 following parts.

Part I – Introduction
In this part we motivate the work presented in this dissertation, present the background and survey previous works on health information retrieval and, more specifically, on the use of context in health information retrieval.

Part II – Context Influence on Consumer Health Information Retrieval: Exploratory Studies
This part comprises Chapters 4, 5, 6 and 7. It begins describing the first user experiment conducted in this work to explore the influence of context features in health information retrieval by consumers. This experiment served as basis for the studies described in Chapters 5 and 7. In Chapter 5, we evaluate how search engine context and task context affects precision in HIR. The latter study involves a wider range of context features, related to the user, task, document and query, and intends to analyze their influence on users' query formulation and relevance assessments. The study described in Chapter 6 is based on the user experiment detailed in Chapter 8. This study investigates how documents' characteristics, like HONcode certification and HON categories, influence the readability and comprehension of documents and also the precision, medical accuracy and motivational relevance of the session.

Part III – Query Formulation: contextualization by English Proficiency, Health Literacy and Topic Familiarity
In this part we describe studies that analyze how the language (Portuguese or English) and terminology (lay or medico-scientific) of a query affect the retrieval process considering users' English proficiency, health literacy and topic familiarity. After describing the user experiment that allowed this analysis, in

Chapter 9 we describe the language analysis and in Chapter 10 the terminology analysis. Chapter 11 explores the interplay between readability, comprehension, precision, medical accuracy and motivational relevance, considering the query terminology. Finally, in Chapter 12 we analyze how users' health literacy, topic familiarity and past search sessions influence query formulation behavior. The results and conclusions gathered in this part of the dissertation motivates the development of the query suggestion prototype that will be described in Part IV.

Part IV – Query Suggestion System: Implementation and Evaluation
This part comprises Chapters 13, 14 and 15 in which we describe the implemented query suggestion system and its evaluation based on the third user experiment conducted in this dissertation. In the evaluation we analyze how suggestion types and user characteristics influence the use of suggestions and how their use influence the outcome of the retrieval session in terms of precision, medical accuracy and motivational relevance.

Part V – Automatic context acquisition
This part has only one chapter that proposes two types of methods to identify health queries and compare them with another method previously proposed in the existing literature. Our proposals involve the use of an existing vocabulary with consumer medical terminology. Moreover, one of our methods allows to relate queries with health categories.

Part VI – Conclusion
This part comprises Chapters 17 and 18. The first aggregate the conclusions acquired in all the conducted studies. In the latter, we identify lines of future work.

HEALTH INFORMATION RETRIEVAL

2.1 INTRODUCTION

The reasons that made the IR field substantially evolve in recent years are also applicable to the increasing interest that health IR has been receiving. The Web has profoundly changed the availability and access to health information, not only by health professionals but also by consumers.

To health professionals, applications providing an easy access to validated and up-to-date health knowledge are of great importance to an adequate dissemination of knowledge and have the potential to impact the quality of care provided by these professionals. On the other hand, the Web opened doors to the access of health information by patients, their family and friends, making them more informed and changing their relation with health professionals. Fox and Duggan (2013) have recently released the results of a study where they concluded that "one in three American adults have gone online to figure out a medical condition". Considering only the Internet users, this proportion rises to 72%. Search sessions usually start at a search engine as stated by 77% of the online health seekers. Half of these users conduct their searches on behalf of someone else.

In this chapter we describe how health information is usually classified and identify resources that exist in each category. Moreover, we characterize health representation systems, namely the Medical Subject Headings and other thesauri; the Unified Medical Language System and its three knowledge sources; and existing ontologies. In the end of this chapter we present a brief overview of works done in the specific area of health information retrieval (HIR). These works are presented according to the following high-level research areas: health information seeking, indexing, retrieval, evaluation, user interfaces and visualization. This literature review does not intend to be exhaustive. In the following chapters of this dissertation, we present detailed literature reviews adjusted to the topics covered in them.

2.2 HEALTH INFORMATION

Information plays a crucial role in healthcare professionals' activities and consumers' attitudes. Two early studies concluded that healthcare personal spent about one-third of their time handling and using information (Jydstrup and Gross, 1966; Mamlin and Baker, 1973). According to Hersh (2002), it is likely that the time dedicated to managing information in healthcare is as large, or even larger, nowadays. In this dissertation we adopt the term Health Infor-

mation instead of Biomedical Information, because it's broader than the latter, including not only concepts from biological and medical sciences but also encompassing related areas like health care facilities, manpower and services, health care economics and organizations.

2.2.1 *Classification*

Health information may be in text, audio, images and video formats. It may be classified in patient-specific information and knowledge-based information (Hersh, 2002; Shortliffe and Cimino, 2000). The first type relates to individual patients and its purpose is to tell health professionals about the health condition of a patient. It typically comprises the patient's medical record and it may contain unstructured data (e.g.: lab results, vital signs) or free/narrative text (e.g.: radiology report). Knowledge-based information derives and is organized from observational and experimental research, providing health professionals the knowledge acquired in other situations so it may be applied to individual patients or used to conduct further research. Similarly to other types of scientific information, in health, knowledge-based information can be subdivided in primary information and secondary information. The former includes direct results of original research and the latter encompasses reviews, condensations and summaries of primary literature. As proposed by Hersh (2002), knowledge-based information can also be classified in bibliographic, full-text, databases/collections and aggregations. Each subcategory is described next along with some of its main examples.

2.2.2 *Bibliographic content*

The bibliographic content includes literature reference databases also named bibliographic databases, web catalogs and specialized registries. The distinction between these subcategories is becoming blurry. For example, since literature reference databases started providing links to the referenced literature they became closer to web catalogs.

Literature reference databases catalog books and periodicals and were the original IR databases in the 1960s, designed to guide the searcher and not to provide the actual resources.

MEDLINE, the Medical Literature Analysis and Retrieval System Online, is probably the best known and the National Library of Medicine (NLM) premier bibliographic database (NLM, 2011). It started as the *Index Medicus*, a print catalog of the medical literature, with its first volume published in 1879 (Hersh, 2002). It now provides over 20 million references to articles of approximately 5,600 journals with a subject scope of biomedicine and health. It is freely available via PubMed[1] and a search generates a list of citations (including authors, title, source, and often an abstract) to journal articles, an indication of free electronic full-text availability, generally through PubMed Central[2], or a link to the website of the publisher or other full text provider. It may also be searched using the NLM Gateway, a single Web interface that searches multiple NLM retrieval systems (NLM, 2011). Other websites also provide access

[1] http://pubmed.gov
[2] http://www.pubmedcentral.nih.gov

to MEDLINE, some for free and others for some fee, usually providing value-added services.

Besides MEDLINE, NLM has other databases organized in three categories: citations to journals and other periodicals since 1966, citations to books, journals and audiovisual material, and citations to journal articles prior to 1966 and scientific meeting abstracts. The first set of databases is accessible through PubMed that is composed by MEDLINE, MEDLINE in-process citations and publisher-supplied citations. The second type of databases is available through LOCATORplus[3]) and the third through NLM Gateway[4].

In addition to NLM, there are other creators of bibliographic databases, both public and private. Some are produced by other USA National Institute of Health's organizations like the National Cancer Institute (NCI). Some of these bibliographic databases tend to be more focused in specific resources and subject types like CINAHL (Cumulative Index to Nursing and Allied Health Literature[5]) – the major non-NLM database for the nursing field.

Web catalogs are web pages that contain links to other pages and sites, sharing many features with traditional bibliographic databases (Hersh, 2002). The number of such catalogs is increasing (Shortliffe and Cimino, 2000). Some well-known catalogs are: MedicalMatrix[6], Hardin[7], HealthFinder[8], HON Select[9] and MedWeb Plus[10].

Specialized registries may overlap with literature reference databases and web catalogs but, generally, they point to more diverse information resources. One famous specialized registry is the *National Guidelines Clearinghouse* (NGC[11]), produced by the Agency for Healthcare Research and Quality, with information about clinical practice guidelines.

2.2.3 Full-text content

This subcategory includes resources with complete online versions of periodicals and books. Originally, full-text databases were mainly online versions of journals and they only started to include books with the decrease of the price of computers and websites.

Most periodicals are nowadays published electronically, which allows an easier access and the provision of additional data like figures, tables, raw data and images and true bibliographic links. Some publishers make their papers freely available on the Web in initiatives like the BiomedCentral[12] (BMC), Public Library of Science[13] (PLOS) and PubMed Central (PMC).

Textbooks are also increasingly publishing versions on the Web. These electronic versions allow several additional features over the printed versions:

[3] http://locatorplus.gov
[4] http://gateway.nlm.nih.gov
[5] http://www.cinahl.com
[6] http://www.medmatrix.org
[7] http://www.lib.uiowa.edu/hardin/
[8] http://www.healthfinder.gov
[9] http://www.hon.ch/HONselect
[10] http://www.medwebplus.com
[11] http://www.guideline.gov/
[12] http://www.biomedcentral.com
[13] http://www.plos.org

high-quality images; multimedia content; links to other resources; interactive self-assessment questions and an easier access to book updates.

The third type of full-text content is composed of web sites that provide full-text information and services like interaction with experts and links to other sites. This excludes web sites that implement services such as bibliographic databases, online versions of books and other printed material, specialized databases and collections and aggregations of these. There are several health full-text information sites that are developed by communities, from consumers to governments. Some examples are: Intelihealth[14], Netwellness[15], WebMD, eMedicine[16], Medscape[17] and Institute for Clinical Systems Improvement guidelines[18]. The first three sites are directed to consumers while the last three are more oriented to health professionals.

2.2.4 Databases and Collections

This category consists of databases and collections of specific information like images (from radiology, pathology and other areas), genomics (gene sequencing, protein characterization and others) and Evidence Based Medicine resources. There are several health-image databases available on the Web. One of the most famous is the Visible Human Project[19], which consists of three-dimensional representations of normal male and female bodies built from cross-sectional slices of cadavers.

Genomics studies the genetic material in living organisms and its research has been evolving rapidly in recent years. One of its main driving forces was the Human Genome Project, led by the National Human Genome Research Institute, that ended in April 2003 with the production of a version of the human genome sequence that is freely available[20]. Several genomics databases are available across the Web and at the center are those produced by the National Center for Biotechnology Information (NCBI). NCBI's databases are linked among themselves, and with PubMed, in the NCBI's Entrez system[21].

On the Web there are also databases that hold the evidence-based medicine principles and try to eliminate the problems of scattering and fragmenting the primary literature. These databases provide systematic reviews (e.g.: The Cochrane Database of Systematic Reviews[22]) or highly-synthesized synopses of evidence-based information (e.g.: Clinical Evidence[23], DynaMed[24], PIER[25], UpToDate[26]).

[14] http://www.intelihealth.com/
[15] http://www.netwellness.com/
[16] http://www.emedicine.com/
[17] http://www.medscape.com/
[18] http://www.icsi.org/guidelines_and_more/
[19] http://www.nlm.nih.gov/research/visible/visible_human.html
[20] http://www.genome.gov/10001772
[21] http://www.ncbi.nlm.nih.gov/Entrez/
[22] http://www.cochrane.org
[23] http://www.clinicalevidence.com
[24] http://www.dynamicmedical.com
[25] http://pier.acponline.org
[26] http://www.uptodate.com

2.2.5 *Aggregations*

This last category includes aggregations of the first three categories for all types of users, consumers, health professionals and scientists. The distinction between this category and some of the above content with several links is blurry but, typically, aggregations have a larger variety of information that serves diverse needs of their users. They are, for example, websites that collect several types of content to generate a coherent resource.

One of the largest aggregated consumer information resources is MedlinePlus, a service provided by the NLM and the NIH. It aggregates information from these entities and from other trusted sources on diseases and conditions. It also has pre-formulated MEDLINE searches to give access to medical journal articles, information on drugs, an illustrated encyclopedia, a medical dictionary, links to clinical trials, interactive patient tutorials and updated health news[27]. CancerNet is another consumer-oriented service from the National Cancer Institute[28] (NCI) that contains information on all aspects of the disease. Healthline[29] focuses on consumers containing online health search, content and navigation features. The Microsoft HealthVault[30] combines the aggregation of contents with other services. It is a hub of a network of sites, personal health devices and other services to let consumers manage their health.

There are also aggregated content more directed to professionals such as: MDConsult[31], a service of Elsevier that aggregates medical resources in an integrated way to help health professionals and Merck Medicus[32], developed by Merck, available to all licensed US physicians, which includes resources like Harrison's Online and MDConsult.

2.3 HEALTH INFORMATION STRUCTURES

IR benefits from the availability of well-defined information structures that can be used in the indexing and retrieval processes. Health information is, by its nature, highly detailed (Shortliffe and Cimino, 2000), having an old tradition in classification dating back to Aristotle's effort in biology and formal descriptions (Pellegrin, 1986). The representation of health concepts is more challenging than in many domains due to its levels of precision, complexity, implicit knowledge and breadth of application (Shortliffe and Cimino, 2000). However, it is also an area where great efforts have been developed and several representation systems have appeared.

Terminologies, controlled vocabularies/thesauri and ontologies are three ways to represent health information with an increasing degree of formalism. A terminology or vocabulary is a list of terms representing the concepts used in a specific area. When simple relationships among different terms are specified, this representation system becomes a controlled vocabulary or a thesaurus. The relationships are typically of three types: hierarchical (terms are broader or narrower), synonymous or related (terms with relations other than hierar-

[27] http://www.nlm.nih.gov/medlineplus/
[28] http://www.cancer.gov
[29] http://www.healthline.com
[30] http://www.healthvault.com
[31] http://www.mdconsult.com
[32] http://www.merckmedicus.com

chy or synonymity). Ontologies are the most formal representation type, used for machine-processable structures where categories of terms and relationships can be freely defined. They must also exhibit internal consistency, acyclic polyhierarchies and computable semantics (Shortliffe and Cimino, 2000). Recently there has been an explosion of health ontologies (Smith and Rosse, 2004).

Health representation systems may be used in several areas such as Information Retrieval, Natural Language Processing, Semantic Interoperability and Decision Support Systems. In Information Retrieval, they may be used in the indexing process (manual indexing is usually done using a thesaurus) and in the retrieval process (e.g.: synonymous terms may be used to improve the expression of the information needs).

This section starts to describe MeSH, the NLM's thesaurus that is used to index most of the NLM's databases, followed by other non-NLM thesauri used in the health area. Most of these structures have origin in the USA and are globally adopted, receiving contributions from worldwide researchers. Then, NLM's Unified Medical Language System, together with its 3 knowledge sources: Metathesaurus, Semantic Network and the SPECIALIST Lexicon and Tools are presented. In the end, some of the main health ontologies and two general ontologies are briefly presented.

2.3.1 Medical Subject Headings

The Medical Subject Headings (MeSH) is the NLM's thesaurus used to index most of the NLM's databases (Coletti and Bleich, 2001). The 2013 version of MeSH has 26,853 descriptors (NLM, 2012b) and its vocabulary files may be downloaded from the NLM site, at no charge, in XML or ASCII format.

The hierarchies in which descriptors are placed are named trees[33]. Each descriptor appears in one or more places of the tree. The XML MeSH is structured in three levels: Descriptor, Concept and Term. A descriptor includes a set of concepts, each described by a set of terms which are synonymous with each other. For example[34]:

```
Cardiomegaly [Descriptor]
   Cardiomegaly [Concept, Preferred]
      Cardiomegaly [Term, Preferred]
      Enlarged Heart [Term]
      Heart Enlargement  [Term]
   Cardiac Hypertrophy [Concept, Narrower]
      Cardiac Hypertrophy [Term, Preferred]
      Heart Hypertrophy [Term]
```

Each concept has a preferred term, which is also the name of the concept, and each descriptor has a preferred concept. The name of the descriptor corresponds to the preferred term of the preferred concept. The terms in one concept are not strictly synonymous with terms in another concept, even in the same record. Additionally, MeSH has two types of relationships: hierarchical and associative (NLM, 2001). The first type is a crucial component of a thesaurus and is formalized by the MeSH tree structure that represents distinct levels of specificity (terms that are broader or narrower). MeSH descriptors[35]

[33] The 2013's MeSH list of trees is available at: http://www.nlm.nih.gov/mesh/trees.html
[34] Obtained from: http://www.nlm.nih.gov/mesh/concept_structure.html
[35] http://www.nlm.nih.gov/mesh/MBrowser.html

are organized in 16 categories (NLM, 2012c). The associative relationships are often represented by the “see related" cross-reference. They can be used to add/suggest terms to a specific search, to point out in the thesaurus the existence of other descriptors, which may be more appropriate, or to point out distinctions made in the thesaurus or in the hierarchical structure of the thesaurus.

Besides descriptors, also called MeSH headings, MeSH has additional types of vocabulary: qualifiers or subheadings, check tags, publication characteristics and supplementary concept records. Qualifiers can be attached to descriptors to narrow the focus of a term (e.g.: drug therapy, diagnosis, etiology, surgery). For example, a deficiency of monoamine oxidase is retrieved by the Descriptor Monoamine Oxidase combined with the Qualifier */deficiency* (NLM, 2012f). The field *Allowable Qualifiers* mentions the rules restricting the attachment of qualifiers for each term. Check tags are a special class of MeSH descriptors representing frequently occurring concepts related to species, gender, human age, historical time periods and pregnancy (NLM, 2012d). They are called this way because, in the past, MeSH indexers used forms that already had these concepts preprinted and indexers only had to check the appropriate box. Publication characteristics (or types) describe the item being indexed instead of its topic. It has 3 main categories: publication components (e.g. English Abstract), publication formats (e.g. lectures, letter) and study characteristics (e.g. clinical trial, meta-analysis). The Supplementary Concept Records contain non-MeSH headings that can also be used in the indexing process (NLM, 2012d). In 2013, MeSH has 214,000 headings in the Supplementary Concept Records within a separate thesaurus (NLM, 2012b).

MeSH can be used in the indexing process of non-NLM databases like the ones used by health libraries and the National Guidelines Clearinghouse[36] (Hersh, 2002).

2.3.2 *Non-NLM thesauri*

In addition to Mesh there are other thesauri in the health area used to index documents.

CINAHL uses the CINAHL Subject Headings, which is based on MeSH and has additional domain-specific terms (Brenner and Mckinin, 1989). EMBASE[37], a European database of biomedical and pharmacological information, has a vocabulary called EMTREE[38] with features similar to those of MeSH.

Other common vocabularies in the health area include the Logical Observation Identifier Names and Codes (LOINC) that provides a universal code system for reporting laboratory and other clinical observations in electronic messages (Mcdonald et al., 2003), the HL7 vocabulary tables[39] that identify, organize and maintain coded vocabulary terms used in HL7 messages and the National Drug Code Directory[40].

The Consumer Health Vocabulary (CHV) is another vocabulary worth of notice, connecting “informal, common words and phrases about health to technical terms used by health care professionals” (NLM, 2012a). It is devel-

[36] http://www.guideline.gov
[37] http://www.embase.com
[38] http://www.embase.com/info/what-is-embase/emtree
[39] http://www.hl7.org/Special/committees/Vocab/vocab.htm
[40] http://www.fda.gov/cder/ndc

oped as an open source and collaborative initiative and can be used to improve IR systems, to help lay-people read and understand health-related information. CHV is part of the Unified Medical Language System (UMLS) since the 2011AA release and is also available from the CHV website[41] in file format or through an online browser[42]. The latest version of CHV has 57,819 health concepts and 158,519 English concept strings. A CHV concept is identified by the UMLS unique identifier and may be associated with several synonyms strings to express that concept. Each CHV concept is also associated with a CHV preferred name and a UMLS preferred name. The CHV preferred name is the string that best represents that concept for health consumers and is defined by the CHV. On the other hand, the UMLS preferred name is the preferred string for that concept as defined by the UMLS.

Another thesaurus is the multilingual glossary of technical and popular medical terms in nine European Languages, developed by the Heymans Institute to the European Commission. This glossary was developed to help overcome the terminology problems of health consumers when reading medication information leaflets and is composed by 1830 scientific and popular medical terms in nine languages (Stichele, 1995). Although having a small number of terms, this glossary is singular for its multilingual characteristics.

2.3.3 Unified Medical Language System

The Unified Medical Language System (UMLS) started at the NLM, in 1986, by the hands of its Director, Donald Lindberg, as a “long-term research and development project" (Lindberg et al., 1993). This project aimed at reducing barriers to the application of computers to the health area and more specifically to the effective retrieval of machine-readable information (Lindberg et al., 1993; Humphreys et al., 1998). Two of such barriers are the variety of ways to express the same concept in different vocabularies and the interchange of useful information between different systems. In fact, the medical informatics field is characterized by a large diversity of vocabularies developed for specific applications (e.g.: epidemiological systems, medical expert systems, indexing literature, codes for billing and procedures). The lack of a common language barred the interoperability of the applications that used these vocabularies and motivated the development of the UMLS.

The UMLS can be downloaded from the UMLS Terminology Services webpage[43]. Each UMLS release includes MetamorphoSys, required to install Knowledge Sources files, and to create, search and browse customized Metathesaurus subsets.

The UMLS consists of three knowledge-sources that can be used separately or together. One is the Metathesaurus that has more than one million biomedical concepts from over 100 sources (including MeSH), another is the Semantic Network with 135 broad categories and 54 relationships between categories and the last is the SPECIALIST Lexicon and Tools which has lexical information and programs for language processing (Kleinsorge and Willis, 2007). The uses of these knowledge-sources can be very diverse (e.g. information retrieval,

[41] http://www.consumerhealthvocab.org
[42] http://consumerhealthvocab.chpc.utah.edu/CHVwiki/
[43] https://uts.nlm.nih.gov/

natural language processing, automated indexing, thesauri construction, electronic health records and others).

Metathesaurus

The UMLS Metathesaurus is a multi-source, multi-purpose and multi-lingual thesaurus of health related concepts, their various terms and the relationships among them. It is called a Metathesaurus as it transcends the thesauri, vocabularies and classifications it covers (NLM, 2012e). The Metathesaurus is not a unified standard vocabulary (Humphreys and Schuyler, 1993) but it establishes conceptual linkages between its source vocabularies preserving their meanings, concept names and relationships.

In the Metathesaurus, synonymous terms are clustered into a concept with a unique identifier (CUI). Each term, identified by a unique identifier (LUI), is a normalized name and may have several strings (identified by SUI), which represent the terms' lexical variants in the source vocabularies. Each string is associated with one or more atoms (identified by AUI) that represent the concept name in the source. For example, as presented by Kleinsorge and Willis (2007), the Headache concept (C0018681) has the following structure:

```
headache (L0018681)
   headaches (S1459113)
      headaches (A1412439) -- BI
   Headache (S0046854)
      Headache (A2882187) -- SNOMED
      Headache (A0066000) -- MeSH
cranial pain (L1406212)
   Cranial Pain (S1680378)
      Cranial Pain (A1641293) -- MeSH
cephalgia head pain (L0290366)
   HEAD PAIN CEPHALGIA (S0375902)
      HEAD PAIN CEPHALGIA (A0418053) -- DxP
```

If a term has different meanings (e.g. cold) it is assigned the same LUI that stays associated with different CUI (e.g. cold temperature, common cold, cold sensation). The same can happen with strings and concepts.

Semantic Network

The Semantic Network is an upper-level ontology in the health field (Sherrilynne, 2005) composed of *Semantic Types*, which may be assigned to Metathesaurus' concepts and *Semantic Relationships* relating a set of Semantic Types. The network has Semantic Types as nodes and Relationships as links. It is provided in the format of a relational table format and in a unit record format (NLM, 2009).

The current scope of the UMLS semantic types is very broad, allowing the semantic categorization of a wide range of terminology (NLM, 2009). Each concept of the Metathesaurus is assigned the most specific semantic type available in the hierarchy. Instead of adding semantic types to the Network to encompass an object in the most appropriate categories, concepts that don't belong in a granularity level, must be associated with a Type of an upper level. For example, the Semantic Type *Manufactured Object* has two child nodes: *Medical Device* and *Research Device*. If an object is neither a medical device nor a research devices, it is simply assigned to the more general type *Manufactured Object* (NLM, 2009).

Semantic relationships may be hierarchical or associative. The *isa* link is the primary link in the Network and is the one that establishes the hierarchy of types and relations (e.g.: animal *isa* organism). The associative relationships are grouped into five major categories (which are also relations): *physically related to*, *spatially related to*, *temporally related to*, *functionally related to* and *conceptually related to*. Whenever possible, relations are defined between the highest-level semantic types and, generally, are inherited by all the children of those types. However, if the inheritance does not make sense, it can be blocked to one or all children of the semantic type. For example, "*conceptual part of* links *Body System* and *Fully Formed Anatomical Structure*, but it should not link *Body System* to all the children of *Fully Formed Anatomical Structure*, such as *Cell* or *Tissue*" (NLM, 2009).

SPECIALIST Lexicon and Tools

The SPECIALIST Lexicon and Tools includes a general English lexicon of common words that includes health terms and was developed to provide information to the SPECIALIST Natural Language Processing (NLP) System (NLM, 2009). Moreover, it includes tools that are programs that process terms. The lexicon and the tools pre-process terms before their introduction in the Metathesaurus and are useful to NLP applications.

The lexicon's entries record the syntactic (how the words are put together), morphological (inflection, derivation and compounding) and orthographic (spelling) information of each term. Lexical items may be composed by more than one word if the multi-word appears in English or medical dictionaries or in medical thesauri.

Each unit lexical record has attributes, or *slots*, and values, or *fillers*, and is delimited by braces ({}). The unit lexical records for "anaesthetic"[44], one as a noun entry and the other as an adjective entry, are presented next:

```
{
base=anesthetic
spelling_variant=anaesthetic
entry=E0354094
cat=noun
variants=reg
variants=uncount
}
```

Every lexical record has the attribute *base* that indicates the base form of the term. Optionally, it may also has one or more spelling variants expressed by slots *spelling_variant*. The *entry* contains the unique identifier (EUI) of the record. The slot *cat* exists in every record and indicates the syntactic category of the entry (e. g.: noun, adjective, verb). The *variants* slot indicates the inflectional morphology of the entry. In the example given, these slots indicate that "anesthetic" is a count noun, which undergoes regular English formation ("anaesthetics"). In the adjective lexical record:

```
{
base=anesthetic
spelling_variant=anaesthetic
entry=E0330019
cat=adj
```

[44] Extracted from NLM (2009)

```
variants=inv
position=attrib(3)
position=pred stative
}
```

the *variants* slot has the filler *inv* that indicates that the adjective “anesthetic" doesn't form a comparative or superlative. The first *position* slot indicates that the adjective is attributive and appears after color adjectives in the normal adjective order. The second *position* slot indicates this adjective can appear in predicate position.

Other slots indicate the complementation (e.g.: in verbs if it is a intransitive or transitive verb), derivation (e. g.: adjective/noun – red/redness) and spelling variants (e.g.: British-American variants – centre/center) of each entry. For more detailed information see (NLM, 2009).

Lexical entries are independent of semantics, representing only a spelling-category pair. If different senses have the same spellings and syntactic category, they are represented by a single lexical entry in the Lexicon.

The Lexical Tools include three Java programs: a lexical variant generator (lvg), a word index generator (Wordind) and a normalizer, abbreviated as NORM. These are designed to help dealing with the high degree of variability in natural language words (e. g.: treat, treats, treated, treating) and even in the order of words in “multi-word" terms.

The lvg program performs lexical transformations of input words. It consists of several flow components that can be combined. For example, the flow `i` simply generates inflectional variants and the flow `l:i` generates the same inflectional variants but in lowercase.

The WordInd breaks strings into words and produces the Metathesaurus word index. This lexical program should be used before searching the word index to assure the congruence between the words to be looked up and the word index. The program outputs one line for each word found in the input string. For example, for the input string *Heart Disease, Acute*, three lines are returned for the three words: *heart*, *disease* and *acute*. The output words are always presented in lowercase.

The NORM program generates normalized strings that are used in the normalized string index (MRXNS). This program must be used before MRXNS can be accessed and it is a selection of lvg transformations (in fact, it is the lvg program with the `N` pre-selected flow option). The normalization process involves stripping possessives (e.g.: *Hodgkin's diseases, NOS – Hodgkin diseases, NOS*), removing stop words (e.g.: *Hodgkin diseases, NOS – Hodgkin diseases,*), lower-casing each words (e.g.: *Hodgkin diseases, – hodgkin diseases,*), replacing punctuation with spaces (e.g.: *hodgkin diseases, – hodgkin diseases*), breaking a string into its constituent words/uninflecting (e.g.: *hodgkin diseases – hodgkin disease*) and sorting the words in alphabetic order (e.g.: *hodgkin disease – disease hodgkin*).

2.3.4 Ontologies

Ontology was initially defined as the set of primitive entities that describes and models a specific knowledge domain and should reflect its underlying reality (Liu and Özsu, 2009). In computer science, ontology is the organization of

concepts in domains, exhibiting internal consistency, acyclic polyhierarchies and computable semantics (Sherrilynne, 2005).

Health ontologies aim to study classes of health significant entities such as substances (e.g.: mitral valve), qualities (e.g.: diameter of the left ventricle) and processes (e.g.: blood circulation) (Sherrilynne, 2005). Ontologies are increasingly playing an important role in medical informatics research (Musen, 2002). Some of the main health ontologies and two other general ontologies will be presented next.

One of the ontologies has already been described. In fact, the Semantic Network of the UMLS acts as a basic, high-level ontology for the health domain (Mccray, 2003).

GALEN (Generalised Architecture for Languages, Encyclopedias, and nomenclatures in medicine) was the name of a European Union project (1992-1999) that illustrated how medical concepts could be represented as a formal ontology and how this could be used in practical applications (Rector and Nowlan, 1994). One of this project's core features is the Common Reference Model, an ontology that aims to represent "all and only sensible medical concepts", whose access is made through OpenGALEN[45]. GALEN provides the blocks required for describing terminologies and a mechanism for combining concepts. For example, it has explicit representations for *adenocyte* and for *thyroid gland* and instead of having one for *adenocyte of thyroid gland*, it has an indication that these concepts can be combined. GALEN has a hierarchy of categories and a "rich hierarchy of associative relationships used to define complex structures" (Sherrilynne, 2005).

SNOMED CT, the Systematized Nomenclature of Medicine Clinical Terms, is a biomedical terminology developed in native description logic formalism (Sherrilynne, 2005). It has a good concept coverage organized in hierarchies. Each concept is described by several characteristics such as unique identifiers, parent(s), name(s) and role(s) or semantic relation(s). It is freely available as part of the UMLS.

Cyc[46] is a general ontology that is built around a core of more than 1 million hand-coded assertions that capture "common sense" knowledge. Groups of assertions that share a common set of assumptions (e.g.: domain, level of detail, time interval) are called *microtheories*. OpenCyc is the upper level and publicly available part of this ontology (Sherrilynne, 2005).

WordNet is a general English resource, freely and publicly available, where nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. This resource may be used in conjunction with other concept representations like the UMLS, that don't include all word-level synonyms and permutations in health concepts, providing another component to medical concept representation and retrieval. This is important because, in WordNet, health concepts may not be related to the health domain. For example, the concept *vasoconstriction* is only related to *constriction*, emphasizing the physical mechanism rather than a pathology (Sherrilynne, 2005).

Sherrilynne (2005) analyze the representation of blood in several systems, showing the differences among them and the richness of ontologies when com-

[45] http://www.opengalen.org
[46] http://www.cyc.com/

pared to taxonomies. The blood concept is interesting because it has two different superordinates: tissue and body substance. This is highlighted by the differing representations that raise issues of compatibility among ontologies. The richness of ontologies is also emphasized by the additional knowledge about blood through concept's properties and through the associative relations to other concepts.

2.4 RESEARCH OVERVIEW

The IR field is increasingly giving more attention to the health domain. Not only have major IR companies given attention to this particular field but the number of specific search engines and publications has also been growing. Globally, HIR research has been dedicated to find more efficient ways to adjust queries to user needs, to index health data, to present search results, to rank results (e.g. based on readability, content) and also to analyze information needs and behaviors. Next, we describe the most recent research directions in HIR, organized by high-level research topics.

2.4.1 Health Information Seeking

The increasing use of the Web to search for health information stimulates studies regarding information needs and behaviors in HIR. Usually, research in this area involves not only computer science but also areas like information science, psychology and medicine. These studies are useful to improve the retrieval process and contribute to understand the Internet's influence on health behaviors and also on the health care system. Some works are focused on consumers, some on health professionals and others on global aspects of the retrieval process, done either by consumers or by health professionals.

Consumers

In the consumer arena, a large number of studies are dedicated to characterize consumers' health information needs and their behavior. These studies are usually based on surveys and log analysis. The Pew Internet & American Life Project[47], that explores the impact of the Internet in several aspects of society, has published several reports related to online health search (Fox and Duggan, 2013; Fox, 2006; Fox and Rainie, 2000). Rains (2007) discusses the perceptions of the Web's use to seek health information gathered from the Health Information National Trends Survey's findings. These studies are centered in the American citizen and there are others that characterize the behaviors of others countries' citizens, like the works focusing on Portuguese (Santana and Pereira, 2007) and Greek citizens (Halkias et al., 2007).

There are also studies that analyze the use patterns of specific systems. For example, Scott-Wright et al. (2006) study the stability of user queries overtime using logs of the MEDLINEplus. Others explore student's search process and outcome in Medline to write an essay for a class (Huuskonen and Vakkari, 2008) and others investigate the influence of technical knowledge and cognitive abilities on health information seeking (Sharit et al., 2008). Yoo and

[47] http://www.pewinternet.org

Robbins (2008) on a more theoretical approach, attempt to explain how and why middle-aged women use health web sites based on two theories: the uses and gratifications approach from mass communication research and the theory of planned behavior from social psychology. An earlier and more comprehensive study is the one from Cline and Haynes (2001) covering aspects like potential benefits of health information seeking on the Internet, information quality problems, criteria for evaluating online health information and problems related to design features such as disorganization and technical language.

Professionals

There are also studies that seek to understand the information needs of health professionals. Revere et al. (2007) do a literature review on these needs, trying to answer the following questions: "What are the information needs of public health professionals?", "In what ways are those needs being met?", "What are the barriers to meeting those needs?" and "What is the role of the Internet in meeting information needs?". González-González et al. (2007) analyze Spanish primary care physicians' information needs that arise in office practice and their information seeking patterns to satisfy these needs (in and after the consultations). Twose et al. (2008), with the same goals but a different approach, include the combination of usage statistics from a web portal allowing the access to a library's electronic resources, self-report and observational data collected during an offered course. Other works are more dedicated to the study of information seeking behaviors, such as the one from Hemminger et al. (2007) that, through a census survey, analyze the information seeking behavior of academic scientists of basic science and medical science departments; and one that determines how good is Google to lead doctors to a correct diagnosis (Tang and Ng, 2006).

Still directed to professionals, other lines of research include the study of information retrieval systems' impact on professionals' performance answering clinical questions (Westbrook et al., 2005), the assessment of the effectiveness of information retrieval systems for professionals (Hersh and Hickam, 1998), modifications in information behavior after changes like the introduction of a clinical librarian service (Urquhart et al., 2007) and collaborative information seeking behavior in the context of medical care (Reddy and Spence, 2008).

Consumers and Professionals

There are also general studies that are neither directed to consumers or professionals. Some papers are more theoretical, globally studying health information seeking behavior like the one from Lambert and Loiselle (2007) that examine scientific literature from 1982 to 2006 on this subject to characterize health information seeking behavior, discuss its operationalizations, antecedents, and outcomes. Others explore specific aspects in health information seeking behavior like user navigation (Graham et al., 2006), education disparities (Lorence and Park, 2007) and changes in user needs (Adams and Blandford, 2005).

Other studies are more focused on web aspects. Spink et al. (2004) report findings from an analysis of health queries to different web search engines,

providing insights into health querying and suggesting implications of the use of web search engines for health information seeking. Eysenbach and Kohler (2003) study the prevalence of health-related searches on the Web, based on the proportion of pages on the web cointaining the search string and the word *health*.

2.4.2 *Indexing*

Indexing in HIR is a research area that can largely benefit from the diversity and quantity of the health concept representations described in the previous section. In fact, a significant number of papers about this topic use at least one of these representations. Douyere et al. (2004) adapt the MeSH thesaurus to the broader field of Internet health resources, using it with the Dublin Core metadata format to catalogue and index French health resources. Hliaoutakis et al. (2006b) combine MeSH with a well-established method for extraction of domain terms in the development of an automatic term extraction method for indexing large medical collections such as MEDLINE. Houston et al. (2000) explore the use of concept spaces, that is, automatically generated thesauri, where concepts are represented as nodes and relationships as weighted links, in HIR. They evaluate and compare the use of terms suggested by MeSH, UMLS Metathesaurus and the automatic generated thesauri with document collection's terms. No statistically significant differences among the thesauri were found and there was almost no overlap of relevant terms suggested by different thesauri suggesting that recall could be significantly improved using a combined thesauri approach.

There are also papers dedicated to the development of health information structures. Zeng and Crowell (2006) describe the development of computerized methods to mark up Web content. Zou et al. (2007) also present research on segmenting and labeling HTML medical journal articles through a hidden Markov model approach. Kipp (2007) suggests the use of social bookmarking, such as the tags from CiteULike, as an additional health concept representation and a way to discover documents not yet indexed in on-line databases.

2.4.3 *Retrieval*

Health Information Characteristics

The analysis of health information quality (e.g. accuracy, timeliness, accessibility) and its adjustment to the user (e.g. readability) are popular lines of research in HIR. The identification of features related to these two factors may be used to rank results.

Berland et al. (2001) evaluated the accessibility, quality and readability of health information on breast cancer, depression, obesity and childhood asthma available in English and Spanish. The accessibility of search engines was assessed using a structured search experiment, the quality of the content was evaluated by physicians using structured implicit review and the reading grade level was assessed using the Fry Readability method. They found that coverage of key information is poor but the accuracy is generally good and that high reading levels are necessary to comprehend health contents.

Another work proposes a model of text readability to allow the adjustment of ranking to experts and non-experts health users (Yan et al., 2006). The

model takes into account two factors: how the domain-specific concepts affect document readability and how it affects the document's textual genres. They propose three readability formulas that are also applied in HIR and compared them to four traditional readability measures. Results show improvements in terms of correlation with users' readability ratings.

Miller et al. (2007) also state that traditional readability formulas are not targeted to specific domains like health as they ignore the use of specialized vocabulary. In their paper they propose a naïve Bayes classifier for three levels of health terminology specificity (consumer, health learner, health professional) created with the lexicon of a medical corpus. This classifier attained an accuracy of 96% and was applied to consumer health web pages. Only 4% of pages were classified as consumer ones, while all the others were included at the professional level. Miller was the second author of a recently published paper (Leroy et al., 2008) that also compares the naïve Bayes classifier with readability formulas and the readability assessment of an expert and a consumer. Results showed that the classifier indicated that documents were harder to read than what readability formulas suggest. A previous paper from Leroy and other authors (Leroy et al., 2006) compared four types of documents: easy and difficult WebMD documents, patient blogs, and patient educational material. They found that it is possible to simplify documents based on terminology in addition to sentence structure. However, this can still be insufficient for difficult documents.

Keselman et al. (2006) evaluate the male's and female's familiarity with terms associated with male-specific, female-specific and gender-neutral health topics. It was found that males were more familiar with neutral and male-specific topics and in females no significant effect was found. In face of these results, the tailoring of health readability formulas to target populations is also discussed. Rosemblat et al. (2006) have explored the relevance of readability predictors in the consumer health domain, based on expert judgment to characterize ratings' patterns across the various predictors. They concluded that the development of health readability tools might require the modification of existing measures (e.g. including health-related vocabulary) and the addition of new predictive features.

Zeng is the author of several papers related to health consumer terminology. In one, Zeng et al. (2002) study the characteristics of consumer terminology used in HIR through the log analysis of two consumer web sites and patients' interviews. They concluded there are significant mismatches between consumer and information source terminologies. Zeng et al. (2005b) created a method to measure the familiarity of medical terms and a predictive model for familiarity, based on term occurrence in text corpora and reader's demographics. In the third paper, Zeng et al. (2005a) develop a systematic methodology using text analysis and human review to assign *consumer-friendly* names to UMLS's concepts. The evaluation of this method was done applying a questionnaire to consumers and the results suggested this methodology is useful in the development of consumer health vocabularies. More recently, Zeng-Treitler et al. (2008) present a method to predict consumer familiarity with a document using contextual information. The method is based on a network of terms and context relationships where some of the nodes, named root terms, have a related evaluation. A term may appear in contexts like the query session, sentence, paragraph or document. The method was applied to MedlinePlus log

files and showed better results than the syllable count, frequency count, and log normalized frequency count.

Query Expansion

Query terms are often related to terms used to index documents but are not index terms. This motivates the development of techniques of query expansion. These methods are used to improve precision in search results by suggesting alternative or additional query terms.

Zeng et al. (2006) develop a tool to assist people in health-related query construction. The suggested terms were selected based on their semantic distance to the original query, calculated through co-occurrences in medical literature and log data and also through semantic relations in medical vocabularies. Authors concluded that semantic-distance-based query recommendations can help consumers with query formulation during HIR.

Fattahi et al. (2008) propose a query expansion method based on the non-topical terms that exist in web documents. Authors define topical terms (TT) as terms that represent the subject content of documents (e.g.: breast cancer); non-topical terms (NTT) as terms that occur before or after topical terms to represent a specific aspect of the subject (e.g. 'about' in 'about breast cancer') and semi-topical terms (STT) as terms that normally do not occur alone, being used in conjunction with topical terms to narrow or further specify the subject (e.g.: 'risk of' in 'risk of breast cancer'). The query expansion method is based on the use of NTT and STT in conjunction with TT.

Ide et al. (2007) describe the algorithms used in a search engine with query expansion and probabilistic ranking, evaluated using data and standard evaluation methods from the 2003 and 2006 TREC Genomics Track.

Ranking

The ranking algorithms used in IR systems, to determine the position of each result in the search results list, can also be used to enhance the initial precision of the systems. Price and Hersh (1999) present a system that ranks results according to the estimated quality of the page health contents. In the work of Price and Delcambre (2005), the approach to improve the ranking of results is different. The authors model queries as relationships between concepts and try to match these relations with the ones existing between documents. Anagnostopoulos and Maglogiannis (2007) use a different approach, which is also typical in the development of ranking algorithms, and analyze users' browsing behavior.

IR Models

The definition of specialized IR models is also common in HIR papers. The use of semantic information is typical of the most recent proposed models. This happens in the model proposed by Price et al. (2006) and Price et al. (2007). In this model, the authors describe the content of documents in domain-specific collections using semantic components, that is, "segments of text about a particular aspect of the main topic of the document that may not correspond to structural elements in the document" (Price et al., 2007)), as a complement

to full text and keyword indexing. In the first paper, Price et al. (2006) introduce the model, present the results of its application to the representation of clinical questions in the medical domain, and present ways to use the model in retrieval. In the second paper, Price et al. (2007) present experimental evidence that the model enhances the retrieval of domain-specific documents in response to realistic users' queries.

Hliaoutakis et al. (2006a) present an IR model, implemented in MedSearch, that discovers similarities between documents containing semantically similar, but not necessarily lexically similar, terms. Dung and Kameyama (2007) present a methodology to build and enhance ontologies in the health domain through the extraction of semantic elements. The information extracted from Web documents is then summarized, indexed and used in retrieval by the IR system.

In a more recent paper, Hung et al. (2008) present an information seeking model to represent human search expertise that may allow the development of an intelligent search agent that generates adaptive search strategies based on human search expertise. The model described is hierarchical and multi-level, each level representing a problem space traversed during the search process and a layer of knowledge required to a successful search.

2.4.4 Evaluation

The evaluation of an IR system is not a simple task, frequently involving human intervention, usually in the judgment of relevance or in user studies. An alternative approach involves the use test collections, like the ones used in TREC[48] with less human intervention. The development of such collections is the target of some papers. For example, Hersh et al. (2006) have developed a test collection to assess visual and textual methods in biomedical image retrieval.

An example of a user-centered study is the one presented by Kushniruk et al. (2002) where a comparative usability evaluation considers an automated text summarization system and three search engines. The evaluation involved audio and video recording of subject interacting with the interfaces. Another paper in which evaluation is done with human intervention is one from Tang et al. (2006). Here, human assessors judged relevance according to a previously defined scheme. The goal was to compare the performance of health-specific and depression-specific search engines with Google on both relevance of results and quality of advice.

2.4.5 User Interfaces and Visualization

The development of user interfaces and the aspects of information visualization in IR systems is another area with a significant quantity of research work. In the health area, with developments in concept representation models, it also became popular the use of this information to improve the way results are presented to the user.

Stuckenschmidt et al. (2004) present a system that implements a concept-based visualization of the results that, according to a user study, is less suitable for searching specific information and more suitable to the exploration of mostly unknown data. Douyere et al. (2004) develop a terminology based on

[48] http://trec.nist.gov

the MeSH thesaurus and metadata elements. This terminology is then used in several tasks like the visualization and navigation through the concept hierarchies. Stapley and Benoit (2000) have built a system for retrieving and visualizing co-occurrences of gene names in Medline abstracts. From the co-occurrence data is built a graph where nodes are genes and edge lengths are a function of the co-occurrence of the two genes in the literature.

2.5 CONCLUSION

In this chapter we describe the background of health information retrieval, the subfield of information retrieval to which our work is dedicated. We begin describing the classical classifications usually applied to health information and enumerate the main existing resources in each category.

We then explain that health information may be represented through terminologies, thesauri and ontologies and differentiate these three types of representation systems. For each type of system, we describe their main existing instances. Of these, the Multilingual Glossary of technical and popular medical terms, the CHV thesaurus and the UMLS are used in subsequent phases of this dissertation. The multilingual glossary is used in the setup of the experiment described in Chapter 4 and the other two resources are used in the implementation of the suggestion system described in Chapter 13 and in one of the methods for health query identification and classification proposed in Chapter 16.

In the end of the chapter we have an overview of the research being done in this specific field. Areas of particular interest for this work are covered in more detail in subsequent chapters.

In the following chapter we introduce the notion of context and review the existing literature on the use of context in the general area of information retrieval and in the specific area of health information retrieval.

CONTEXT IN HEALTH INFORMATION RETRIEVAL

3.1 INTRODUCTION

Context is one of the most abused terms in IR, being associated to a large range of ideas (Finkelstein et al., 2002). Brézillon (1999) enumerates twelve different definitions from several authors and the lack of consensus is evident. To clarify some aspects regarding *context*, we do a literature review on the definition of context and context taxonomies in Section 3.2.

Afterwards, we describe the growing importance of context in IR and, through an existing taxonomy for context features and a proposed taxonomy for context uses, we analyze a set of IR papers using context to get an overview of what and where are context features being used in this domain.

Finally, we describe the main research being done in the application of context to HIR. Using the same taxonomies mentioned above, we start by classifying HIR research works that use context according to the used features and to the stage of the retrieval process into which they were incorporated. Further, we also identify the specific features used in each context category and stage of the process. We then describe these works with more detail in Section 3.4.2.

3.2 CONTEXT

As Dervin (1997) says, "context has the potential of being virtually anything ... [it is] a kind of container in which the phenomenon resides". The concept crosses several areas of knowledge from cognitive sciences to engineering. This section reports on definitions in domains related to IR and does not intend to do a thorough review of definitions in other areas. The work of Brézillon (1999) presents a more thorough review of context's definitions in five areas connected to artificial intelligence.

In the literature, some authors have gone further in the characterization of context, defining contextual taxonomies. These structures facilitate the understanding and exploration of context. Some of the main context taxonomies are also described in this section.

3.2.1 Context Definition

According to Dourish (2004), *context* may be defined in two perspectives: as a representational problem or as an interactional problem. In the first per-

spective, it is viewed as a form of information that is delineable, stable and independent of the activity. It consists of implicit attributes that describe the user and the environment in which information activities occur. The second perspective sees context as arising from the activity, from which it can't be separated.

Dey and Abowd (2000) also do an extensive review on context's definitions. They propose their own definition that encompasses other authors' definitions. Context is: "any information that can be used to characterize the situation of entities (e.g. a person, a place or an object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves". This definition matches the first perspective of Dourish. Dey and Abowd are not the only authors defining context in line of the first perspective of Dourish. Göker and Myrhaug (2002) present a short and comprehensive definition: "description of aspects of a situation", similar to the one from Dey and Abowd (2000). Marchionini (1997) had already defined it as a "setting" that has "physical and conceptual/social components, including whether the task is done in collaboration or alone and the information seeker's physical and psychological states". The first definition proposed by Johnson (2003) also equals context to "situation".

The second definition of Johnson (2003) goes beyond the enumeration of factors to the specification of the active ingredients in context, noting that they have predictable effects on processes. In this view, context is defined as a relation between the specific ingredients and the processes, which is closer to the second perspective of Dourish. Similarly, Winograd (2001) says "something is context because of the way it is used in interpretation". Sato (2003) defines context as "a pattern of behavior or relations among variables that [...] potentially affect user behavior and system performance". Ingwersen and Järvelin (2005) say "actors and other components function as context to one another in the interaction processes. There are social, organizational, cultural as well as systemic contexts, which evolve over time".

3.2.2 Context Taxonomies

Ingwersen and Järvelin (2005) developed the most comprehensive taxonomy for context features in the IR domain. The first version of their nested model of contexts has six dimensions. The first and second dimensions represent the intra and inter object contexts and constitute the central component of the cognitive Information Seeking & Retrieval (IS&R) framework, also proposed by the authors. The other four dimensions are: the interaction (session) context; the context provided by the remaining components of the framework; the societal infrastructures and, across the stratification, the historic context of all actors' experience. Later, and by the same authors, the social/organizational/cultural context dimension was divided in two sub-dimensions: an individual and a collective one (Ingwersen, 2006).

This model may be centered on the information space, on the cognitive actor (e.g.: searcher), on the interface, on the information technology (engines, logics, algorithms) or on the social/organizational/cultural context. This choice will affect the nature of the interaction context and the context of the individual and collective dimensions.

In this classification we decided to center the model on the information space as can be seen in Figure 3.1. The cognitive actor was another potential alternative but we felt the specificities of the information space in the health domain would be better described if placed in the first two dimensions of the model. Searcher's context is therefore included in the fourth dimension. We also felt that the cognitive actor in the core would result in a more ambiguous model. In fact, depending on the use given to context features, the cognitive actor could be the searcher or another actor (e.g.: person contributing to the indexing process).

With the information space as the center of this model, the intra-object structures refer to context obtained from each document where, for example, images are contextual to a surrounding text and paragraphs act as context for their own lines and words. Inter-object contexts are concerned with the properties of documents, like references, citations, outlinks and inlinks, that give and take context from other objects. The interaction dimension is about the search/authoring evidence like eye and mouse movements, descriptions of the work tasks and explicit relevance feedback. The component-dependent individual context, the fourth dimension in Figure 3.1, includes information about the elements that are not in the core of the model. In this case, it can be about the engine logic/algorithms, the interface functionality or the searcher. The component-dependent collective context, fifth dimension in Figure 3.1, includes all the local socio-organizational structures and conditions like domain vocabulary, searchers' work tasks perceptions and their implicit relevance feedback behavior. The sixth dimension is about infrastructural context, including aspects like network type and speed. Finally, historic context is a temporal form of context including actor's previous experience.

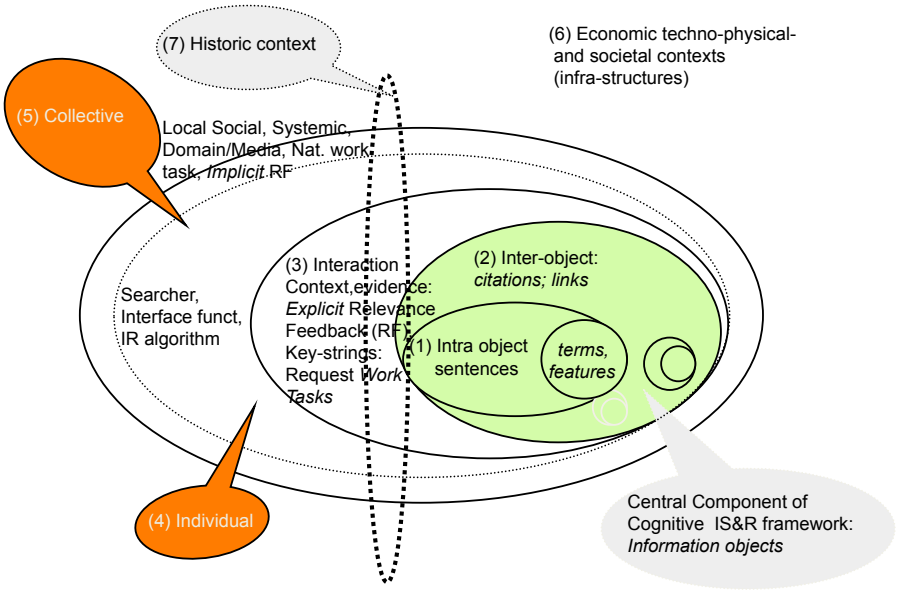

Figure 3.1: Ingwersen and Järvelin's nested model of contexts with the information space as the central component.

Dey and Abowd (2000) propose a classification of context information based on the entities in which the context is assessed and on categories of context (Figure 3.2). They define three entities: places like regions of geograph-

ical space such as rooms or offices, people including individual or groups, co-located or distributed, and things (e.g. physical objects or software components and artifacts like a computer file). Primary and secondary context characterize these entities. Primary context types are: identity (ability to assign a unique identifier to an entity), location (all information that can be used to deduce position and spacial relations between entities such as position information, orientation, elevation, co-location, proximity, containment), status/activity (intrinsic characteristics of the entity that can be sensed like temperature, light or noise for a place; physiological factors, mood or the activity the person is involved in for people; state of files in a file system for things) and time (context information that helps characterize a situation like a time span indicating when is the information relevant). These context types may be used to infer secondary context like the address of a person by her identity. This work also proposes categories for uses of context: presentation of information/services to the user, execution of a service and tagging of context to information for later retrieval.

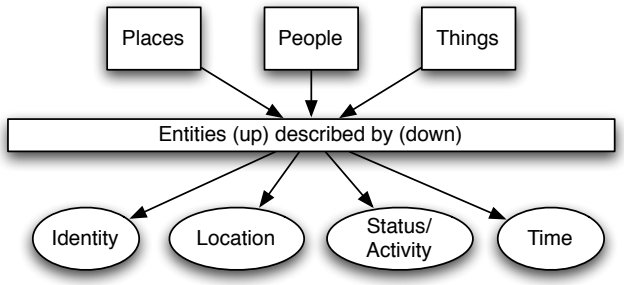

Figure 3.2: Dey and Abowd (2000) context taxonomy

Göker and Myrhaug (2002) present a context taxonomy in which context elements are divided into five main categories (Figure 3.3). The task category is about what the user is doing, his goals, tasks and activities. The social one refers to the social aspects of the user, such as information about friends and family or his role. Personal context aggregates mental and physical information about the user such as mood, expertise and disabilities. In the spatio-temporal category are included attributes like time and location and the environmental context is about user surroundings like things, light, people and information accessed by the user.

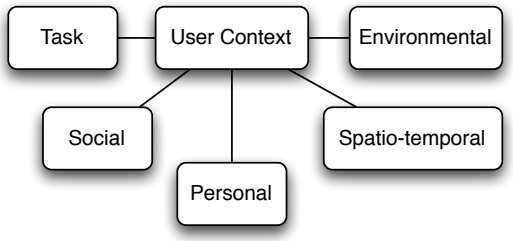

Figure 3.3: Göker and Myrhaug (2002) context taxonomy

Bricon-Souf and Newman (2007) propose a framework to analyze the use of context in health care applications. Their framework, shown in Figure 3.4, has three main axes to characterize context. The axis *purpose of use of context* presents the three types of context uses proposed by Dey and Abowd (2000). The second axis, *items for context representation*, identifies three main classes to split items of context into: people, environment and activities. The third axis, *organization of context features* proposes other ways to organize context features such as an hierarchical organization that draws from general to local aspects of context, an organization according to the internal and external dimension of context, an organization in accordance with the focus of the current activity and an organization according to the usefulness of context (relevant or non relevant for the current action).

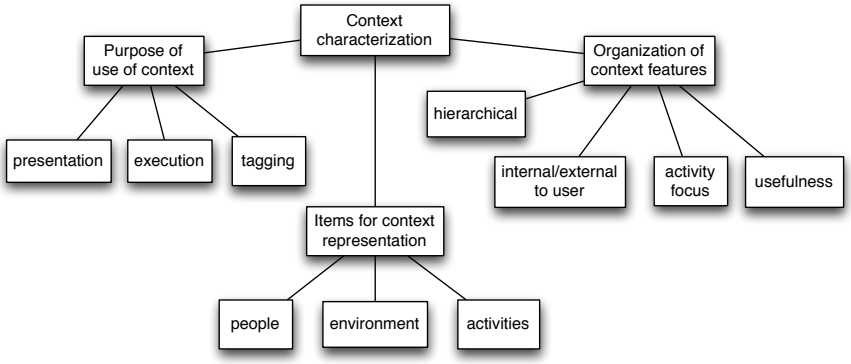

Figure 3.4: Bricon-Souf and Newman (2007) context model

Mansourian (2008) has also developed a taxonomy for the contextualization of web search (Figure 3.5). This taxonomy was built upon an inducted analysis of a set of interviews. Five categories were identified as the main contextual elements that affect web search. The *web user* axis is divided in feelings, thoughts and actions during the search. The features of the *search tool* (generalized search engine versus specialized database) and the *search topic* (work-related and everyday life searches) are two other axes of the taxonomy. The fourth axis, *search situation*, includes the place of search, type of search, immediacy of search and importance of search. The last axis is about the retrieved *information resources* and consists of searchability, level of provision and presentation format.

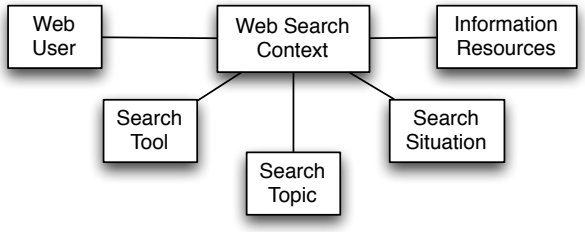

Figure 3.5: Mansourian (2008) context model

From all the reviewed taxonomies, only the Ingwersen and Järvelin (2005)

one has been made for IR. This is the most exhaustive taxonomy, even though it doesn't propose a classification for uses of context. Only the Dey and Abowd (2000) and Bricon-Souf and Newman (2007) taxonomies include this categorization. The Göker and Myrhaug (2002) taxonomy is a well-known taxonomy in the field of IR.

3.3 CONTEXT IN INFORMATION RETRIEVAL

Several authors agree that context, often ignored, might be used to improve the retrieval process (Ingwersen et al., 2005; Bierig and Göker, 2006). A contextualized IR system has the potential to learn and predict what information a searcher needs, learns how and when information should be displayed, shows how information relates to other information that has been seen and how it relates to other tasks the user was engaged in and decides who else should be informed about new information. Moreover, context should also be used in the evaluation of IR systems. Very recently, *contextual evaluation* has been defined as one of the six main challenges for the experimental evaluation of multilingual and multimedia information access and retrieval systems (Agosti et al., 2012). According to this report, contextual evaluation is the "integration of users, tasks, search applications and underlying information retrieval systems in a holistic perspective to ensure that the global impact of an information retrieval system on a user [...] can be assessed".

In IR, the interest to adapt the search process towards the user's needs and context has been increasing (Bierig and Göker, 2006). Contextual IR, Adaptive IR or Interactive IR are names usually given to a research area that combines search technologies and search context in order to provide the most appropriate answer for a user's information need (Allan et al., 2003). Several journals and conferences have given attention to this topic in the past few years. The Information Processing and Management journal has dedicated a special issue to context in IR (Cool, 2002), a special issue to adaptive IR (Jose et al., 2008), a special issue on Personalization and Recommendation in Information Access (Fernández-Luna et al., 2011) and a special issue on Human-Computer Information Retrieval (White et al., 2012). The Information Retrieval journal has also dedicated a special issue to Contextual Information Retrieval Systems (Crestani and Ruthven, 2007).

In 2004 and 2005, two workshops entitled Information Retrieval in Context (IRiX) were held in association with the ACM SIGIR Conference (Ingwersen et al., 2004, 2005). In 2006, a new set of biennial conferences entitled Information Interaction in Context (IIiX) has begun as a spin-off of these workshops (Borlund and Ingwersen, 2006). A workshop about Adaptive IR had two editions, one in 2006 in SPIRE'06 (Rijsbergen et al., 2006) and the other in 2008 in conjunction with IIiX 2008 (Joho et al., 2009). The Information Seeking in Context (ISIC) biennial conference (Macevičiūtė and Wilson, 2008) is a conference that started in 1996 and still occurs. In conjunction with the European Conference on Information Retrieval, two editions of a workshop dedicated to Contextual Information Access, Seeking and Retrieval Evaluation (CIRSE) were held (Doan et al., 2009, 2010). A series of workshops entitled Workshop on Context-awareness in Retrieval and Recommendation (CaRR) has also been occurring since 2011. The two first editions were held in

conjunction with the International Conference on Intelligent User Interfaces and its third edition occurred in conjunction with the Sixth ACM International Conference on Web Search and Data Mining (WSDM '13). An isolated workshop entitled *Context-Based Information Retrieval - CIR* was also held in the Sixth International and Interdisciplinary Conference on Modeling and Using Context (Doan et al., 2007).

The publications in these journals, conferences and workshops publications provide an overview of the relation between context and information retrieval, present case studies, propose evaluation and research methodologies, offer new ways for modeling context and provide frameworks for doing research in this area. The number of special issues and events dedicated to this topic show the attention that context is receiving from the IR research community.

3.3.1 Overview on the use of context in IR

To get an overview of the context features used in IR research we decided to adopt the Ingwersen and Järvelin (2005) nested model of contexts with the information space at the core as shown in Figure 3.1. The adequacy of this model to the IR domain and its comprehensiveness influenced our decision. The absence of a *uses of context* taxonomy in this model and the inadequacy of the taxonomies that do so (Dey and Abowd, 2000; Bricon-Souf and Newman, 2007) made us decide to propose our own taxonomy for the uses of context.

A taxonomy for uses of context in IR

We propose the taxonomy showed in Figure 3.6. In the existing taxonomies, only Dey and Abowd (2000) have proposed such an organization that was later included in the Bricon-Souf and Newman (2007)'s taxonomy. Their organization has three categories: presentation of information and services to a user, automatic execution of a service and tagging of context to information for later retrieval. With the Dey and Abowd (2000) categories in mind and with IR as this work's focus, the proposed top-level categories of uses of context in IR are: indexing & searching, query operations, ranking, user interface. These categories are the components of an IR system where context may be used.

The proximity of techniques used in the index construction and searching phases, stimulated their fusion into a single category. The *query operations* category is divided in two major subcategories: Relevance Feedback and Query expansion. However it can also include operations that don't fit in these subcategories like the generation of queries and their use to gather information from other systems. Relevance feedback and Query expansion are query refinement mechanisms that work fully automatically or with the help of the user.

Manning et al. (2008) consider relevance feedback a local method, that is, a method that adjusts a query "to the documents that initially appear to match the query". The main methods of this type are relevance feedback, pseudo relevance feedback and implicit relevance feedback. In relevance feedback the user marks returned documents as relevant or non-relevant and the system builds a better representation of the information need based on his feedback (Manning et al., 2008). Pseudo relevance feedback assumes the k ranked documents as

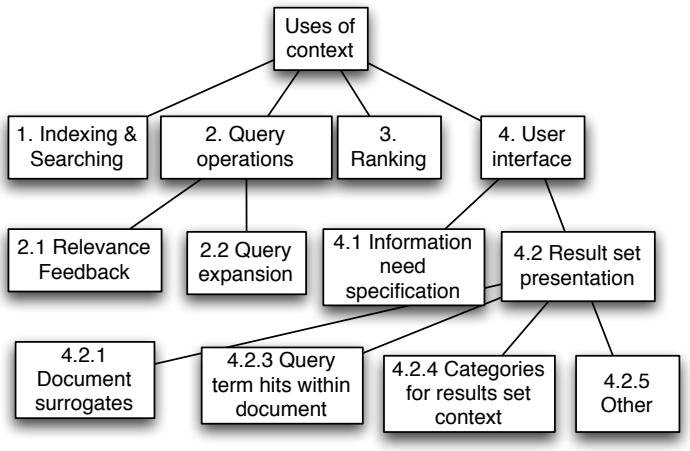

Figure 3.6: Proposed taxonomy for Uses of Context

relevant and implicit relevance feedback uses indirect sources of relevance like clicks on documents. Manning et al. (2008) consider query expansion a global method because it is independent of the query and the results retrieved with it. It can be based on collection-independent knowledge structures (Efthimiadis, 1996) like domain-specific thesauri or general-purpose thesauri (e.g.: WordNet), automatic thesauri generation and techniques like spelling correction.

In the IR process, the ranking phase is usually straight connected to the searching phase. Yet, we preferred to keep them as two distinct categories to help differentiate systems that have their own index and implement a retrieval model from systems that just reorder existing result sets based on some specific criteria.

The user interface category is divided in two subcategories: the interface associated with the specification of the user's information need and the presentation of the result set. The latter is also divided in document surrogates (e.g. snippet - short summary of the document), query term hits within document (e.g. keyword-in-context snippets), categories for results set context and other type of strategies.

The categories we propose map well to the categories defined by Dey and Abowd: the indexing & searching fits in the tagging category; the query operations may fit in the presentation of information and services (e.g. relevance feedback) or in the automatic execution of a service (e.g. implicit relevance feedback); the ranking fits in automatic execution of a service; and the user interface fits in the presentation of information and services. Despite the similarities between taxonomies, our taxonomy has a more operational perspective, being more focused on system development.

Literature analysis

The taxonomy proposed in the previous section is the basis for the analysis of a sample of contextual IR research papers. In 2009, we selected 24 papers from all the papers classified with the tag *context*[1] in *CiteULike*, a social web

[1] Available through: http://www.citeulike.org/tag/context

service for management of bibliographic references. From this list, the IR papers, published in 2008, that made use of context features were included in our sample. The sample is composed of papers from: Abel et al.; Ahn et al.; Bai and Nie; Chahine et al.; Deng et al.; Dhanapal; Fonseca et al.; Freund and Butterworth; Gyllstrom and Soules; Gyllstrom et al.; Inskip et al.; Kelly et al.; Kumaran and Allan; Li and Belkin; Martinez et al.; Martins et al.; Mylonas et al.; Pandey and Luxenburger; Rahurkar and Cucerzan; Ritchie et al.; Shtykh and Jin; Skov et al.; Ukkonen et al.; Zhang.

Each paper was examined according to: (1) the adopted context definition, (2) the exploited context taxonomy, (3) the context features used in the experience and (4) their specific use. Only four papers define context and only one present the underlying context taxonomy. In Figure 3.7 we present the proportion of papers using features from each layer of the Ingwersen and Järvelin (2005) nested model of contexts. The category used more often is the *collective* one, which was expectable because it is a very comprehensive one. Further analysis shows that from papers that use features from this dimension, a large majority (63%) uses topic context features such as TREC topics' descriptions, context documents, domain thesauri/ontologies and conceptual maps. Less used are the social (19%), environmental (13%) and spatio-temporal (6%) context features. As can be seen in Figure 3.7, the use of interaction and intra-object context features is also very popular in the existing IR literature. The interaction features range from desktop and web user behavior to users' submitted queries and descriptions of work tasks.

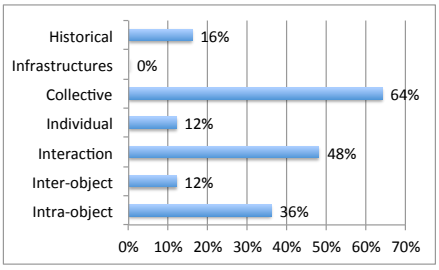

Figure 3.7: Context uses in Contextual IR literature

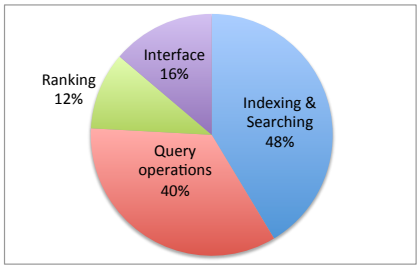

Figure 3.8: Context features used in Contextual IR literature

In Figure 3.8 we present a pie chart with the proportion of implemented context uses. With this chart it is possible to see that the *Indexing and searching* and the *Query operations* are the two stages where context features are used more often.

3.4 CONTEXT IN HIR

According to Lin and Fushman (2005), "the domain of clinical medicine is very well-suited for experiments in building richer models of the information seeking process". In fact, it's not difficult to predict how context features in the health domain can enrich HIR models. In this domain, the search process usually occurs in well-defined scenarios like treatment or diagnosis (Liu et al., 2007) and context may be extremely rich. Similarly to any visit to the doctor, where the patient doesn't just say "itch", but explains the context of the "itch" to the doctor, context is relevant to HIR. Possible context features that can be

used in this domain are the search scenario and its specificities (e.g.: treatment of a disease), the searcher's personal health record, the clinical case in hands and the searcher's knowledge in the health domain.

In this section we start by classifying the HIR research that uses context according to the used features and to the stages of the retrieval process where they are used. We then analyze in further detail the classified papers.

3.4.1 Classification of Research

To understand how context is being used in HIR, we gathered a set of research papers in this field using context features. To define a sample of papers, we considered all the documents classified with the tags context and health in CiteULike[2]. From this set we excluded papers not related with IR and papers in which IR was out of their main focus. For example, papers on Information Extraction and papers proposing readability formulas for health documents were excluded from this analysis. In addition, papers without an innovative contribution (e.g.: literature reviews or comparisons of IR systems) were also excluded. The final set included 27 papers.

To classify the research articles we use the Ingwersen and Järvelin's nested model of contexts with the information space at the core and the taxonomy for uses of context proposed in Section 3.3.1. The results of our analysis are presented in Figure 3.9 showing the distribution of papers by categories.

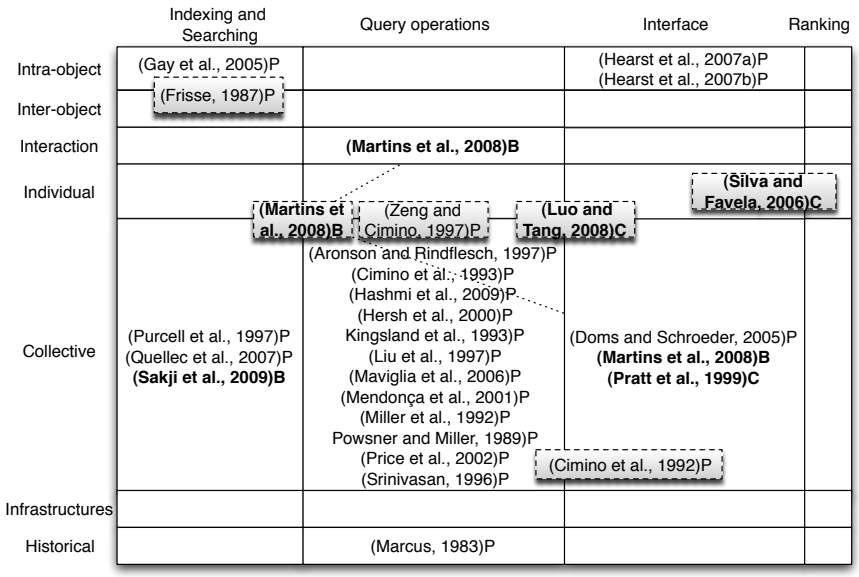

Figure 3.9: Papers classified based on the used context features and their specific use.

For convenience of representation, we switched the order of the interface and ranking categories. Each paper is represented by its bibliographic reference and a letter (P, C or B) representing the type of users to whom the system is targeted: professionals, consumers or both. When a paper crosses more than one category, its reference is represented in the categories intersection area. In some cases, it may also be connected with a dotted line to another cell of the

[2] Available through: http://www.citeulike.org/search/all?q=tag:context+tag:health

matrix. For example, the paper with reference (Martins et al., 2008) uses interaction, individual and collective context features in Indexing and Searching, Query operations and Interface stages.

Figure 3.9 shows that research is more intense on Query Operations, using mainly context features from the individual and collective dimensions. We were surprised with the weak use of the interaction context. This may be explained by the preference to use context features more related to the health domain. Typically, interaction context is more generic and not so health-related as individual and collective context features. On the other hand, we already expected to have a large number of papers using collective context features since this category is exhaustive, covering the characteristics of all the components from the cognitive framework that are not at the center of the model.

In Figure 3.9 we highlight, in bold, the papers focused on health consumers systems (letters C and B). It is clear that research is mostly dedicated to health professionals. The small number of consumer-dedicated research papers use interaction, individual and collective context features. To show which exact context features are used, we built Table 3.1 where we included the specific features in a structure similar to the one in Figure 3.9. In this table, EHR stands for Electronic Health Record and PHR for Personal Health Record. Health institutions hold the former and the latter is managed by the patient.

Table 3.1 shows that the health domain is very rich. The collective dimension gives an overview of the variety of structured information available, namely terminologies, thesauri and ontologies. Note that the same type of information may be included in different dimensions. For example, in systems for professional users, patient's clinical data incorporate the collective dimension of context and, in systems designed for patients, the use of clinical data about the searcher or patient is considered individual context.

3.4.2 Examination of Research

In this section, the papers included in Figure 3.9 are described in greater depth, and we enlarge the set to some theoretical papers. This analysis contributes to a better understanding of how context features are being acquired and used and how studies are conducted. Excluding the first section where theoretical papers are described, research works are presented according to the IR phase in which the context features are used. In each section, whenever possible, research is also separated according to context dimensions. Research using context features in more than one phase is split across categories.

Theoretical papers

In this section we analyze three papers that were excluded from the classification presented in Figure 3.9 due to their theoretical nature. Two of these papers propose a set of context features that can be used in the health domain, organized by categories. Cimino and Li (2003) use the following contextual parameters when the clinical system gives access to information resources: user type (nurse, physician, patient), patient age (newborn, infant, child, adolescent, young adult, middle aged and elderly), patient gender (male, female), concept of interest (medication, test result, organism that generated the user's request, mapped to concepts in the Medical Entities Dictionary), institution

Table 3.1: Context Features used in HIR.

	Indexing and Searching	Query Operations	Interface	Ranking
Intra-Object	Document contents and structure (e.g. abstract, conclusions, title, HTML structure).		Document images and captions.	
Inter-Object	Links between documents.			
Interaction		Browsing behavior.		
Individual	Authoring context.	Searcher's clinical data and user interest.	Searcher's clinical data and PHR.	PHR.
Collective	UMLS, domain categories, tasks, ontologies, taxonomies and patient data (age, sex and clinical context).	UMLS, MeSH, domain questions and terminologies, clinical practice guidelines, retrieval feedback, task context and patient data (clinical data, medical appointment reports, examination reports, EHR).	UMLS, MeSH, domain questions, Gene Ontology and patient data (clinical data, EHR).	
Infrastructures				
Historical		Search history.		

(used to determine which resources are available/preferred at a given institution). Lin and Fushman (2005) propose five context elements to better capture user's information needs: the work task, the search task, the process, the problem structure and the domain. The work task is composed by the user's broader activities like the evidence-based medicine therapy, diagnosis, etiology and prognosis. The search task is more detailed and specifies aspects like therapy selection, differential diagnosis, diagnostic methods selection, cause determination and patient outcome prediction. The process specifies how the information gathered is integrated in the work task. The EBM framework, for instance, has several tools to evaluate the confidence a health professional should have in the results and mentions the Strength of Recommendations Taxonomy (SORT) to do this. The problem structure is organized according to the PICO (Problem, Intervention, Comparison and Outcome), a well-known method in medicine. The domain is described using the UMLS.

The third paper reviews the integration of on-line bibliographic resources into patient care systems (Cimino, 1996), a very common functionality. In this paper, 12 systems were evaluated regarding the user question, the source of the answer and the composition of the retrieval strategy. All the systems use clinical data and the UMLS to generate the information requests. Some also use clinical data to directly compose a query, and others to identify topics of interest that will help in the query construction and on the source selection. Cimino concludes that interfaces need to be improved in systems that integrate several sources of information.

Indexing and Searching

As can be seen in Figure 3.9, the *Indexing and Searching* retrieval phase uses context features from the objects (Frisse, 1987; Gay et al., 2005) and from the collective dimensions (Martins et al., 2008; Quellec et al., 2007; Purcell et al., 1997; Sakji et al., 2009). In the latter, terminologies, thesauri and ontologies prevail (see Table 3.1). In this set of papers, only two (Frisse, 1987; Quellec et al., 2007) are dedicated to search and retrieval models, all the others focus on the indexing phase.

Object dimension

More than 20 years ago, Frisse (1987) used a prototype of a hypertextual medical therapeutics handbook to approach retrieval. He divided a therapeutics handbook into what he called individual hypertext cards, assigned a label to each card (the first sentence of that section in the book) and created links between the cards using the hierarchical structure of the book. He considered two approaches for indexing the documents: the small document approach and the graph traversal approach, the first emphasizes pattern matching and the second browsing. The uniqueness of Frisse's approach stands on using the usual `tf*idf` weighting with the average weight of all immediately linked nodes. He used hypertext as a belief network where the value of a card depends on its linked nodes.

A more recent study, done in the NLM, analyzed if, in term suggestion, there are benefits in using the article's full-text instead of using only the title and abstract (Gay et al., 2005). Authors used the structure of the document in their experiments and found advantages in using terms from captions and

from the following sections: *results*, *results and discussions*, *conclusion*, and *no header* sections. This last type of section was obtained through the division of articles into sections having no titles.

Individual and collective dimension

Martins et al. (2008) propose a semantically built index using natural language tools to analyze each document and link its terms to concepts of the UMLS ontology. They also mention the use of an authoring context that is gathered from the user and stored in the index. This is the only *individual* context feature used in Indexing and Searching. Their system uses ontologies to represent domain activities in the indexing and retrieval process.

Quellec et al. (2007) describe a system to retrieve medical cases based on images with contextual information. In the context of diabetic retinopathy patients, authors have defined context attributes to be stored with the images: age; sex; general clinical context (familiar, medical, surgical, ophthalmologic); circumstances, examination and diabetes context (diabetes type, diabetes duration, diabetes stability, treatments); eye symptoms before the angiography test (ophthalmologically symptomatic, ophthalmologically asymptomatic) and maculopathy. The authors used decision trees with images and context attributes as features and found promising results about the combination of numerical and contextual data in retrieval frameworks.

Purcell et al. (1997) propose context models for three types of medical publications: clinical research articles, case reports and review articles. These models outline the contexts that characterize each publication and provide the basis for understanding potentially ambiguous terms or phrases. In the indexing stage they mark up content assigning contexts (from the models) to sentences in documents, through specific tags. To evaluate their indexing scheme and, specifically, the ability of different people to reproduce the indexing for a set of documents, they conducted studies of inter-indexer consistency for each context model. They concluded that context models for clinical research articles and case reports could easily be learned and applied.

Another work that uses collective context features in the indexing and searching process is the one from Sakji et al. (2009). Working on the Catalogue and Index of the French-speaking Medical Sites (CISMeF), authors move from MeSH-only indexing towards the use of several health terminologies arguing that will lead to a better user adequacy.

Query Operations

According to our classification, the *query operations* phase is where context features are mostly used, mainly features from the collective dimension.

Interaction dimension

Only Martins et al. (2008) used interaction context features in the implemented relevance feedback approach. Besides the explicit evidence acquired when users browse the results, this system also uses implicit evidence provided by contextual information, although "implicit evidence" is not defined.

Individual and collective dimensions

A large number of research papers in this category describe systems that com-

pose, or help the user compose, queries and submit them to bibliographic retrieval systems. Of these, a large number are based on *infobuttons*, initially developed by Cimino et al. (1997), that take the form of context links from a clinical system to information resources related to the initial context. The differences between them reside essentially on the context they use to generate the query and on how they do it. Only two of the works use individual features (Zeng and Cimino, 1997; Luo and Tang, 2008), all the others exclusively use collective features.

A work of Zeng and Cimino (1997) is an application of the *infobuttons* to radiology results. To generate questions, this system uses individual context features like user interest and collective features like patient clinical information, generic questions' templates and the UMLS. The questions are then submitted either to the clinical system, the Medline or Web resources.

IMed is a system that helps users generate queries through an interactive questionnaire and the background use of medical knowledge (Luo and Tang, 2008). Based on a searcher's description of his condition, the system suggests the top-ranked symptoms and signs. The system then continues the iterative process asking questions and ranking answers not selected by the user according to their medical probabilities. In this process, the system uses diagnostic decision trees written by medical professionals. To help the searcher redefine his situation, the system also presents MeSH medical phrases related to his initial inserted condition. Evaluation with real medical case records and medical exam questions suggested the effectiveness of the system.

In a way similar to the not yet created *infobuttons*, a 20-year old paper already used the context of the clinical situation of a patient, more precisely a psychiatric consultation report, to gather a set of bibliographic references relevant to the reported case (Powsner and Miller, 1989). The PsychTopix system presents the user important topics extracted from the psychiatric report that, if selected by the user, are submitted to Medline. A similar approach (Miller et al., 1992) extracted UMLS Metathesaurus Main Concept terms and their synonyms from a medical text (e.g.: discharge summaries, lab results, radiology reports) in natural language. The system, Chartline, then suggests a set of Medline queries to the user. According to Hersh (2008a), these approaches were limited by the low specificity of the data in dictated reports. Cimino et al. (1992) developed an approach without the extraction component. In this work, using UMLS, authors have converted diagnosis and procedures coded in the International Classification of Diseases, Ninth Edition, Clinical Modifications (ICD9-CM) to MeSH terms. Between the translation and the submission to Medline, the user had to select a question from a set of generic queries generated by the system based on diseases and procedures selected in the patient admission form.

Later, Cimino et al. (1993) explain how they generated the generic questions using the UMLS Metathesaurus, Semantic Network and Information Sources Map (a fourth UMLS knowledge source that existed from 1991 to 1998). They do a syntactic and semantic analysis of a set of real queries. Syntactic analysis uses NLP techniques to identify the interrelated medical concepts in the query and librarians do semantic analysis to determine the type of query and the semantic relationships between the concepts. The system described in this paper works in two ways, either the user inserts a query and the system identifies the most relevant generic query or the user indicates patient data of

interest and selects one of the queries suggested by the system.

SmartQuery is another system that provides context-sensitive links from patient data to five online health resources (Price et al., 2002). These links appear next to lab tests and dictated reports. The system uses three different sources of query terms: MeSH terms translated from the ICD9 codes existing in the patient's diagnosis list; MeSH terms obtained from the lab results or dictated reports; and user inserted terms. From the terms suggested by the system, the user selects the terms he wants. A query is then constructed for each information source. Evaluation was done through a questionnaire applied to hospital residents after they have completed a set of three tasks in the system. Results showed that users liked the system, learned how to use it easily and were moderately satisfied when comparing the system with traditional methods.

A more recent work links pieces of computerized clinical practice guidelines to PubMed medical literature (Hashmi et al., 2009). This framework transforms the clinical practice guidelines into small chunks of knowledge components that are later enriched with context, semantics and metadata about the original practice guideline's content. To select the medical phrases that will integrate the final query, the knowledge components are analyzed, processed and filtered. In the end, the set of the selected medical phrases will be classified into one of four categories: diagnosis, etiology, prognosis and therapy. This label will also integrate the final query that will be submitted to PubMed. To evaluate the systems, domain specialists assessed the relevance of the retrieved literature (87% of the results were assessed as relevant) and the query classification given by the system (89% of the queries were assessed as correct).

KnowledgeLink is the name of another application that uses *infobuttons* in places where drug names appear in the EHR, providing links to web resources. The use of the infobuttons was assessed in a study conducted by Maviglia et al. (2006) in which they concluded "although used infrequently and for brief sessions, KnowledgeLink was positively received, answered most users' questions, and had a significant impact on medical decision making".

Mendonça et al. (2001) reviewed studies of clinicians' information needs and the role of terminologies on the integration of clinical systems with literature resources and also describe a model to accomplish this integration.

Coach is an NLM project designed to assist users during search in the medical domain (Kingsland et al., 1993). The system's main goal is to deal with the problem of boolean combinations with null results. Analyzing the *null retrieval* searches and using the UMLS knowledge sources, this system suggests new query terms to the user.

Srinivasan (1996) is the author of a study that examines three pseudo-relevance feedback (RF) methods on Medline: expansion on the MeSH query field alone, expansion on the free-text field alone and expansion on both the MeSH and the free-text fields. The study also intended to analyze the dependence on the availability of relevant documents for feedback information. The strategy that showed best results was the one adding only MeSH concepts. Moreover, authors concluded that the RF method is independent of the availability of relevant results. After Srinivasan's work on pseudo RF and its comparison to the use of thesauri for query expansion, Aronson and Rindflesch (1997) replicated their UMLS query expansion method in the test collection used by Srinivasan. They argued that their previous work had not been done

in a MeSH-indexed text and showed that it compares positively with pseudo RF.

In another study, Hersh et al. (2000) evaluated the use of UMLS Metathesaurus definitions and relationships in query expansion methods. Authors used the OHSUMED collection in their experiments. Expansion was done using term definitions, synonyms and hierarchical/related information. Their experiments showed bad aggregate retrieval performance in terms of recall and precision although improvements have been verified in individual queries.

Liu et al. (2007) describe a query expansion method that takes into account the scenario/task context (e.g. treatment or diagnosis of a disease) in retrieving medical free text. This expansion method uses UMLS to add terms relevant to the query's scenario. This method showed improvements when compared with a method that performs expansion with terms that are statistically correlated terms but not necessarily scenario-specific.

Historical dimension

CONIT is the only system exploring historical context in query operations. It assists the user when searching in heterogeneous retrieval systems (Marcus, 1983). The authors enhanced the system with search aids in three major areas: search history and reconstruction, automatic keyword/stem searching and individualized database searching. They concluded that the effectiveness of the CONIT at least approximates that of human intermediaries in some contexts.

Interface

Systems that use context to improve the *interface* explore intra-object structures and features from the individual and collective dimensions.

Intra-object dimension

Two of the analyzed papers use intra-object context features. In one of the papers, Hearst et al. (2007a) present a characteristic of the BioText Search Engine that allows browsing and searching of article's figures and captions. This BioText Engine's feature is based on the fact that the elements that get more attention in a scientific paper are the title, abstract, figures and captions. In the other work, Hearst et al. (2007b) support the search in the captions of the article's figures and returns the article, its figures, captions and also other related figures. Participants in a pilot study reported positive reactions to this idea.

Individual dimension

The iMed system (Luo and Tang, 2008), introduced above, helps the users construct a query through an interactive questionnaire and a personalized interface. Based on the patient clinical information and medical knowledge, the iMed search advisor makes suggestions with alternative symptoms and signs, alternative answers to the questions and a set of related medical phrases to help refine the description of the information need.

Silva and Favela (2006) use data from PHR as context to deliver health information search results adapted to the user health conditions. In their system, MISearch, several contextual aids were added to the interface. On the search results page, together with the title, snippet and URL, there is also a context line with the items from the PHR that were found in each document and a link

to a cached version of the document where context keywords are highlighted. They also have contextual filters in the results page, namely a drop-down list where the user may choose if he wants to see all the results, results relevant to his context or results relevant to a specific context of his PHR. The system also provides options to restrict context intrusiveness where the user may select the ordering criteria: search query, context relevance or domain name. An evaluation of the system showed that the context was more useful when searching for difficult or misleading information.

Collective dimension

DynaCat (Pratt et al., 1999) is a well-known system in the healthcare domain. This system uses UMLS knowledge and MeSH terms to dynamically categorize search results. Another ontology, Gene Ontology, is used by a system proposed by Doms and Schroeder (2005) to categorize search results, to show terms related to the original query, to highlight terms in the abstract and to present definitions of terms.

In the work of Martins et al. (2008) search results are organized into a taxonomy of UMLS concepts. In his work regarding the hypertextual medical therapeutics handbook, Frisse (1987) also gives generic suggestions for electronic books interfaces.

Ranking

The only system that uses context features to rank results is the one proposed by Silva and Favela (2006). Google is used to obtain an initial list of relevant results to a query string, which are then reordered according to their relevance to context keywords of the PHR.

3.5 CONCLUSION

In this chapter we described the second part of the background behind the work presented in this dissertation. The work here described is related to context and its use in IR and, more specifically and intensively, in Health IR.

In this dissertation, context is considered an interactional problem, as defined by Dourish (2004). It not only includes the environmental features surrounding the user and his activities, but also the interaction in which he is involved. We believe context is dynamic and might change each time a new search is made, a new set of results is reviewed or a new document is viewed (Harper and Kelly, 2006). Therefore, "it arises from and is sustained by the activity itself" (Dourish, 2004).

The classification of Health IR research showed a weaker use of interaction context features than we expected, a different reality from the one we found in the IR research analysis where interaction was the second most used category. Generally and specifically in Health IR, research makes an extensive use of collective features. This was not a surprise because this dimension is very comprehensive, including several types of context features. In addition, it is the dimension where all the health-related structured knowledge sources (e.g.: thesauri) are included. A considerable number of papers use context to support query formulation. In Health IR this is the stage where context is more frequently used and, in general IR, it is only surpassed by the Indexing & Searching stage.

We noticed that research has been more focused on health professionals than on consumers. Of the 27 Health IR papers analyzed, only 3 are dedicated to health consumers and 2 are dedicated to both professionals and consumers. This difference may be explained by the longer tradition of information retrieval in health professionals when compared to consumers. Only recently, with the advent of the Web, has search become more popular among health consumers. Other possible reasons include the large number of medical knowledge sources, the possibilities open by the integration of search systems with clinical systems and the difficulties associated with user studies in consumer health retrieval. The lack of research on the use of context in health IR by consumers highlights the importance of focusing research on health consumers.

In Part II we describe three exploratory studies conducted to explore the influence of context on health information retrieval. In each study we use different context features. In the study described in Chapter 5 we analyze the outcome of different types of search engines (individual context according to the Ingwersen and Järvelin's nested model of contexts centered in the information space), namely generalist and health-specific ones, with different types of work tasks (session context according to the same model). In Chapter 6 we compare the outcome of search engines retrieving only certified documents with search engines retrieving also non-certified documents (individual context). Moreover, we compare the influence of documents features like their certification and HON categories in the retrieval process (intra-object context). In the study described in Chapter 7 we explore the influence of several user (individual context), task (interaction and collective contexts) and document features (intra-object context) on their interaction behavior, namely, query formulation and relevance assessments.

PART II

CONTEXT INFLUENCE ON CONSUMER HEALTH INFORMATION RETRIEVAL: EXPLORATORY STUDIES

USER EXPERIMENT 1

4.1 INTRODUCTION

We conducted an interactive light IR experiment, run on the laboratory, with 41 participants choosing 2 information needs associated with one or two of five possible simulated work tasks. To satisfy each information need, users chose 4 of 7 possible search engines, 4 of which are generalist and 3 are health-specific. More details regarding this study are given in the following sections.

This experiment was the basis for two studies, one described in Chapter 5 and the other in Chapter 7. The first study evaluated and compared search engines in health searches and the second study explored how context affects query formulation and subjective relevance in the health domain.

4.2 WORK TASKS

Following the framework proposed by Borlund (2003b), we defined five simulated work tasks based on popular (most viewed) questions submitted to web health support groups. Each work task acts as the context of four information needs (IN) that are linked to it. The defined work tasks are transcribed next.

1. You are the sibling of a 5-year old child who, usually, is irritable throughout the day. There are times when you feel you cannot keep up with the situation any longer but, on the other hand, you also feel sorry for her. You think she may suffer from bipolar disorder and you want to know more about this disease. For example, (IN1.1) to know what characterizes the disease, (IN1.2) if children can have this disease, (IN1.3) how to deal with people affected by the disease and (IN1.4) to know treatments for it.

2. You are going on vacation to Maldives in the next month with a 6-month old baby and are worried with the dangers of sunlight exposure. You are even thinking on changing your vacation destiny. To make a final decision you need to know more about sunlight exposure dangers. For example, (IN2.1) what are the problems of being at the sun, (IN2.2) what are the problems of exposing a baby to the sun, (IN2.3) does it matter what is your skin type in the first place and (IN2.4) is the danger associated only with getting burned or simply with being in the sun.

3. Your mother, 55 years old, says that while she does some heavy work her right side breast is paining heavily. You are afraid it might be a symptom

of breast cancer and you need to know more about this type of cancer. For example, (IN3.1) you want to know what symptoms are associated with this disease, (IN3.2) how is the diagnosis done, (IN3.3) how is treatment done and (IN3.4) what are its survival rates.

4. You are going through tough times and believe you are suffering from a major depression brought on by a multitude of factors: stress, work, family and poor diet. This has been affecting you deeply. You have bouts of sudden mood swings, feelings of worthlessness, thoughts of suicide, a lack of interest in favorite hobbies, no social activities and lack of sleep patterns. You need to know more about depression. For example, you need to know (IN4.1) what types of depression exist and how are they characterized, (IN4.2) how is a depression diagnosed, (IN4.3) how to recover from a depression and (IN4.4) what drugs are usually used in the treatment.

5. Your brother, 42 years old, had kidney stones some years ago and, recently, is having similar pains. Besides these pains, he also bleeds when he pees. You need to know more about kidney stones. For example, you need to know (IN5.1) if the bleeding is related to the kidney stones, (IN5.2) what are the symptoms, (IN5.3) causes and (IN5.4) major problems.

Work tasks are associated with the following medical specialties: gynecology, dermatology, psychiatry and urology. They are also categorized as severe or non-severe. Any life threatening or long-term, chronic illness is considered a severe condition. Each information need is associated with one of the following types of clinical questions: overview, diagnosis/symptoms, treatment, prevention/screening, disease management and prognosis/outcome. These categories were defined upon the categories of clinical questions presented by Hersh (2008a) and the information categories available in MedlinePlus topics. The associations between information needs and the above categories are presented in Table 4.1.

4.3 SEARCH ENGINES

We included 4 generalist web search engines and 3 health-specific ones in our study, as expressed in Table 4.2. Google, Bing and Yahoo! were selected for their popularity. At least in two rankings (alexa.com and hitwise.com) they are positioned as the top-3 search engines. Sapo was included because it is the main Portuguese search engine. MedlinePlus is a service of the U.S. National Library of Medicine and was included for its credibility. We also included WebMD because, according to the US market share of visits[1], it is one of the main services of this type and SapoSaúde for the same reason in what concerns Portugal. Sapo and SapoSaúde are both owned by the same company.

[1] http://www.marketingcharts.com/interactive/top-10-health-medical-information-websites-july-2010-13919/. Archived at http://www.webcitation.org/6DeU6s6c6.

Table 4.1: Work tasks used in this study

Task	Specialty	Severe?	IN	Clinical question	#Users[#F,#M]
1	Psychiatry	Yes	IN1.1	Overview	8[5,3]
			IN1.2	Overview	2[0,2]
			IN1.3	Disease Management	2[1,1]
			IN1.4	Treatment	5[4,1]
2	Dermatology	No	IN2.1	Prevention/Screening	8[3,5]
			IN2.2	Prevention/Screening	2[1,1]
			IN2.3	Prevention/Screening	5[3,2]
			IN2.4	Prevention/Screening	0[0,0]
3	Gynecology	Yes	IN3.1	Diagnosis/Symptoms	7[5,2]
			IN3.2	Diagnosis/Symptoms	6[5,1]
			IN3.3	Diagnosis/Symptoms	1[1,0]
			IN3.4	Diagnosis/Symptoms	1[1,0]
4	Psychiatry	Yes	IN4.1	Overview	8[6,2]
			IN4.2	Diagnosis/Symptoms	8[5,3]
			IN4.3	Treatment	5[4,1]
			IN4.4	Treatment	3[1,2]
5	Urology	Yes	IN5.1	Diagnosis/Symptoms	2[2,0]
			IN5.2	Diagnosis/Symptoms	3[3,0]
			IN5.3	Diagnosis/Symptoms	4[4,0]
			IN5.4	Diagnosis/Symptoms	2[2,0]

Table 4.2: Search engines included in this study

	URL	Type
Bing	http://www.bing.com/	Generalist
Google	http://www.google.com/	Generalist
	http://www.google.pt/	
MedlinePlus	http://www.nlm.nih.gov/medlineplus/	Health-specific
Sapo	http://pesquisa.sapo.pt/	Generalist
SapoSaúde	http://saude.sapo.pt/	Health-specific
WebMD	http://www.webmd.com/	Health-specific
Yahoo!	http://www.yahoo.com/	Generalist

4.4 PROCEDURE

Each user chose two information needs, belonging to the same or different tasks and four search engines of any type. For each information need, users had to formulate a query and submit it to the selected search engines. Users were asked to, whenever possible, use the same query in every search engine. However, they were allowed to change it if the query did not return enough results or if its language needed to be adjusted to the language of the search engine's contents.

Following the pooling approach, each user assessed the relevance of the top-30 documents returned by each engine in a 3-graded scale (0-non relevant; 1-partially relevant and 2-totally relevant).

Before the relevance assessments, users answered an initial questionnaire, available in Appendix A, focused on the collection of demographic data, web search experience, health seeking behavior, previous searches on the topic and knowledge on the work task. After assessing relevance, users answered a final questionnaire, available in Appendix B, with questions about the information needs they chose, why they chose them and about the task completion status.

4.5 PARTICIPANTS AND THEIR CHOICES

Forty-one undergraduate students participated in this study (27 females; 14 males) with a mean age of 27.2 years (Standard-Deviation – SD = 10.02). These students evaluated 9,572 documents, less than $41 \times 2 \times 4 \times 30$ because some queries returned less than 30 documents. There was a total of 82 sets of judged documents, one for each pair of user and information need, from which repeated documents obtained in different search engines were excluded.

The average number of years users have been searching the Web is 8.37 years (SD = 3.05), most of the students (61%) do one or two web searches a day (4 in `ws_fre` in Figure 4.1) and more than 80% of the students say they find what they want almost all the time (4 in `ws_suc`).

The Web is not used to search for health information by 22% of the students. As can be seen in Figure 4.1, the frequency of health searches (`hs_fre`) is much lower than the frequency of web searches (`ws_fre`). The majority of the students (40%) does this type of searches one or twice a month and 33% say they do it one or two times a year. In these searches, users feel less successful (`hs_suc`) than in general web searches. Globally, students consider they have a good health condition (`hstat` in Figure 4.1).

As can be seen in Table 4.1, with the exception of one information need, all were associated with at least one user. Only 25% of the selected information needs were about a previously searched topic. In a global perspective, as can be seen in variables `clar`, `comp` and `fam` of Figure 4.1, students found the tasks clear, moderately complex and were somehow familiar with the topic.

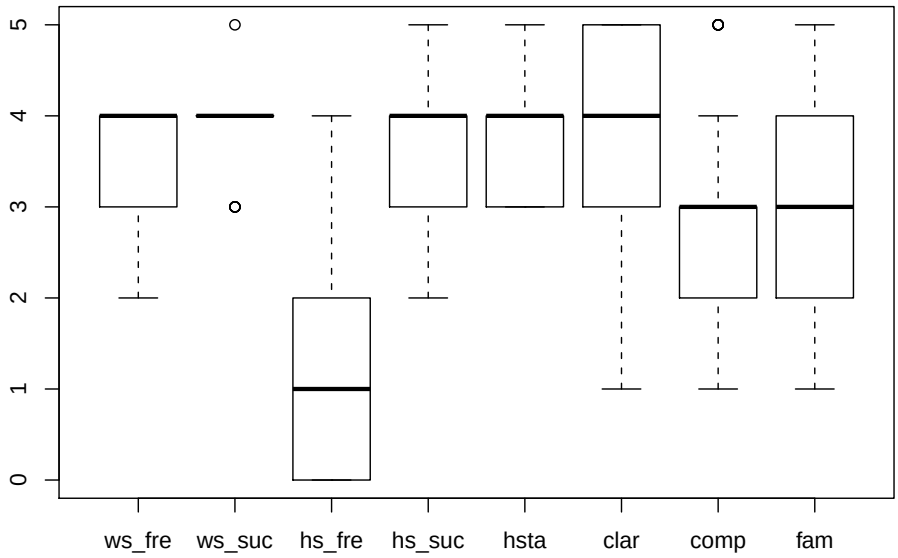

Figure 4.1: Distributions of ordinal variables. Variables' descriptions and scales in Table 4.3.

Every user chose Google as one of the four engines. In Figure 4.2, we present the number of users selecting each search engine.

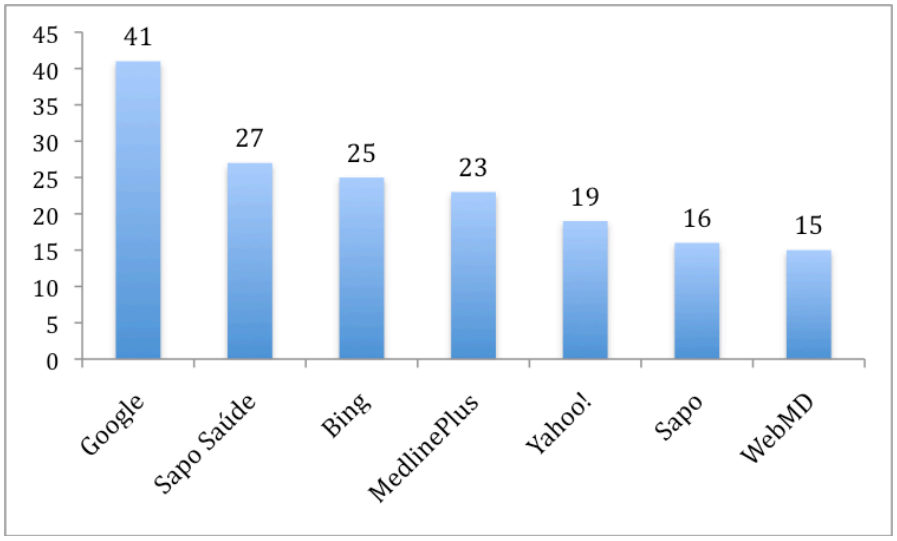

Figure 4.2: Number of users selecting each search engine

Users' search behavior was controlled in the sense that they could only choose search engines from a predefined list and they had to focus on the top-30 documents.

4.6 SUMMARY OF CONTEXT FEATURES

The context features involved in this experiment are summarized in Table 4.3.

4.7 CONCLUSION

In this Chapter we have described an interactive light IR experiment run with 41 undergraduate students. The data collected in this experiment is used in two studies described in this Part of the dissertation, namely Chapter 5 and Chapter 7. Different context features have been used in each mentioned study. In the next chapter we describe the first study involving the comparison of several health-specific and generalist search engines in tasks with different medical characteristics.

Table 4.3: Context features used in the experiment.

Dimension	Feature	Description	Scale	Values
user	age	-	ratio	-
	gender	-	nominal	female, male
	hstatus	Health status self-evaluation.	ordinal	1 (not healthy) to 5 (healthy)
web search	ws_freq	frequency	ordinal	1 (twice a year) to 5 (more than twice a day)
	ws_success	success rate	ordinal	1 (never find) to 5 (always find)
	ws_years	years of experience	ratio	-
health search	hs_freq	frequency	ordinal	1 (twice a year) to 5 (more than twice a day)
	hs_success	success rate	ordinal	1 (never find) to 5 (always find)
	hs_webuse	Web use for these searches?	nominal	no, yes
	usualengine	Is this engine typically used?	nominal	no, yes
topic's familiarity	familiarity	self-evaluation of familiarity	ordinal	1 (not familiar) to 5 (familiar)
	prev_search	previous searches	nominal	no, yes
task	clarity	-	ordinal	1 (not clear) to 5 (clear)
	easiness	-	ordinal	1 (difficult) to 5 (easy)
	qtype	question type	nominal	overview (o), disease management (dm), treatment (t), prevention/screening (p/s), prognosis/outcome (p/o), diagnosis/symptoms (d/s)
	specialty	medical specialty	nominal	psychiatry (p), dermatology (d), gynecology (g), urology (u)
	taskstat	completion status	ordinal	1 (failure) to 5 (success)
query	medterms	use of medico-scientific terminology?	nominal	no, yes
	nterms	number of terms	ratio	-
	qlang	query language	nominal	EN, PT
	qadv	advanced or boolean operators	nominal	no, yes
document	docrank	position in the ranking	ordinal	-
	doctype	file type	nominal	doc, html, pdf, ppt, swf
	snippet	snippet length	ratio	-
	title	title length	ratio	-

COMPARATIVE EVALUATION OF SEARCH ENGINES IN HEALTH INFORMATION RETRIEVAL

5.1 INTRODUCTION

Health consumers are increasingly using the Web to search for health information. Fox and Duggan (2013) found that 77% of the health searches start at generalist search engines and 13% at health-specific websites.

According to Hersh (2008a), the amount and quality of evaluation research didn't follow the changes that Information Retrieval has suffered with the ubiquity of the Web. In his opinion, the number of studies evaluating the performance of web search systems in health is "surprisingly small". We focus on health consumers because, as we have reported on Chapter 3, they receive less attention when compared to professionals.

The study described in this Chapter evaluates the performance of 4 generalist search engines (Google, Bing, Yahoo! and Sapo) and 3 health-specific search engines (MedlinePlus, WebMD and SapoSaúde). The evaluation is based on the data collected in a user study with undergraduate students and work tasks defined according to the framework proposed by Borlund (2003b). Besides an overall comparison, search engines are also differentiated by their performance on different clinical questions, medical specialties and levels of severity.

We start to review previous works on evaluation of web search engines and, more specifically, on their evaluation on the health domain. Next, we describe our methodology, present the study and discuss our results. We conclude with some final remarks.

5.2 SEARCH ENGINES EVALUATION

5.2.1 Evaluation in Information Retrieval

Information Retrieval is a highly empirical field in which evaluation is essential to demonstrate the performance of new techniques (Manning et al., 2008). The use of test collections is the dominant evaluation standard, being used since the early 1950s along with evaluation measures (Sanderson, 2010). Since 1992, TREC (Text REtrieval Conference) has been a major forum to discuss research evaluated through this model. The use of test collections is particularly well suited to system-oriented performance evaluations that focus on specific aspects of systems. In fact, this model allows a greater control of sources of

variability and, therefore, an increasing power of experiments that can be run with lower costs (Voorhees, 2008).

Although popular, evaluations with test collections are restrictive in terms of the cognitive and behavioral features of the IR system's environment (Borlund, 2003b). In fact, users, their interaction with the system, their intentions with the query and the process used to judge relevance are not included in these experiments. Ingwersen and Järvelin (2005) do an interesting systematization of the problems of this evaluation type in the form of 10 objections and responses. In summary, they point that this model lacks (1) users and tasks, (2) interaction and dynamic requests, (3) tactical variability, (4) uncertainty, (5) user-oriented relevance, (6) variety in collections. Moreover, (7) it assumes document independence and neglects overlap, (8) it is based in recall and precision that are insufficient, (9) does heavy averaging and (10) is just document retrieval with little attention to the presentation or use of information. In spite of the attempts to incorporate users in TREC, as done in the Interactive and HARD tracks, context is still "reluctantly and minimally acknowledged" (Jones, 2006).

Experimental methods involving the user also exist and have been promoted by Ingwersen and Järvelin (2005) and by Borlund (2003b). Ingwersen (2009) identifies three major types of research methods involving users: ultra-light IR interaction experiments, interactive light IR experiments and naturalistic IR field studies in the context of, for instance, an organizational setting. The first focus on short-term IR interaction composed of 1 to 2 retrieval runs. This kind of experiments allows the use of existing test collections in the form of laboratory studies provided that the number of iterations is limited to avoid learning effects by test persons. Interactive light IR experiments entail session-based multi-run interaction with more intensive monitoring like log analysis, interviews and observation. They can be run in a laboratory, in naturalistic settings or in the Internet through what Sanderson (2010) calls Live Labs. The Interactive IR evaluation model proposed by Borlund (2003b), that integrates the simulated work task situations, is an example of an interactive light experiment. Naturalistic studies assume live tasks in natural environments. They may be a field experiment if an experimental situation with test persons is used, field studies if they use real persons' natural behavior or even a case study. Field studies are typically used in social sciences and case studies in the human computer interaction area. Kelly (2009) has a good compilation of contents useful for those who need to conduct Interactive IR evaluations with users. Finally, another method that has been growing since the appearance of web search engines involves the study of user behavior using query logs (Kelly, 2009). The high cost of conducting studies with users, like the ones described above, is motivating efforts to create adequate test collections to Interactive IR (Voorhees, 2008).

The two most popular measures for IR effectiveness are precision and recall (Manning et al., 2008). Precision is the fraction of retrieved documents that are relevant and Recall is the fraction of relevant documents that are retrieved. Along with the F measure, that is, the weighted harmonic mean of precision and recall, these are the most used measures in unranked retrieval. In a ranked retrieval context, precision-recall curves can be plotted. Moreover, a very common measure is the Mean Average Precision (MAP) and, in scenarios like the Web in which it is important to have good results on the first

pages, precision is also measured at fixed levels of retrieval (e.g.: precision at 10). In situations where non-binary scales of relevance are used, Normalized Discounted Cumulative Gain (nDCG) is popular. This measure considers that highly relevant documents are more valuable than marginally relevant ones and their value decreases as the position in the ranking increases (Järvelin and Kekäläinen, 2002). In 2010, Robertson et al. proposed a new measure named Graded Average Precision (GAP) that generalizes average precision to multi-graded relevance. Because GAP and Graded Precision (gP), also proposed by Robertson et al. (2010), are used not only in the study described in this chapter, but also in other studies of this dissertation, we will describe these measures a little more deeply.

The GAP and gP measures consider a user model in which the user has a binary view of relevance even when using a non-binary scale of relevance. In this model, each point of relevance in the scale has a probability g_i of being the grade from which the user considers the documents relevant. The GAP and gP@n measures are defined as stated in Equations 5.1 and 5.2, where $\delta_{m,n}$ is defined by Equation 5.3. In these equations, g_i is the probability that the user sets the threshold at grade i, i.e., in a relevance scale of 0..c, he considers grades i..c as relevant and the others as non-relevant. R_i is the total number of documents in grade i for this query and i_n is the relevance grade of document at rank n. If $i_n>0$, document at rank n contributes to the calculations. More details on the calculation of these measures can be seen in the cited paper. Based on the evaluation results presented by the measure's proponents, an equally balanced g_1 and g_2, i.e., $g_1=g_2=0.5$ makes GAP always more informative than nDCG and Average Precision (AP).

$$GAP = \frac{\sum_{n=1}^{\infty} \frac{1}{n} \sum_{m=1}^{n} \delta_{m,n}}{\sum_{i=1}^{c} R_i \sum_{j=1}^{i} g_j} \tag{5.1}$$

$$gP@n = \frac{1}{n} \sum_{m=1}^{n} \frac{\sum_{j=1}^{min(i_n,i_m)} g_j}{\sum_{j=1}^{i_n} g_j} \tag{5.2}$$

$$\delta_{m,n} = \begin{cases} \sum_{j=1}^{min(i_n,i_m)} g_j, & \text{if } i_m > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.3}$$

5.2.2 *Evaluation of Web Search Engines*

Although we focus here on the health domain, we present a brief overview of previous works that aim to evaluate and compare the performance of generalist web search engines. In Table 5.1 we present a list of research papers with these goals, along with the number of search engines evaluated and the type of measures used to compare them. All studies, except the one from Shang and Li (2002), involve users in their evaluation, either to define the information needs, the queries or to judge the relevance of the documents. Shang and Li (2002) compute relevance scores using three traditional algorithms (cover density ranking, Okapi similarity measurement and vector space model) and an additional one developed by the authors. Besides the differences presented in Table 5.1, other distinctions lay in the users characteristics, the information needs, the queries generated (e.g.: number, how, any restrictions?) and the

method used to judge results (e.g.: number of results judged, by whom, relevance scale). Contrasting the other studies, Vaughan (2004) uses a continuous relevance scale (from the most relevant to the least relevant result) instead of a discrete relevance scale. Statistical comparison of the measures was the standard method to analyze the results.

Table 5.1: Previous studies on Evaluation of Web Search Engines (SE)

Study	# SE	Evaluation Criteria
(Chu and Rosenthal, 1996)	3	Search capabilities, output options, documentation and interface. Response time, precision.
(Gordon and Pathak, 1999)	8	Recall and precision at varying numbers of retrieved documents.
(Gwizdka and Chignell, 1999)	3	Precision, presentation, user effort and coverage.
(Hawking et al., 2001)	20	Precision oriented measures: P@n at n≤20, mean reciprocal rank of first relevant document and TREC-style average precision.
(Shang and Li, 2002)	6	Precision oriented measures.
(Su, 2003b)	4	16 performance measures in 5 criteria: relevance, efficiency, utility, user satisfaction and connectivity.
(Tang and Sun, 2003)	4	*First 20 full precision*, as proposed by Chignell et al. (1999), search length and rank correlation.
(Vaughan, 2004)	3	Quality of result ranking, ability to retrieve top ranked pages and 3 stability measurements.
(Lewandowski, 2008)	5	Precision measures and recall-precision graphs applied to results and to their descriptions.

The work from Chu and Rosenthal (1996) also includes an evaluation methodology for web search engines. In their opinion, search engines should be evaluated considering 5 aspects: composition of web indexes, search capability, retrieval performance, output option and user effort. Su (2003a), in a work previous to the one presented in Table 5.1, proposes a set of criteria and measures as a systematic model to evaluate web search engines.

5.2.3 Evaluation of Web Search Engines in the health domain

Hersh (2008a) does a broad review of studies that evaluate search systems in the health domain in terms of system and user performance. The majority of the studies focus on professionals' systems, mostly using MEDLINE. The number of web search systems evaluated in the literature is, as Hersh says, "surprisingly small". In this section we will review previous studies that, at least, evaluate one web search engine. Studies that, for example, evaluate and compare two MEDLINE search systems will not be reported here. We will give more attention to papers that focus on health consumers.

That we are aware of, only three papers explore the performance of web search engines in the health domain when used by professionals. This might

be explained by these users' preference on sources like MEDLINE instead of the Web to satisfy their information needs. In Table 5.2 we present the web search engines included in each study and also their evaluation criteria. All studies compare web search engines with other kind of resources. In the study from Johnson et al. (2008), users were randomly assigned to Google or other web resource of their choice. Graber et al. (1999) selected 10 questions posed by physicians and Yu and Kaufman (2007) chose 12 physicians questions in the format "What is X?". On the other hand, Johnson et al. (2008) used 10 medical questions extracted from a multiple choice exam. All papers evaluated the medical quality of the contents retrieved and the number of links used to get to the answer. A few other criteria were used, as can be seen in Table 5.2. Results of these studies report that health specific search engines behave poorly when compared with generalist engines. In studies where Google was used, authors concluded that this is an effective engine for health information.

Table 5.2: Studies evaluating Web Search Engines (SE) in the professional health domain

Study	SE	Evaluation Criteria
(Graber et al., 1999)	1 site, 4 generalist engines, 9 medicine-specific engines, 2 medical meta-lists	Number of questions answered, correctness of the answers, number of links followed to get an answer and how well documented the answer was using Health on the Net criteria.
(Yu and Kaufman, 2007)	Google, MedQA, Onelook, PubMed	Quality of answer, ease of use, time spent, and number of actions taken.
(Johnson et al., 2008)	Google vs. other web resources	Resource efficiency (inversely related to number of links used to identify the correct answer) and correctness (# correct answers/# answered questions).

We analyzed seven papers that evaluate web search engines on the health domain in the consumer's perspective. A summary of the main differences between these works is presented on Tables 5.3 and 5.4.

All papers evaluate and compare several search engines and, if we exclude the work from Jones and Timm (2008), all include, at least, one generalist search engine. The works from Kumar (2005) and Tang et al. (2006) are user studies. Only the Wu and Li (1999) study had the contribution of two librarians. As seen in Table 5.4, the authors, either selecting questions posed to librarians/clinicians or consumers' popular questions, formulated information needs. The method used to evaluate popularity was not mentioned in any of the papers. Bin (2001) consider two search types: single keyword searches (SKS) in which the authors want to retrieve information about a term and question-answering (QA) to evaluate the answer to a clinical question. In QA, authors used the questions previously defined by Graber et al. (1999). Tang et al. (2006) employ two types of information needs, related and non-related to treatments. The first type of queries is based on treatments' names to which they had evidence ratings. This allowed the quality evaluation of the documents retrieved without the intervention of health professionals. The second type of queries was extracted from the search logs of a depression search en-

gine and from the suggestions given by a tool based on common queries. In the first type of queries, users judged documents relevance and treatment recommendations in retrieved documents. In the work of Kumar (2005), the relevance of the documents was assessed by users in an aggregated way with a post-search questionnaire. Health professionals judged the quality of documents' content in an individual way.

As can be seen in Table 5.5, the criteria to evaluate the search engines included the relevance of the results, quality of results from a medical point of view, usability and search engines' features. Results' quality is evaluated in different ways. Some analyze documents' characteristics like authorship, evidence of citation, disclosure and currency (Wu and Li, 1999; Ilic et al., 2003). Kumar (2005) asked health professionals to judge the accuracy and trustworthiness of results and Tang et al. (2006) used treatments' evidence ratings to validate their quality. Finally, Knight et al. (2009) used the FA4CT algorithm proposed by Eysenbach and Thomson (2007) for the same purpose.

A common pattern emerges from all studies including generalist search engines. All conclude that the performance of generalist search engines is equal (Bin, 2001), or better (all the others) than the performance of health-specific ones. Regarding information quality, some studies concluded that health-specific ones outperform the generic ones (Kumar, 2005; Tang et al., 2006) and the other said there were no differences (Ilic et al., 2003).

From all these studies, the one from Jones and Timm (2008) stands out for its qualitative nature. The work of Tang et al. (2006) is the closest to work presented in this chapter. It is a user study, it has objective methods and it uses well-known measures in Information Retrieval. Our study differs in the larger involvement of users in the experiment, the use of different measures to evaluate performance and the inclusion of tasks of different medical specialties, types of clinical questions and levels of severity. More specific differences are detailed in the sequel.

5.3 DATA ANALYSIS

The experiment described in Chapter 4 served as the basis for this analysis.

5.3.1 Overall strategy

Our analysis was done in four perspectives, as depicted in Figure 5.1. Initially we did a global analysis with 4 goals: (1) to find if and where are the differences in the performance of each category of search engine (health and non-health) and, more specifically, on each search engine; (2) if and where are the differences in the answers to the several types of clinical questions; (3) to the different medical specialties and (4) to find if severe and non-severe conditions are associated with different global performances. We then focused on the differences related to the types of clinical questions. Here, we investigate (5) if, in each type of search engine and in each specific search engine, there are differences in the performance for different types of clinical questions. We also studied (6) the differences that exist in each type of clinical questions (e.g.: in treatment information needs, are there differences between types of search engines and between search engines?). We followed this same strategy to analyze the differences in the medical specialties and within levels of severity (7-10).

Table 5.3: Studies evaluating Web Search Engines (SE) in the consumer health domain - part I. Generalist SE signed with *.

	(Wu and Li, 1999)	(Bin and Lun, 2001)	(Ilic et al., 2003)	(Kumar, 2005)	(Tang et al., 2006)	(Jones and Timm, 2008)	(Knight et al., 2009)
SE							
#	7	8	9	3	4	6	9
Which?	Altavista*	Altavista/Health	AltaVista*	Google*	4sites	Healia	Dogpile*
	Excite*	Excite/Health	DrKoop	Healia	Blue Pages (BPS)	HealthFinder	Healia
	Hotbot*	HardinMD	Excite*	MedHunt	Google*	Healthline	Healthline
	Infoseek*	MedHunt	HealthInsite		HealthFinder	MedlinePlus	Google*
	Medical World*	MediAgent	HON			Medstory	Jux2*
	Northern Light*	Medical World	Google*			Yahoo!/Health	Kosmix Health
	Yahoo!*	Medical Matrix	MedlinePlus				RevolutionHealth
		Yahoo!/Health	NHS				WebMD
			Yahoo!*				Yahoo!*
Users							
#	2	-	-	66	Not specified	-	-
Type	Librarians	-	-	Volunteers	Research assistants	-	-
IN							
#	5	SKS: 4, QA: 8	1	6	Not specified	Not specified	5

Table 5.4: Studies evaluating Web Search Engines (SE) in the consumer health domain - part II.

	(Wu and Li, 1999)	(Bin and Lun, 2001)	(Ilic et al., 2003)	(Kumar, 2005)	(Tang et al., 2006)	(Jones and Timm, 2008)	(Knight et al., 2009)
IN							
How	Questions posed to librarians	SKS: not specified; QA: questions posed to clinicians	Not specified	Popularity	Not specified	Not specified	Popularity
By whom	Authors	SKS: authors; QA: Graber et al. (1999)	Authors	Authors	Not specified	Not specified	Authors
Categ.	5	-	-	-	2: treatment and others	-	-
Queries							
#	5	SKS: 4; QA: 8	20	Not specified	101	Not specified	5
How	Keywords linked by operators	Not specified	Phrases and Boolean	Not specified	Treatments' names; O: BPS logs + suggestion tool	Not specified	Following the characteristics of popular queries
By whom	Librarians	Authors	Authors	Users	Authors	Not specified	Authors
Nr. Judg.	30	SKS: 100; QA: 5	50	Not specified	10	Not specified	10
Judgments							
By whom	Librarians	Authors	Authors	Health professionals	Users	Not specified	Authors
Nr. levels	2	Not specified	Not specified	10	4	Not specified	Not specified
Total Nr.	150	440	4927	Not specified	Not specified	Not specified	Not specified

Table 5.5: Criteria used to evaluate search engines in the consumer health domain

Study	Evaluation Criteria
(Wu and Li, 1999)	Relevance (relevant hits per queries topic), source reliability (authorship, source, disclosure, currency), duplicate and inactive links, search engine features.
(Bin, 2001)	SKS: number of medical resources about the topic. QA: number of questions answered, number of links followed.
(Ilic et al., 2003)	Relevance. Quality: target audience, authorship and evidence citation.
(Kumar, 2005)	Usefulness, ease-of-use, relevance, overall satisfaction, accuracy, trustworthiness.
(Tang et al., 2006)	Relevance (MAP, NDCG). Quality of advice according to evidence-based medicine (Quality Score).
(Jones and Timm, 2008)	Major features, navigation, timeliness and quality of retrieved items, search interface and strategy, search results/display, deficiencies or disadvantages, overall effectiveness.
(Knight et al., 2009)	Popularity, usability of the landing and results page, relevance (precision and relative recall), results quality and features.

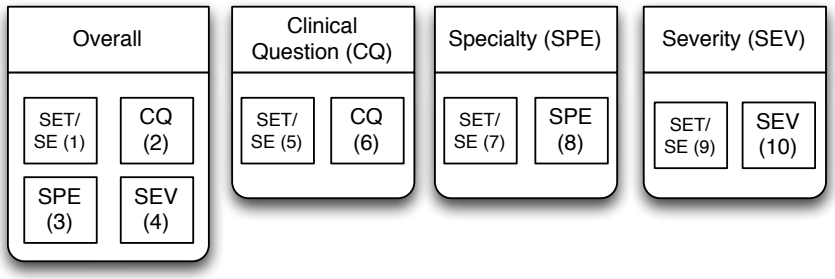

Figure 5.1: Conducted analysis (SET=Search Engine Type; SE=Search Engine)

5.3.2 *Evaluation measures*

To evaluate and compare the engines we use binary and graded relevance measures. The first type of measures includes the Average Precision (AP), precision at 5 documents retrieved (P@5) and P@10. For computing these measures we converted the 3-graded scale into a binary one. All the partially relevant and totally relevant documents in our user study were considered relevant and the others non-relevant. The second type of measures, described with greater detail in Section 5.2.1, include the Graded Average Precision (GAP), graded precision at 5 documents retrieved (gP@5) and gP@10.

All measures, binary and non-binary, will be averaged over assessment cycles. An assessment cycle is composed by relevance assessments of a specific user for a certain information need in one search engine. Since we are comparing means, we will in fact be comparing the Mean Average Precision (MAP) and Mean Generalized Average Precision (MGAP).

We decided to use GAP and gP in our study because they are recently proposed measures and GAP consistently outperformed nDCG and has the

properties of AP that led to its predominance (Robertson et al., 2010). Since Robertson et al. (2010) found that an equally balanced g_1 and g_2 made GAP always more informative than nDCG and AP, we decided to use these threshold probabilities, i.e., g_1=g_2=0.5. We decided to use binary and graded relevance measures because we want to evaluate the impact of using this new type of measures, comparing it to MAP, one of the most common measures. In fact we are comparing different threshold probabilities in the model: g_1=g_2=0.5 in GAP and g_1=1 in AP because all partially and totally relevant documents convert to relevant.

In the AP and GAP calculations we had to estimate the size of the set of relevant documents. In this sense, we considered the set of relevant documents (assessed with 1 or 2 in the relevance scale), in each pair of information need and search engine, regardless of the user. It is defined as expressed in Equation 5.4.

$$Rel(in, se) = \{d : d \in P(in, se) \wedge \exists_j(\mathrm{doc}(j, d) \wedge (\mathrm{RJ}(j) = 1 \,|\, \mathrm{RJ}(j) = 2))\} \quad (5.4)$$

In this formula *in* is the information need, *se* is the search engine, $P(in, se)$ is the pool of judged documents to information need *in* and search engine *se* and j is a relevance judgment. $\mathrm{doc}(j, d)$ holds if j is a judgment for document d and $\mathrm{RJ}(j)$ designates the value of the judgment. The set $Rel(in, se)$ contains documents from the pool for which there is a relevance judgment with value 1 or 2.

In the computation of GAP, we considered the proportion of documents in $Rel(in, se)$ classified with 1 and the proportion of documents assessed with 2.

To prevent biases, as each assessment cycle contains at most 30 judgments, if $|Rel(in, se)| > 30$, we only considered the existence of 30 relevant documents in the MAP computation. In these cases, to MGAP, we multiply the proportion of partially relevant documents by 30 and we do the same to totally relevant documents. Without this upper limit, the evaluation would be unfair to the search engines with larger number of selections that, probably, have larger collections and only 30 judgments in each assessment cycle.

5.3.3 Statistical strategy

As previously stated, our analysis is based on six measures: AP, P@5, P@10, GAP, gP@5 and gP@10. This set of measures was computed for each assessment cycle, defined by the triplet: user, information need and search engine. The mean of each of these measures was then compared between different groups using hypothesis tests. We followed the strategy presented in Figure 5.2, in each measure. Whenever possible, we applied a parametric test instead of a non-parametric one due to its greater statistical power. To select the appropriate statistical test we considered the number of groups being compared. When more than two groups were being compared, we initially applied a one-way ANOVA or a Kruskal-Wallis test to detect if there were differences between the groups. If differences were found, we either applied the Tukey's test or a pairwise comparison in which we divided the α value by the total number of comparisons to minimize the type I error. These comparisons allowed us to detect where the differences are located. In comparisons between two groups

we either applied the t-test or the Mann-Whitney to detect if and in what way there are differences. In all the comparisons we only considered groups with at least 5 assessment cycles in it.

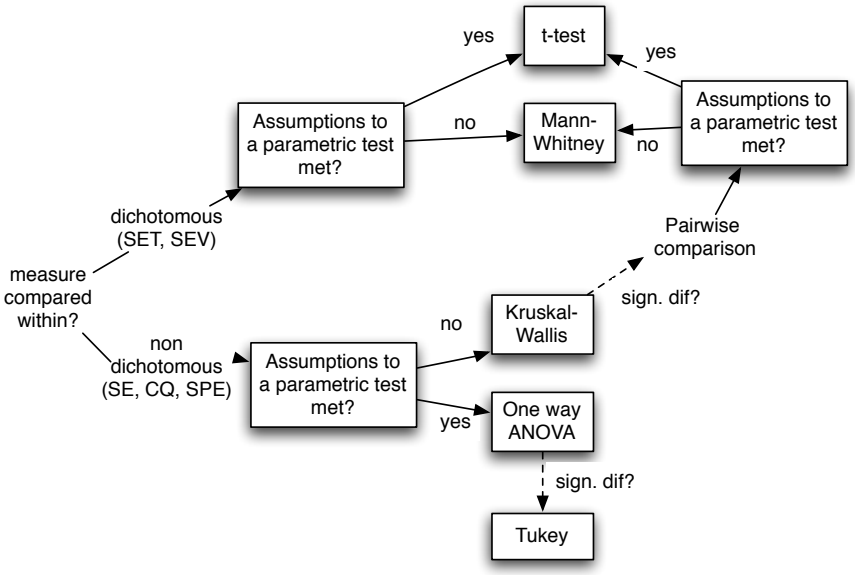

Figure 5.2: Statistical strategy

As a result of the large number of hypothesis tests performed, in the next sections we will only report significant results at $\alpha = 0.05$ or $\alpha = 0.01$. Detailed results of the hypothesis tests are presented in Appendix C. In the overall analysis we also present boxplots to graphically depict the GAP differences between groups. We chose GAP because it is an average of graded measure, therefore conveying a more stable and genuine result.

5.4 OVERALL ANALYSIS

In the broad analysis of differences between types of search engines, we can see that generalist search engines clearly have better performance than health-specific ones. This is not only visible in the boxplots presented in Figure 5.3, but also a significant difference found in all measures as indicated in Table 5.6.

In Figure 5.4, two search engines stand out, Google in a positive way and SapoSaúde in a negative way. These differences are significant in several measures and in several pairs of engines as can be seen in Table 5.6. Google is significantly better than all the other engines, mainly in top-5 and top-10 measures, and SapoSaúde worst than Bing, Google, MedlinePlus and Yahoo! in the top-5 and top-10 measures. It is interesting to note that, in average measures, Google is significantly better than all the health-specific engines. Sapo is the engine with the largest statistical dispersion.

In the clinical question analysis, we found that precision at the top of the ranking is significantly better in the overview and diagnosis/symptoms questions than in the prevention/screening ones (Table 5.6). As can be seen in Figure 5.5, differences in GAP are less evident.

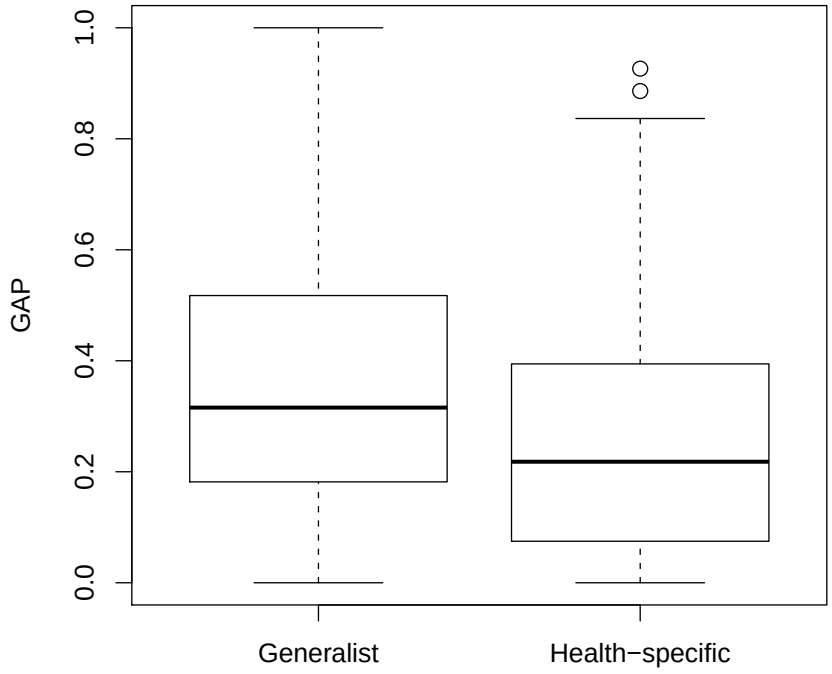

Figure 5.3: GAP comparison between search engine type

Through Figure 5.6, we can see that Urology is the specialty with highest GAP mean and also the one with lower dispersion. Yet, there are no significant differences in GAP, only in top-5 and top-10 measures where it is clear that psychiatry is better than dermatology (Table 5.6).

Finally, we also found that information needs associated with severe conditions have significantly higher performance than non-severe in all measures. This suggests there is more online information about severe health topics than non-severe ones, which agrees with White and Horvitz (2009) who, in their study about cyberchondria, conclude: "Web search engines have the potential to escalate medical concerns".

5.5 CLINICAL QUERY TYPE ANALYSIS

Our analysis by search engine type (Table 5.7) shows that, in generalist search engines and documents at the top of the ranking, overview questions have higher precision than the prevention/screening and treatment ones. In overview, diagnosis/symptoms and prevention/screening questions, almost all measures show that generalist engines have a better precision.

In Table 5.8 we see that, in Yahoo!, the overview questions have better precision in the top-10 documents than treatment questions. An analysis on types of clinical questions repeatedly shows that Google is better than the other en-

Table 5.6: Significant differences in the overall analysis.

Comparisons	**@5**	**@10**	**Average**
Generalist>Health-specific engine	gp,p	gp,p	gap,ap
Bing>SapoSaúde	gp	gp,p	
Google>Bing	gp,p	gp,p	gap,ap
Google>MedlinePlus	gp,p	gp,p	gap,ap
Google>SapoSaúde	gp,p	gp,p	gap,ap
Google>Sapo	gp,p	gp,p	
Google>WebMD	p	p	gap,ap
Google>Yahoo	gp,p	gp,p	
MedlinePlus>SapoSaúde	gp	gp,p	
Yahoo>SapoSaúde	gp	gp,p	
Overview>Prevention/Screening	gp,p	p	
Diagnosis/Symptoms>Prevention/Screening	gp	p	
Gynecology>Dermatology	gp		
Psychiatry>Dermatology	gp,p	gp,p	
Severe>Non-severe	gp,p	gp,p	gap,ap

Table 5.7: Significant differences in the query type analysis by search engine type

Where	**Which**	**@5**	**@10**	**Avg**
Generalist engine	Overview>Prevent./Screen.	p	gp,p	
Generalist engine	Overview>Treatment	p	gp,p	
Overview	Generalist>Health-specific SE	gp,p	gp,p	gap,ap
Diagnosis/Symptoms	Generalist>Health-specific SE	gp,p	gp,p	gap,ap
Prevention/Screening	Generalist>Health-specific SE	gp,p	gp,p	ap

gines. This is more evident on the top-5 and top-10 measures. MGAP and MAP show that Google is better than SapoSaúde in the overview and diagnosis/symptoms questions. It has also a larger MAP than MedlinePlus in diagnosis/symptoms questions.

5.6 MEDICAL SPECIALTY ANALYSIS

Table 5.9 shows that generalist search engines have better precision in the top documents in gynecology questions when compared to dermatology ones. All specialties have higher top-10 measures on generalist search engines. In average, this happens in psychiatry and urology.

In Table 5.10 we see that, in MedlinePlus and WebMD, psychiatry questions have higher graded precision in the top-5 documents when compared with dermatology ones. In the latter type of questions, Google surpasses 4 en-

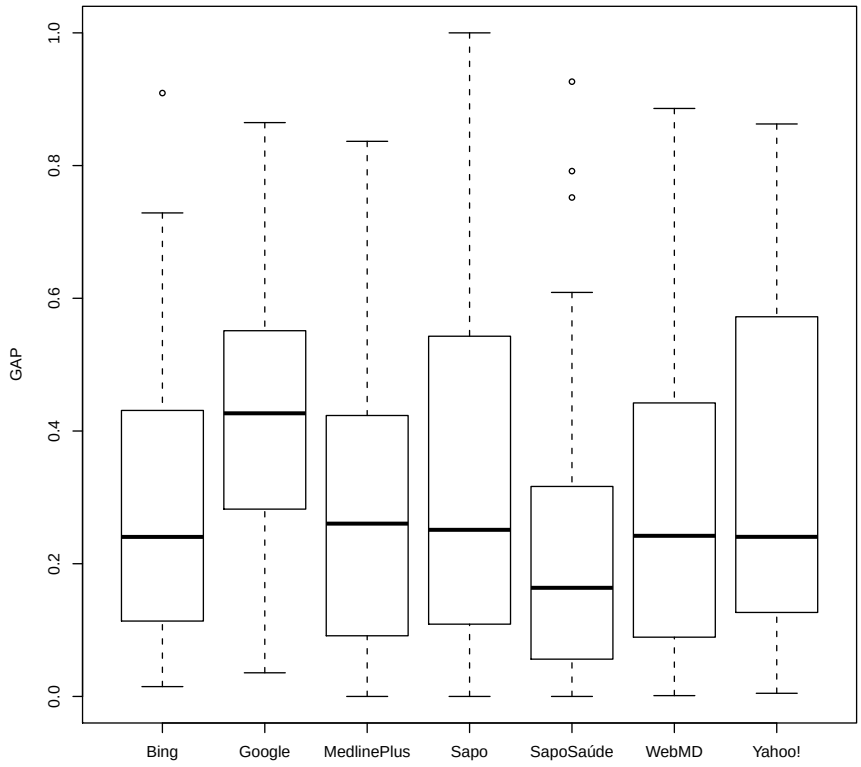

Figure 5.4: GAP comparison between search engines

Table 5.8: Significant differences in the query type analysis by search engine

Where	**Which**	**@5**	**@10**	**Average**
Yahoo!	Overview>Treatment		p	
Overview	Google>Sapo Saúde	gp,p	gp,p	gap,ap
Diagnosis/Symptoms	Google>Sapo Saúde	gp,p	gp,p	gap,ap
Diagnosis/Symptoms	Google>MedlinePlus			ap
Prevention/Screening	Google>Sapo	p	p	
Prevention/Screening	Google>WebMD	p	gp,p	
Prevention/Screening	Google>Sapo Saúde	gp	gp,p	
Treatment	Google>Yahoo!		p	

gines in the top-5 precision. All measures show us that Google is better than SapoSaúde in gynecology and psychiatry questions. With the top-10 measures Google is also better than Sapo.

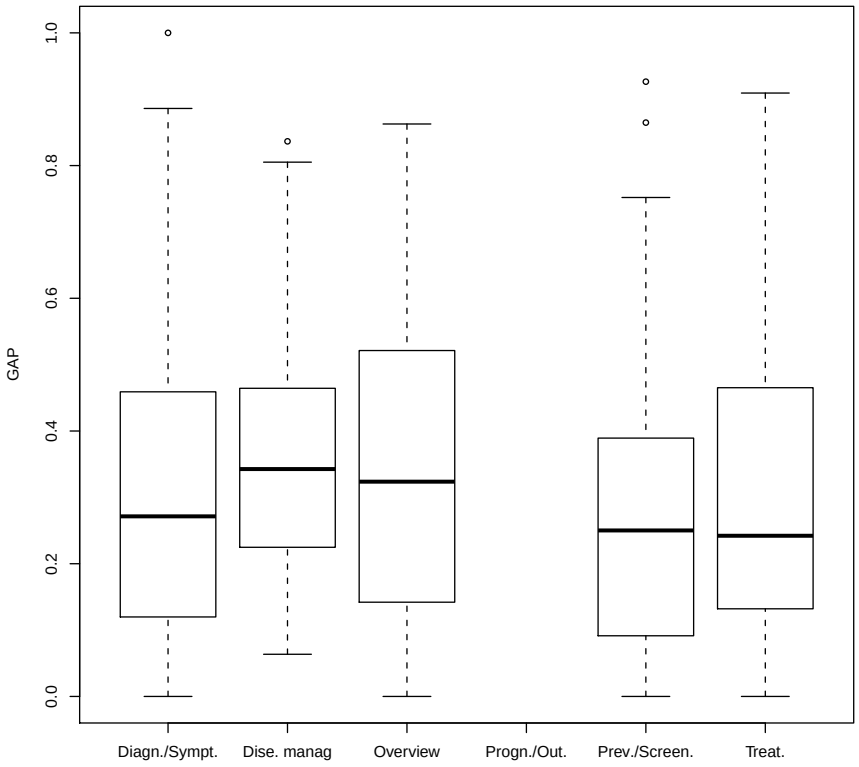

Figure 5.5: GAP comparison between query types. (Not enough data for Progn./Out.)

Table 5.9: Significant differences in the medical specialty analysis by search engine type

Where	**Which**	**@5**	**@10**	**Average**
Generalist engine	Gynecology>Dermatology	gp,p	p	
Dermatology	Generalist>Health-specific engine	gp,p	gp,p	
Gynecology	Generalist>Health-specific engine	gp,p	gp,p	
Psychiatry	Generalist>Health-specific engine		gp,p	gap,ap
Urology	Generalist>Health-specific engine		gp,p	gap,ap

5.7 CONDITION SEVERITY ANALYSIS

As can be seen in Table 5.11, severe questions have better results in both types of engines but this is more expressive in generalist ones. The tendency of generalist engines to have better performance is also visible in both levels of severity, although more in severe ones. Does this mean that health search engines have concerns on the balance of health information?

The tendency expressed above is also found on the analysis by search engine (Table 5.12), i.e., severe questions have better performance than non-severe ones. In Bing, Google, Sapo and WebMD, average measures are significantly higher in severe questions. In MedlinePlus, Sapo and WebMD, this superior-

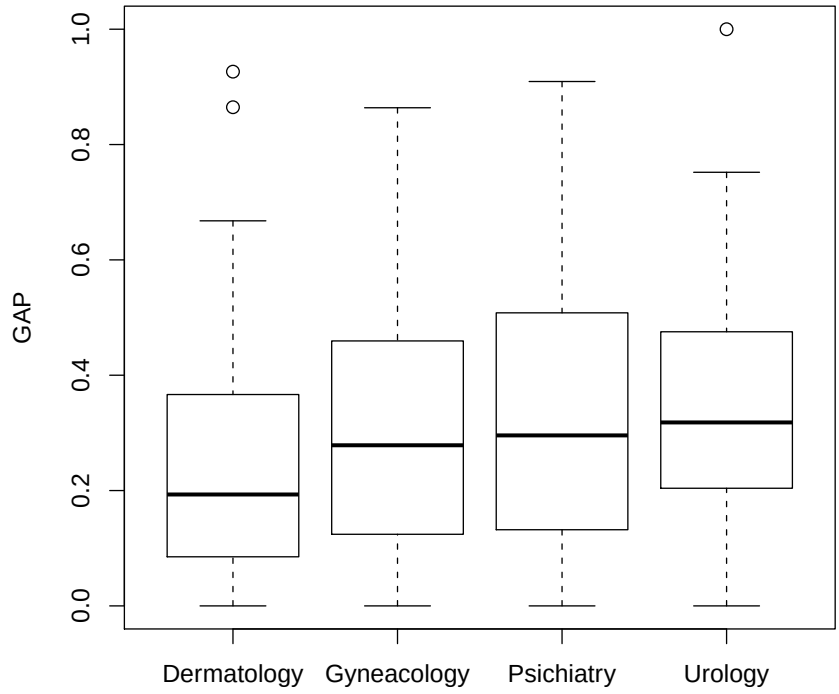

Figure 5.6: GAP comparison between specialties

Table 5.10: Significant differences in the medical specialty analysis by search engine

Where	**Which**	**@5**	**@10**	**Average**
MedlinePlus	Psychiatry>Dermatology	gp		
WebMD	Psychiatry>Dermatology	gp		
Dermatology	Google>MedlinePlus	p		
Dermatology	Google>Sapo	p		
Dermatology	Google>SapoSaúde	p		
Dermatology	Google>WebMD	p		
Gynecology	Google>MedlinePlus		p	
Gynecology	Google>SapoSaúde	gp,p	gp,p	gap,ap
Gynecology	Yahoo!>SapoSaúde		p	
Psychiatry	Google>Sapo	gp,p	gp,p	
Psychiatry	Google>SapoSaúde	gp,p	gp,p	gap,ap
Urology	Google>SapoSaúde		p	
Urology	Google>Sapo		gp	

ity is also expressed in top-5 and top-10 measures. In non-severe questions,

Table 5.11: Significant differences in the severity analysis by search engine type

Where	Which	@5	@10	Average
Generalist engine	Severe>Non-severe	gp,p	gp,p	gap,ap
Health-specific SE	Severe>Non-severe	gp,p	gp	
Non-severe	Generalist>Health-specific SE	gp,p	gp,p	
Severe	Generalist>Health-specific SE	gp,p	gp,p	gap,ap

Google is better than Sapo, SapoSaúde and WebMD in top documents. In severe questions, we can also see that Google is consistently the one with better precision in pairwise comparisons and the opposite happens with SapoSaúde.

Table 5.12: Significant differences in the severity analysis by search engine

Where	Which	@5	@10	Average
Bing	Severe>Non-severe			gap,ap
Google	Severe>Non-severe			gap,ap
MedlinePlus	Severe>Non-severe	gp,p		
Sapo	Severe>Non-severe	p		gap
WebMD	Severe>Non-severe	gp,p	gp	gap,ap
Non-severe	Google>Sapo	gp,p		
Non-severe	Google>SapoSaúde		gp	
Non-severe	Google>WebMD	p	gp	
Severe	Bing>SapoSaúde	gp	gp,p	
Severe	Google>Bing	gp,p	p	
Severe	Google>MedlinePlus		gp,p	gap,ap
Severe	Google>Sapo	gp,p	gp,p	
Severe	Google>SapoSaúde	gp,p	gp,p	
Severe	Google>Yahoo!	p	p	gap
Severe	MedlinePlus>SapoSaúde	gp	p	
Severe	WebMD>SapoSaúde	gp	gp,p	
Severe	Yahoo!>SapoSaúde	gp	gp,p	

5.8 DISCUSSION

We compared the performance of generalist and health-specific engines satisfying health information needs. Results will be discussed next along with their implications to the user and to the development of search systems. A secondary goal of our work was to compare a recently proposed measure based on graded assessments with the traditional average precision. This comparison will be made in the end of this section.

Users' preference by Google was clear since all the participants chose it as one of the search engines to use. American habits and preferences are similar.

Considering that 77% of the health sessions start on generalist engines (Fox and Duggan, 2013) and that Google's market share in 2013 desktop searches is 84%[1], we can predict that 64.7% of all American health sessions start on Google. According to our results, this is a good habit since Google has shown significantly higher precision than other search engines. Differences are even more expressive in the top documents which means Google's first results page is a good place to start a health search session.

In a global perspective, generalist search engines surpass health-specific ones in precision, and this is in accordance with almost all the studies mentioned in the literature review. Yet, health-specific engines may be more balanced in the type of contents they provide in terms of severity. Indeed, although both type of engines show higher precision in severe conditions, a smaller number of significant differences is found on health-specific ones.

Therefore, in order reduce the bias of the results, it might be a good practice to complement the results gathered from generalist search engines with the ones given by health-specific engines.

The higher precision obtained for severe conditions makes us suspect there is more online information about severe health topics than non-severe ones, and this may raise the problem of escalations on medical concerns. This finding alerts to the potential danger of online health information and should be considered in systems' development.

Overview clinical questions tend to have higher precision, mainly in the top results. In the complete set of search engines, they are better than the prevention/screening questions and, in generalist engines they are also better than the treatment ones. Conceptually this type of question is more comprehensive than the others and this may explain the better results. When other clinical queries have bad results, a good strategy may be their conversion to an overview type with which a user may get the specific information they want.

In the top-5 and top-10 results, gynecology and psychiatry medical specialties have better performance than dermatology questions in the complete set of engines. This difference is more evident in the psychiatry specialty. Has the Web more and better information on this topic? Is it easier to discuss this kind of topics online? In generalist engines, only the gynecology superiority stands.

To evaluate the relation of our results with the topics/medical specialties popularity, we have estimated the popularity of the medical condition behind each work task in two axes: number of web pages and number of searches on that topic. The number of web pages estimate was based on Google's total number of results for a query with the medical condition. On the other hand, the number of searches was estimated using on Google Trends the same expression/query. Results were aggregated by medical specialty and then normalized through the division by the maximum value found in the set of medical specialties for each axis. Figure 5.7 presents these values and also the mean of Google GAP and the mean of the overall GAP for each specialty. These two last measures were also depicted to help analyze the relation between popularity and search engines' performance. In particular, Google GAP was included

[1] http://marketshare.hitslink.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0&qptimeframe=Y. Archived at http://www.webcitation.org/6EhBXOb2O.

because our popularity estimates were based on Google information.

In Figure 5.7 we also see that psychiatry and gynecology topics are the most popular, which may help explain the significant differences mentioned above. In this figure, the urology specialty contradicts this tendency, being an unpopular specialty but having the highest GAP mean. Although this superiority in performance is not significant, this led us to analyze the correlation between GAP mean and popularity. We found a correlation of 0.34 with the number of pages and of 0.41 with the number of searches. The correlation is not high, which may imply that the search engines' performance is explained not only by topics' popularity but also by other factors like users' context.

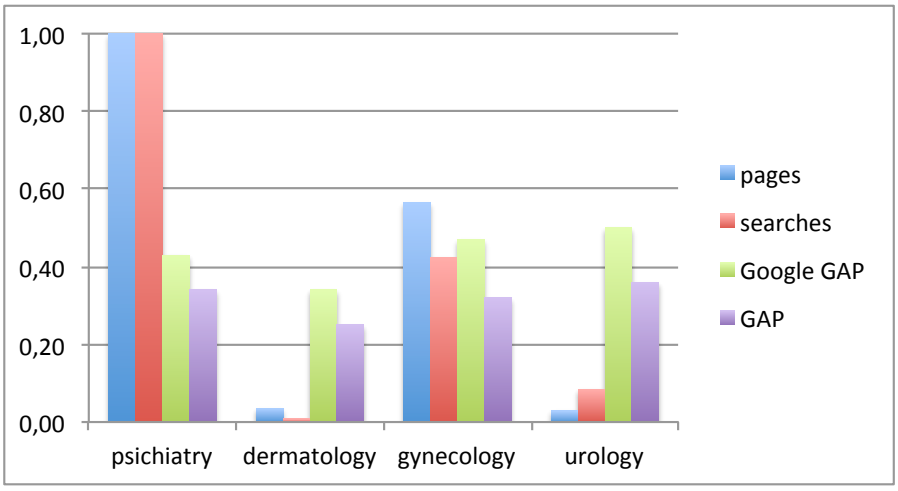

Figure 5.7: Popularity of the topics' medical specialties

One of our goals was also to compare graded average precision and average precision or, in other words, to compare different threshold probabilities in the model underlying GAP. The first threshold probabilities were defined based on the results of the measure's proponents ($g_1 = g_2 = 0.5$) and the second is associated to the commonly used average precision ($g_1 = 1$, $g_2 = 0$). In Figures 5.8 and 5.9 we see that both types of measures have a very similar behavior across search engines. The main difference lays in the magnitude of values. Generally, precision values are 0.1 higher than graded precision ones. This is natural, since the first type considers all the documents assessed with 1 and 2 relevant and, in the second, a document assessed with 1 has only a 0.5 probability of being relevant. In each type of measure, and also as expected, precision at 5 is higher than precision at 10 which, in turn, is higher than average precision.

We also analyzed the significant differences found with each measure. In Figure 5.10 we present the number of differences found with each measure. This number tends to decrease as the number of results in the calculation increases. In GAP and AP the number of differences is smaller than on the other measures. This was expected since these measures are more aggregated and stable. They not only average but also consider more results. The exception to this trend happens with p@10 in which we found more differences than with p@5.

In Figure 5.11 we present the proportion of differences found with both or only one of the measures. More than 60% of measures are significant in

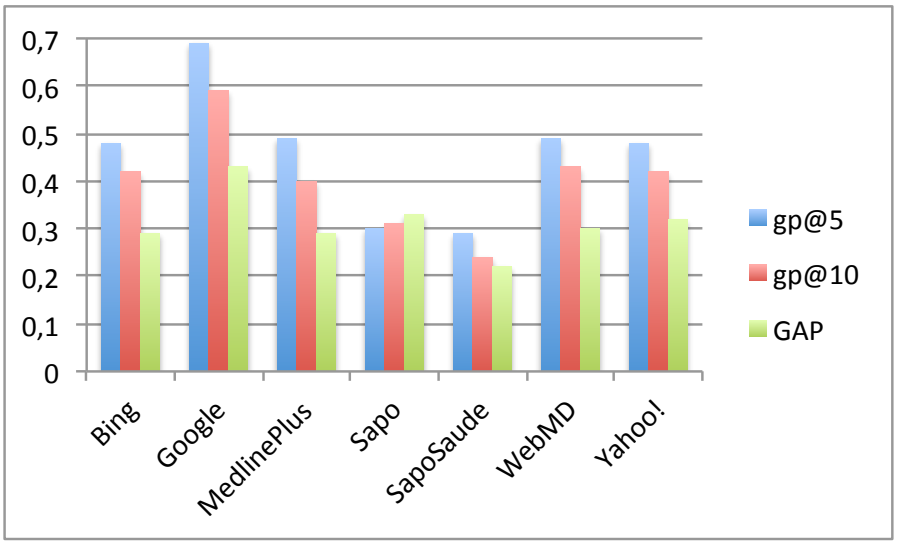

Figure 5.8: Average of graded measures in each search engine

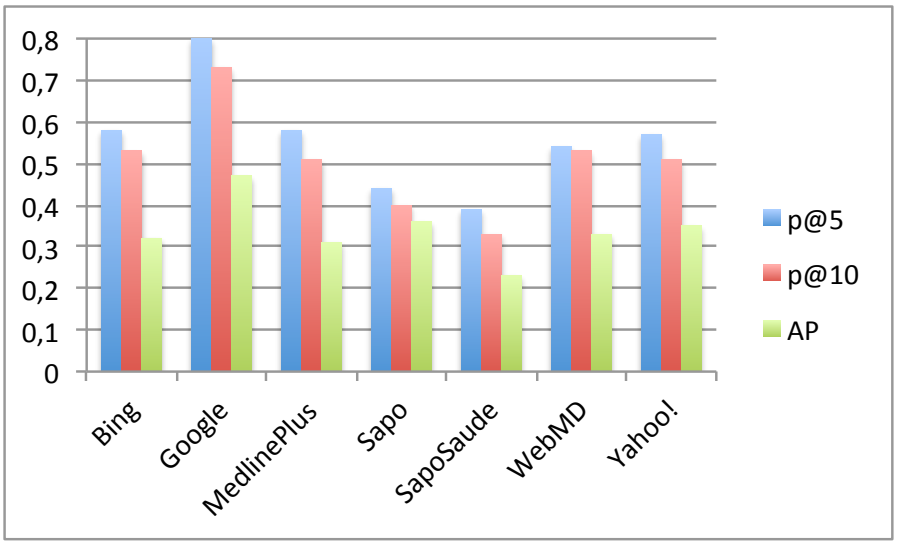

Figure 5.9: Average of non-graded measures in each search engine

both types of precision (p and gp). This proportion rises to more than 80% if we use more complex measures like GAP and AP. This is in line with the previously commented stability of these measures. From this analysis we can conclude that, in evaluations that use simple measures like precision at certain rank cut-offs with graded relevance assessments, it is more critical to have an appropriate threshold definition in graded precision. In our case, we think the first set of threshold probabilities ($g_1 = g_2 = 0.5$) is more sensible and genuine because it is defined over the space of users and considers the differences between them.

5.9 CONCLUSION

We have conducted a user study that allowed the evaluation of seven different search engines on the health domain. Four are generalist search engines and

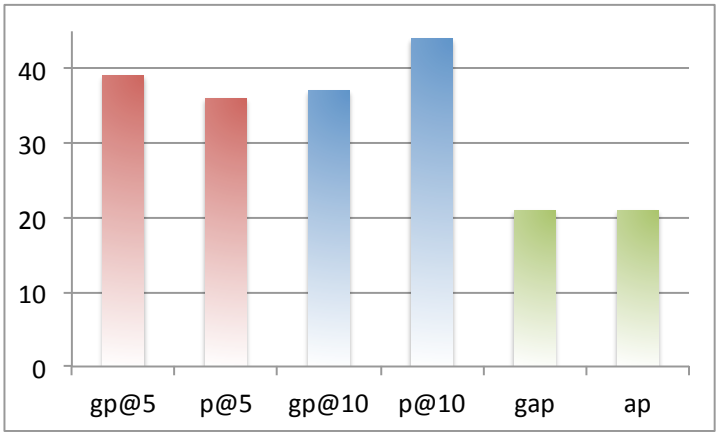

Figure 5.10: Number of significant differences in each measure

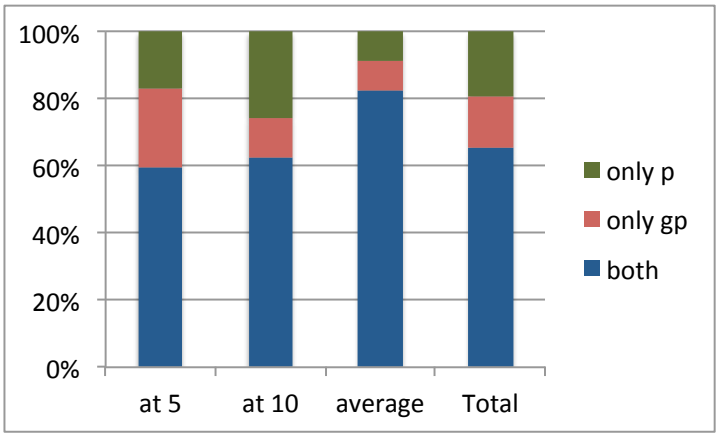

Figure 5.11: Proportion of types of significant differences found in each level

the others are health-specific. We have compared the precision of the search engines using 6 different measures in a global perspective and in specific types of information needs.

Our results show that, in precision, generalist search engines surpass health-specific ones. Google is users' preferred engine and it is also the one with better precision. The superiority of this engine is more expressive in the top of the rank which means Google's first results page is a good place to start a health search session. To reduce the bias towards severe topics, it might be a good practice to use a health-specific engine to refine results. In fact, health-specific engines seem more balanced in severity in their collections. The higher precision of severe conditions make us suspect there is more online information about severe topics than non-severe ones and this may lead to the escalation of medical concerns.

About measures, we found that complex measures like AP and GAP are less vulnerable to thresholds definition in graded precision. In evaluations using only simple measures like precision at certain rank cut-offs, it is important to have an adequate definition of these probabilities.

In the next chapter we compare different types of search engines, ones retrieving only health certified documents and others retrieving also non-certified documents. Moreover, we compare certified with non-certified documents

and also compare documents belonging to different health certification categories.

DATA CERTIFICATION IMPACT ON HEALTH INFORMATION RETRIEVAL

6.1 INTRODUCTION

The use of the Web to search for health information is gaining popularity among patients, their family and friends. The characteristics of the Web make it a medium where publishing is easy and accessible to everyone. This, allied with the impact that online health resources have on people's life and well being, emphasize the importance of mechanisms that help identify the quality of online health information. A Pew Internet report (Fox and Jones, 2009) found that "about one in ten online health inquiries have a major impact on someone's health care or the way they cared for someone else".

The problem of finding quality information exists since the first developments on information retrieval. Health domain specificities have triggered research initiatives parallel to the general ones. A systematic review of studies that assess the quality of health information for consumers on the Web has been done by Eysenbach et al. (2002). Initiatives like the Health on the Net Foundation Code of Conduct (HONcode) certification or the URAC's Health Web Site Accreditation Program have emerged to address the problem of health information quality. They both intend to help the user identify reliable and credible content through a seal that identifies the sites that satisfy their code of conduct or quality standards. HONcode certification is considered the most successful initiative (Baujard et al., 2011).

Typically, a search session starts in a generalist search engine instead of health-specific websites (Fox, 2006) and Google is commonly the chosen search engine (Schembri and Schober, 2009). Studies that compare the performance of generalist and health-specific search engines mostly conclude that the former outperform the latter. Regarding the quality of information, as reported in the previous chapter, some studies find that health-specific search engines provide higher quality contents while fewer conclude that quality is the same in both types of search engines.

With this context in mind, we conducted a user study to analyze the impact of limiting the collection of a search engine to certified health documents, having the HONcode certification as a base. This impact is measured in terms of precision, medical accuracy, documents' comprehension by users, documents' readability and users' motivational relevance. In the end, our findings indicate how medical certification can help generalist search engines provide a better service to their users in consumer health retrieval. A second goal of our study is to evaluate how useful are the HONcode categories for person-

alizing the search experience in a generalist engine. For example, we want to know if sites "for patients" are preferred to sites "for professionals" or if sites "for women" are actually more valued by women.

This chapter is structured as follows. After briefly explaining health information certification, we describe the user study in Section 6.3. Results are presented in Sections 6.4 and 6.5. The study's findings, along with their implications, are discussed in Section 6.6 and the conclusions follow in Section 6.7.

6.2 HEALTH INFORMATION CERTIFICATION

As previously said, health websites may be certified by external entities that assure that every site that has a certification seal respects a certain code of conduct. There are two widely known certification programs, one promoted by the Health on the Net Foundation (HON) and the other promoted by URAC. They are both non-profit organizations and they differ in scope. URAC intends to promote health quality in a global way, not only through the quality of online information as is the case with HON.

The URAC Health Web Site Accreditation Program evaluates websites against 48 quality standards[1]. A search of URAC accredited companies on their web site returns only 19 records. This confirms the greater popularity of the HONcode certification program that has 7,200 HONcode certified websites (Baujard et al., 2011).

Details about the HONcode certification system can be found on their website[2]. Briefly, any health site can request the certification, free of charge, whether or not it has a health focus. Requests are examined by a committee including health professionals that verifies if all the HONcode ethical principles are respected. The ethical principles are: authority, complementarity, confidentiality, attribution, justifiability, transparency, financial disclosure and advertising. A certified website is subjected to regular monitoring.

6.3 CASE STUDY

This is the only study of this Part of the dissertation that has not been based on the experiment described in Chapter 4. Instead, it is based on the experiment that is described in Chapter 8. Since Chapter 8 ensues this chapter, in this section we briefly describe the experiment that allowed the analysis presented in this chapter.

Our user study involved 40 undergraduate students (25 females, 15 males) of a programme in Information Science. Users are medically lay people and have a mean age of 22.25 years (sd = 6.42). We defined 8 information situations[3] based on questions submitted to the health category of the Yahoo! Answers service. Each information situation requires finding a treatment for a particular disease or condition and is associated with 4 different queries formulated by the researchers. We have used Google as a black-box search engine with two collections, Google's entire collection and Google's indexed webpages

[1] Available at: http://www.urac.org/docs/programs/URACHW2.1factsheet.pdf
[2] Available at: http://www.hon.ch/HONcode/Patients/Visitor/visitor.html
[3] Available in Chapter 8.

with HONcode certification. We filtered the collection through Google custom search, a tool provided by Google in which it is possible to create custom search engines that work with specific sets of websites or webpages. Henceforward, we will call WebSys to the system working with the first collection and HONSys to the system working with the HONcode certified collection.

For each query and system, we collected the top-30 results. To reduce the risk of Google learning from the previous submitted queries, we ensured that returned links were never clicked. Further, to prevent changes in the search engine or in the HON collection, we submitted all queries within a very short time span.

A query run on one of the retrieval systems leads to a task that a user can execute. Each user was assigned a set of 8 different tasks in which he had to assess, in a 3-value scale, the relevance and comprehension of the top-30 documents and to answer a post-search questionnaire. A Latin-square like procedure was adopted during task assignment to guarantee that each user assessed the relevance of every information situation and was exposed to each retrieval system the same number of times. We have also guaranteed that each system is associated with each information situation the same number of times. To prevent possible bias owing to human behavior, we have also permuted the order of tasks and forced users to complete them in the prescribed order. Additionally, to preempt users' fatigue, each task had to be performed in different days, that is, tasks had to be separated by an interval of, at least, 24 hours. Users did not have time limits to perform each task.

6.4 IMPACT ANALYSIS

Our analysis will focus on three aspects: comparison of search systems, comparison of certified and non-certified documents and comparison of shared and non-shared documents. A shared document is a document that is retrieved by both systems.

Henceforward, we will use * and ** to sign significant results at the levels of significance (α) of 0.05 and 0.01, respectively. Additionally the following nomenclature will be used to identify information about hypothesis tests: $\chi^2(\mathrm{df})$ corresponds to the Chi-square distribution with df degrees of freedom; W is the statistic calculated for the Wilcoxon rank-sum test, a non-parametric test for assessing whether two independent samples of observations have equally large values; $\mathrm{t}(\mathrm{df})$ is the Student's t distribution with df degrees of freedom; $\mathrm{F}(\mathrm{df})$ is the F-distribution with df degrees of freedom used in the ANOVA test to compare the means of several distributions; TukeyHSD is presented before the confidence limits of the Tukey's Honestly Significance Difference test, used in multiple comparisons after the ANOVA test; and p is the abbreviation of *p-value*.

6.4.1 Precision

To analyze precision we use the Graded Average Precision (GAP) and Graded Precision (gP) with an equally balanced g_1 and g_2. These measures are described in Section 5.2.1.

As can be seen in Figure 6.1, in terms of precision, the HONSys had a worse performance in every measure: GAP, gP@5 and gP@10. As expected,

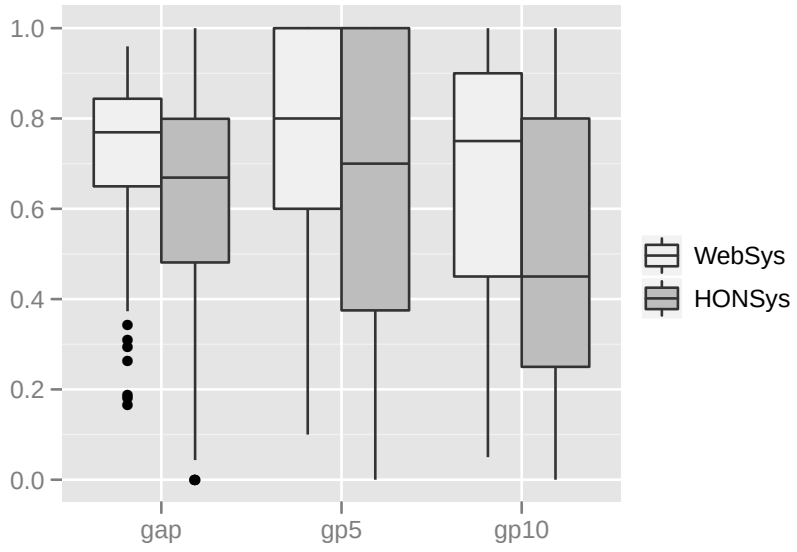

Figure 6.1: GAP, gP5 and gP10 boxplots on both systems.

the statistical dispersion is lower with GAP, since this is an average measure that considers the top-30 results. Differences between systems are significant in every measure at $\alpha = 0.01$ which means that users prefer the WebSys even including non-certified documents.

In the WebSys, we computed the correlation between each measure and the proportion of certified documents in each session (for GAP), in each top-5's session (for gP5) and in each top-10's session (for gP10). In GAP and gP10, the correlation is approximately 0.18 and, at $\alpha = 0.05$, is significantly higher than 0. The almost null correlation in gP5 and the low correlation values in gP10 and GAP make us believe that the HONCode certification is not a major factor influencing relevance assessments, mainly in the top-ranked results.

Since GAP, gP5 and gP10 are measures that evaluate the performance of a set of documents, we cannot use them to compare certified with non-certified documents. For that reason, we will compare these two sets of documents using documents' individual relevance assessments. In the WebSys, the non-certified documents (column *No* in Figure 6.2) are almost equally distributed in terms of relevance assessments having, each level of the relevance scale, about 33% of the non-certified documents. Since the number of documents in each category presented in the x-axis of Figure 6.2 is variable, in the y-axis we plotted the proportion of not-relevant, partially relevant and totally relevant documents instead of the documents' counting. In certified documents, the proportion of non-relevant documents is much lower (27%) and significantly lower than partially relevant ($\chi^2(1) = 11.89$, p=3e-04**) and totally relevant documents ($\chi^2(1) = 19.1$, p=6.24e-06**). In WebSys certified documents, the most likely is to find a *totally relevant* document (38%). We also conclude that, in WebSys, certified documents are associated with higher relevance scores than non-certified ones. In fact, the proportion of *not relevant* documents is higher in WebSys non-certified documents ($\chi^2(1) = 13.54$, p=1e-04**) and the proportion of *totally relevant* documents is higher in WebSys certified docu-

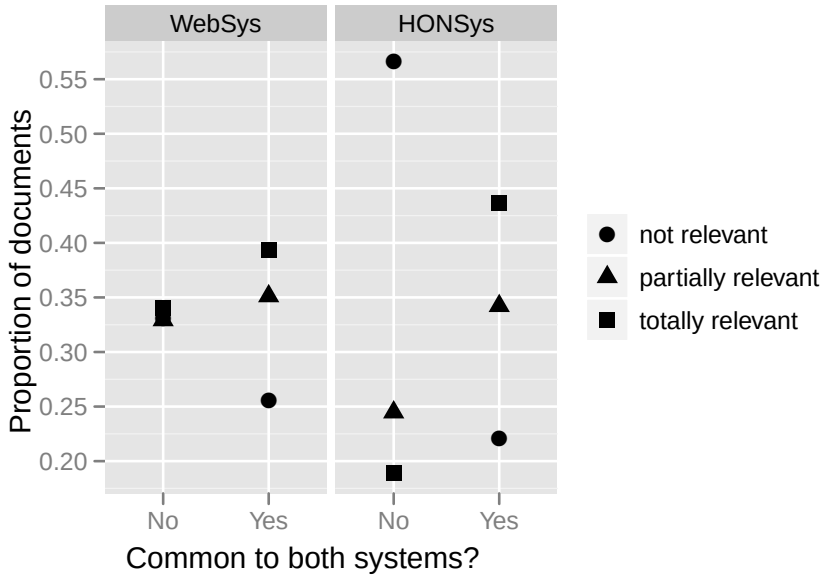

Figure 6.2: Proportion of documents by search system, share status and relevance assessment.

ments ($\chi^2(1) = 6.47$, p=5e-03**).

In the HONSys, all documents are certified but we can distinguish two groups of documents, the ones that are also retrieved by the WebSys and the ones that are not. As can be seen in Figure 6.2, these two groups of documents have an opposite behavior in terms of relevance. When both groups are compared, shared documents are associated with a significantly higher relevance, expressed by a lower proportion of *not relevant* documents ($\chi^2(1) = 283.37$, p<2.2e-16**) and a higher proportion of *partially relevant* ($\chi^2(1) = 28.77$, p=4.07e-08**) and *totally relevant* ($\chi^2(1) = 203.78$, p<2.2e-16**) documents.

It is also interesting to note that, although having a similar pattern, shared documents are assessed with a higher relevance on HONSys than on WebSys, perhaps influenced by the relative comparison to the other results. Comparing rank positions of shared documents in both systems we notice that, as expected, these documents appear first in HONSys ranks (rank median of 4) than in WebSys' ranks (rank median of 11). This difference is statistically significant (W=304150, p<2.2e-16**). Certified documents appearing in the top-30 in the WebSys have characteristics beyond certification that distinguish them from the other certified documents that do not appear on these top-30 and appear on the HONSys top-30 ranks. These characteristics are intimately related with Google's criteria for ranking. Comparing the results of certified and non-certified documents on WebSys, we can conclude that the certification is a criterion that should be integrated in the set of criteria used by search engines.

6.4.2 Medical accuracy

After each task, users were asked to write the treatment(s) they found for the condition mentioned in the information situation. Each answer was evaluated by a medical doctor regarding their correct and incorrect contents. The

combination of these two measures leads to a variable which we named *medical accuracy* that varies between 0 (lowest accuracy) and 4 (highest accuracy). The median of the medical accuracy on the HONSys is significantly higher than the one on the WebSys (W = 11326, p=0.03*). An analysis by answers' correctness and incorrectness shows that, in terms of correct contents, both systems have a similar behavior. Both systems have a median of 1 (*answer with some value*) and have no significant proportion differences in each level of the correctness scale. In terms of incorrect contents, the WebSys leads to more incorrect answers than the HONSys (W = 10815.5, p = 0.004**). The proportion of answers classified with *some incorrect content* is significantly higher in the WebSys (39% against 28% - $\chi^2(1) = 3.59$, p=0.03*) and the proportion of answers with *no incorrect content* is significantly higher in the HONSys (57.5% against 42% - $\chi^2(1) = 7.2$, p=0.004**).

Based on the previous finding, we investigated if, on the WebSys, the medical accuracy, correctness and incorrectness of contents increase with the number of certified pages. In terms of medical accuracy and correct contents, we detected no significant differences and no pattern inline with our hypothesis. In terms of incorrect contents, we found significant differences in the mean number of certified pages between levels of the incorrectness scale (F(2)=3.63, p=0.03*). Surprisingly, a pairwise comparison showed that sessions of answers with *no incorrect content* have less certified pages than sessions of answers with *some incorrect content* (TukeyHSD: (-2.61; -0.04), p=0.04*). This is the opposite of what we expected and, assuming there are no incorrect contents in certified documents, we conclude there are several non-certified documents that have *no incorrect contents* and suspect that a few documents with incorrect contents have the power to damage the knowledge acquired in the overall search session. This strengthens our previous conclusion that certification must occupy a prominent place in the set of criteria used by search engines.

6.4.3 Readability

Documents readability was automatically evaluated using the Simple Measure of Gobbledygook (SMOG) metric. A higher SMOG means the document contains more polysyllables and is, therefore, more difficult to read. We found that documents retrieved by HONSys (mean SMOG of 7.55) are more complex (W = 9287206, p=0.001**) than WebSys documents (mean SMOG of 7.38). However, if we make this comparison by certification status, we find that non-certified documents have a higher SMOG mean (7.46 against 7.4). Although this last difference is not significant, these two results show a contradictory trend. We also compared the SMOG mean according to the system and the URL share status (Table 6.1). In the WebSys, the non-certified documents have a mean SMOG higher than the one in certified documents, a difference that is statistically significant (W=1413680, p=8.88e-16**). In the HONSys, documents that are also retrieved by the WebSys have a lower SMOG when compared to non-shared ones, a significant difference (t(911.76)=6.0029, p=1.4e-09**) that evidences that document's readability may be used by Google to rank documents. Since shared documents appear in the HONSys rank upper than in the WebSys rank (rank median of 4 against rank median of 11), we conclude this criterion predominates in smaller collections, like the collection of only-certified documents, where other criteria may probably not be met.

Table 6.1: Mean SMOG by system and share status.

URL	WebSys	HONSys
Shared	6.9	6.84
Not shared	7.46	7.59

6.4.4 *Comprehension*

The comprehension of the documents was assessed by the users during the search task using a 3-value scale: 0 - *I did not understand*, 1 - *I partially understood* and 2 - *I totally understood*. We found that users understand better the documents retrieved by the WebSys than the HONSys documents. In fact, the former system has a significantly lower proportion of *not understood* URL ($\chi^2(1) = 15.07$, p=5.18e-05**) and a significantly higher proportion of *totally understood* URL ($\chi^2(1) = 10.98$, p=5e-04**). Although the complexity of the text is not the only factor affecting the comprehension of a document, this is in agreement with the readability results reported in the previous section. An analysis by documents' certification status revealed that non-certified documents are better understood by users when compared to certified ones. The former have a significantly lower proportion *not understood* assessments ($\chi^2(1) = 7.16$, p=4e-03**) and a significantly higher proportion of *totally understood* assessments ($\chi^2(1) = 4.78$, p=0.01*). Specifically in the WebSys, we detected no significant differences between certified and non-certified documents comprehension. In the HONSys, shared documents have a significantly lower proportion of *not understood* classifications when compared with non-shared documents ($\chi^2(1) = 10.89$, p=5e-04**). This is a sign that shared documents are better understood by users and is inline with the readability results.

6.4.5 *Motivational relevance*

After the search task, users evaluated their degree of satisfaction with the task in a scale of 1 (*I did not succeed in this task*) to 5 (*I completely succeed in this task*). We compared both retrieval systems and found that the WebSys is associated with a higher degree of satisfaction (W=14265, p= 0.03*). In the WebSys, we did not find significant differences on the mean number of certified pages between the 5 levels of satisfaction.

6.5 CONTEXTUAL ANALYSIS

During the HON certification process, websites are classified according to their purposes. Some of the categories are: "for health professionals", "for patients", "for women" and "for men". In this section we compare users' relevance and comprehension assessments in the four categories mentioned above. In the two last categories, we also consider the users' characteristics. Additionally, in the first two categories, we compare documents' readability.

The categories "for health professionals" and "for patients" are not mutually exclusive, with some documents being classified for both audiences. There

Table 6.2: Proportion tests performed by level of comprehension, relevance and the membership to the Professional (P) HON Category. n= not. $\chi^2(1)$ value in parenthesis.

Level	Comprehension	Relevance
0	nP<P** (11.28)	nP>P** (8.92)
1	nP<P** (86.46)	nP<P** (21.94)
2	nP>P** (115.38)	nP>P (1.74)

Table 6.3: Proportion tests performed by level of comprehension, relevance and the membership to the Consumer (C) HON Category. n= not. $\chi^2(1)$ value in parenthesis.

Level	Comprehension	Relevance
0	nC>C** (56.6)	nC>C** (22.93)
1	nC>C** (34.78)	nC<C* (3.43)
2	nC<C** (88.89)	nC<C** (12.85)

are also documents that do not belong to any of these categories. In our sample there are 118 URL for health professionals and 456 directed to consumers. We have applied several proportion tests to verify if these documents differ in terms of comprehension and relevance scores. In Tables 6.2 and 6.3, for each level of comprehension and relevance, we present the proportion difference found (< or >) along with its significance and test value. For example, regarding comprehension, in level 0 we found that "nP<P", i.e., documents "for health professionals" (P) have a higher proportion of "not understood" (level 0) ratings than documents not belonging to the "health professionals" group (nP).

Through the results presented in Tables 6.2 and 6.3, we conclude that documents that are not "for health professionals" (nP) are better understood by users than the ones that are. The "for health professionals" category has a smaller proportion of documents that are not or are partially understood and a larger proportion of totally understood documents. On the other hand, documents belonging to the "for patients" (C) category are better understood than the ones that do not (nC). The behavior of the "for patients" category is similar to the one presented above for the documents that do not belong in the "for health professionals" category.

Regarding readability, the significant differences we found between SMOG means agree with the previous results. We found that documents "for health professionals" are more complex (W=1338060, p=2.55e-07**) than documents that do not belong to this group just like documents that are not "for patients" (t(3116.6)=5.9, p=1.86e-09**) in comparison with documents in the "for patients" group.

In terms of relevance, documents "for patients" are more relevant than the documents that are not associated with this category. On the "for health professionals" category, the behavior is opposite but less clear. The professional documents seem to be more relevant than the documents that do not belong

to this category.

In our sample there are 60 assessments of 8 documents "for women" and 50 assessments of 6 documents "for men". In these categories we do an analysis similar to the previous one but we also consider the gender of the user. In terms of comprehension, we found that documents "for women" are better understood by the general user and, more specifically, by the women. In fact, they have a lower proportion of documents classified with 1 ($\chi^2(1) = 4.54$, p=0.02* in the general user and $\chi^2(1) = 3.71$, p=0.03* in the women) and a higher proportion of documents assessed with 2 ($\chi^2(1) = 7.75$, p=0.003** in the general user and $\chi^2(1) = 5.94$, p=0.007** in the women). In terms of relevance, the only significant result we found is that men assess documents "for men" as *not relevant* more frequently than they do in documents not classified as "for men" ($\chi^2(1) = 5.56$, p=0.009**).

6.6 DISCUSSION

In the comparison between retrieval systems, the one having the Web as a collection, including certified and non-certified documents, has a better performance in every aspect but medical accuracy. Users assess relevance higher in this system, understand better its documents, feel more satisfied after the search sessions and its documents are less difficult to read. However, this system is associated with more incorrect contents than the one only including certified documents. In this matter, we found that this difference is not due to the higher number of non-certified pages in this system because the proportion of certified documents in sessions with *some incorrect content* is significantly higher than sessions with *no incorrect contents*. Assuming there are no incorrect contents in certified documents, either the problem is on the comprehension of certain documents or in a few non-certified documents that may have the power to damage the knowledge acquired in the overall search session. We found that non-certified documents are better understood than certified ones. Since readability is not significantly different between both types of documents, the comprehension differences may be related with the existence of medical concepts that are not apprehended. We also found that certification does not affect relevance assessments, mainly in the top-rank results. This may be justified by a previous study (Fox, 2006) finding: "three-quarters of health seekers do not consistently check the source and date of the health information they find online".

A deeper analysis was done in three groups of documents: non-certified documents (WebSys non-shared documents), certified documents retrieved by both WebSys and HONSys (shared documents) and certified documents retrieved only by HONSys (HONSys non-shared documents). We found that shared documents are the ones with better performance in terms of relevance and readability. This is not strange since this set of documents meets the general criteria of Google search engine and follows the quality standards defined in the HONCode. In terms of comprehension, these documents are better understood than HONSys non-shared ones but have no significant differences when compared with WebSys non-shared ones. As expected, shared documents rank higher in the HONSys than in the WebSys. After this set of documents, the ones with better performance in relevance, readability and com-

prehension are the WebSys non-shared documents.

In the contextual analysis we also considered the HONCode categories. We found that documents "for patients" and documents that are not "for health professionals" are easier to read and understand. The readability of a document can be a good evidence of its adjustment to health consumers. In terms of relevance, users find documents "for patients" and "for health professionals" more relevant. Although users understand worse the latter type of documents, we believe these documents convey a professionalism and confidence that makes users rate their relevance higher. In addition, documents "for women" are better understood by the general user and, more specifically, by the women.

6.7 CONCLUSION

In this work we analyze the impact of medical certification on several aspects of health information retrieval performance. We conclude that users value the diversity provided by generalist search engines even if this means including non-certified documents. Yet, we found that the medical accuracy of generalist search engines may be in risk if users do not understand documents or if the session has a few documents with unreliable information. As we have seen, to assure the comprehension of the documents, besides their readability, engines must also guarantee that document terminology is adjusted to the users' knowledge. To improve the performance of generalist search engines on health tasks and to assure the credibility of the top results, the ones receiving more attention, it is advisable to incorporate the medical certification in the set of criteria currently in use by the search engine. As we have shown, the documents retrieved by the WebSys with HON certification are the ones with the best overall performance. Supported by findings of previous studies and the fact that certification has no impact on users' relevance judgments, we have reasons to believe that health consumers do not consistently check if the health information they find is certified. Since users' unawareness of information reliability may be associated with some dangers, the inclusion of medical certification in the set of search engines' criteria becomes even more important.

We also concluded that the classification of documents as "for patients" and "for health professionals" might be useful to personalize the search experience. This categorization might be useful to identify documents characteristics that, later, could be used to automatically classify documents in these two categories. On the other hand, there is also the need to predict if the user is a layperson or a professional and his level of expertise on the topic. Since documents' readability is tightly connected with the HONCode categorization and it proved to be discriminating in several comparisons, it may be a good indicator of documents that are valued and understood by users.

In the following chapter we describe the second study conducted with the data collected in the experiment described in Chapter 4. In that study we examine how user, task and document features are related with user's behavior, in search tasks of several types. User's behavior is analyzed in terms of query formulation and relevance assessment.

CONTEXT EFFECT ON QUERY FORMULATION AND SUBJECTIVE RELEVANCE IN HEALTH SEARCHES

7.1 INTRODUCTION

Several authors agree that context, often ignored, might be used to improve the retrieval process (Bierig and Göker, 2006; Ingwersen et al., 2005). As we have shown in Chapter 3, context is a loose concept and is defined in the literature in many different ways. Dey and Abowd (2000) present a comprehensive definition, describing context as: "any information that can be used to characterize the situation of entities (e.g. a person, a place or an object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves". Here, context is considered an interactional problem, as defined by Dourish (2004). It not only includes the environmental features surrounding the user and his activities, but also the interaction in which he is involved. We believe context is dynamic and might change each time a new search is made, a new set of results is browsed or a new document is viewed (Harper and Kelly, 2006).

In the retrieval process, the interaction of the user with the system is concentrated in the formulation of the query and in the relevance assessment of the retrieved documents. With a large human involvement, we expect these stages to be largely influenced by context. Understanding how context affects the formulation of queries can help delineate new ways, with or without the user intervention, to improve the queries. On the other hand, it is crucial to comprehend what factors affect relevance judgments, in which ways and how can these be incorporated in IR systems. These factors would certainly be useful as an input to algorithms that match information needs and documents and to help IR systems move to a concept of relevance that encompasses the search context. Also, these features can be used to improve existing interfaces, either in the first stage where the user transmits the system his information need or in the latest stage, in which he accesses the retrieved documents.

There is an increasing tendency of patients, their family and friends to use the Web to search for health information and, according to Lin and Fushman (2005), this domain is extremely rich and "very well-suited for experiments in building richer models of the information seeking process".

This work intends to analyze the influence of user and task features on the formulation of queries. Moreover, it also analyzes how the above features together with query and document features affect the relevance assessment stage.

The work presented here is based on the user study described in Chapter 4. We focus on user features like age, gender, health status, web search experience, health search experience and familiarity with the topic. Regarding task features, we focused on its clarity and easiness and also on its medical specialty and clinical type (e.g. diagnosis, treatment).

This work is broader than the existing research on the influence of context features on query formulation. On the one hand it covers context features not explored before, like the health-specific ones. On the other hand, existing research is mainly focused on user expertise and type of search (e.g. exploratory, fact-finding). When compared with research that explores relevance judgments, this work is innovative because it is not based on criteria explicitly gathered from users but on implicitly gathered characteristics. Existing research is essentially based on users' explicit descriptions of what affects their relevance judgments. As users have often difficulty discussing their criteria (Ingwersen and Järvelin, 2005), we feel implicit methods might give different insights.

In the two following sections we describe the main research done in query formulation and relevance assessment in IR. Context influence is then analyzed in two stages: query formulation and relevance assessment. Section 7.4 is focused on query formulation according to the query language, the use of advanced and boolean operators, the use of medico-scientific terminology and the number of terms. Regarding relevance, that section gives more emphasis on motivational relevance, assessed through users self-evaluation of web search success and health search success. In Section 7.5 we analyze users relevance assessments by categories of context features. This section focuses on situational relevance, evaluated through users relevance assessments. It also compares both types of relevance. In Section 7.6 we discuss the results described in the previous sections and, in Section 7.7, we present our conclusions.

7.2 QUERY FORMULATION IN IR

Query formulation is the process of transforming an information need into a request according to the rules of the IR system. When communicating, humans are influenced by their previous experiences and their social, organizational and cultural environment (Ingwersen and Järvelin, 2005). Inevitably, the same happens when queries are formulated to express information needs.

Research in query formulation is usually based on analysis of log files and is traditionally more quantitative. Jansen and Pooch (2001) do a good review of studies focused on web search and report that queries are often short, having only 1 or 2 terms and lack structure and language operators. Only 9% of the queries use advanced operators and only 8% use boolean operators.

Research exploring context features affecting web search is not abundant and often ignores features related to the user, the task or the concepts presented in the query (Aula, 2003). In the existing studies, the most examined features are the user's expertise and the type of search.

Aula (2003) conducted a user study to analyze which factors affect query formulation in web search and grouped them in three main classes: media expertise (e.g. computer, Web, search engine), domain expertise and type of search task (fact-finding, exploratory and comprehensive). In her study, me-

dia expertise is correlated with more precise and longer queries and domain expertise presumably leads to higher quality terms in queries. In fact-finding search tasks, precision is an important measure of success and, therefore, the use of precise terms or phrases is usually a good strategy. In exploratory tasks, simple queries may be enough as the goal is to obtain a general idea of the search topic and not to have high recall and precision. On the other hand, on comprehensive search tasks, a high recall is expected and a good strategy involves the use of broader terms and manual truncation.

7.3 RELEVANCE IN IR

The main goal of any IR system has always been the retrieval of *relevant* information. The concept of relevance is recognized as a central concern of any IR system and is related to the perceived topicality, pertinence or usefulness of documents to a particular information situation. After a large interest in the 1960s and 1970s (Ingwersen and Järvelin, 2005), research has been stimulated again in 1990s with the work of Schamber (1990).

Three insightful reviews of research on relevance are done by Saracevic in three parts (Saracevic, 1975, 2007a,b), by Borlund (2003a) and by Ingwersen and Järvelin (2005). The section *Effects of Relevance: What Influences are Related to Relevance Judges and Judgments* in the work of Saracevic (2007b) is particularly pertinent as a literature review of the work reported here. For this reason, we only describe the most relevant concepts and research works.

7.3.1 Nature of relevance

Borlund (2003a) describes relevance as multidimensional and dynamic. It is multidimensional because it depends on the perceptions and assessments of different users and it is dynamic because it changes over time for the same user. This study only focuses on the exploration of the multidimensionality characteristic of relevance. Research in this area has been focused on the identification of the criteria used to judge the relevance of a document. A study of Schamber (1994) identifies 80 criteria as a reasonable sample of the factors used to judge relevance. Barry (1994) founds 23 criteria that were grouped in 7 categories, including the characteristics of the documents, user's previous experience, user's preferences and user's situation. The first work is a review of others' work and, in the second, users are explicitly asked to explain the rationale for the relevance assessment in an interview.

7.3.2 Types of relevance

Relevance can be of two main types: objective/system-based relevance and subjective/user-based relevance (Borlund, 2003a). The first is described by Saracevic (1996) as the relation between a query and a document in an IR system and it is considered independent of the user, it just depends on the characteristics of the documents. IR systems are mainly based on this type of relevance because it is objective, stable and it has an easier implementation in automatic systems. This is also the concept used by the mainstream method of evaluating IR systems that incorporates a document collection, a set of re-

quests and a set of relevance assessments, ignoring the user and his subjacent tasks.

The subjective relevance is user and context dependent and is divided by Saracevic (1996) in four major categories: topical, pertinence, situational and motivational. Topical relevance is associated with *aboutness*, that is, the relation between the topic expressed in a query and the topic expressed in a document. This type of relevance involves an assessment of the topic related to a query and a document. Pertinence is the relation between the information need and the documents, taking into account the user's cognitive state and knowledge at the moment. This is especially significant in health information retrieval done by consumers, in which the document's medical terminology has to be adequate to the user's knowledge to be considered relevant. Situational relevance is expressed by the usefulness of the information objects to the user's work task. Motivational relevance relates the user's goals and motivations with the information objects. It is expressed by the user's feeling of success and his satisfaction.

We believe that a system that incorporates features representing "persons and their interpretations/perceptions, work tasks, interaction, situations and contexts" (Ingwersen and Järvelin, 2005) is more realistic and, therefore, we focus on subjective types of relevance. More specifically, we focus on situational relevance, because the study involves the user and also his interpretations of the work tasks.

7.3.3 Values of relevance

The scales of relevance used to judge documents are typically of two types: binary and non-binary. Binary scales are closely associated with traditional evaluation methods of IR systems using the Cranfield model. In these evaluations, documents are usually judged as *relevant* or *non-relevant*.

On the other hand, non-binary scales are more common on user-oriented IR research, becoming popular in the 1990s (Ingwersen and Järvelin, 2005). The number of rating values in non-binary scales differs from study to study (e.g. 11-points, 7-points, 3-points). The 3-points scale, used in this study, is the most used in IR experiments (Borlund, 2003a) and usually describes categories as: relevant, partially relevant and non-relevant.

7.4 QUERY ANALYSIS

The experiment described in Chapter 4 served as basis for this analysis.

Queries formulated by the users are analyzed in four perspectives: the language of the query terms, the use of advanced and boolean operators, the use of medico-scientific terminology and the number of terms. The language of the query was manually labeled and the use of medico-scientific terminology was identified based on the multilingual glossary of technical and lay medical terms described in Chapter 2. The analysis was done according to some of the context dimensions presented in Table 4.3, namely, user, web search, health search, topic's familiarity and task. From these dimensions, excluding the `hstatus`, `usual-engine` and `taskstat`, it considers the all their context features included.

The language, use of advanced and boolean operators and use of medico-scientific terminology are all nominal variables. Therefore, we follow the strategy presented in Figure 7.1 in these three dimensions. We compare the distributions of the variables belonging to the groups mentioned above in the two categories defined by the nominal variables (e.g. Portuguese and English in the language variable). Using the one-tailed test in nominal and dichotomous variables, we are able to detect the direction of the differences (e.g. higher or lower).

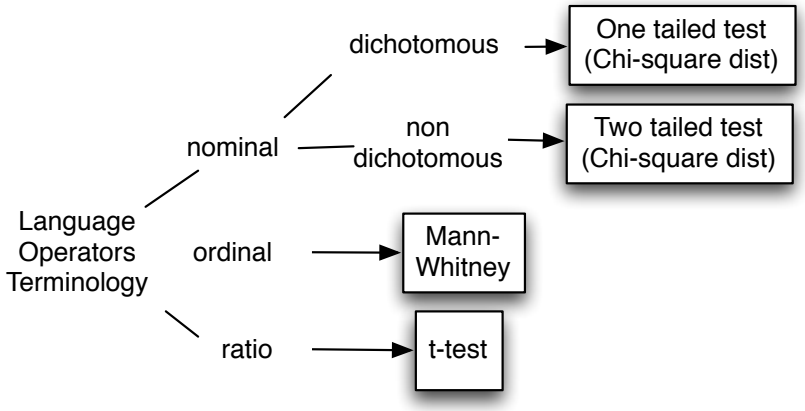

Figure 7.1: Statistical analysis of the language, operators and terminology variables.

The strategy to analyze the impact of context features on the number of terms is presented in Figure 7.2. It is different because the number of terms is a ratio variable.

We compare the average number of terms in the groups defined by nominal and ordinal variables and analyze its correlation with ratio variables. Whenever we found significant differences with the Kruskal-Wallis test, we also did a pairwise comparison in which we have divided the α value by the total number of comparisons.

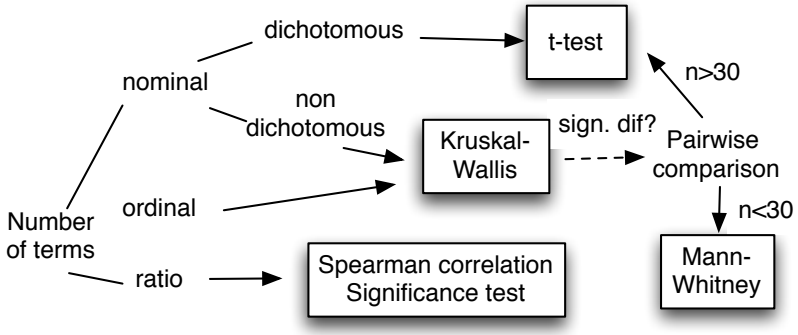

Figure 7.2: Statistical analysis of the number of terms variable.

7.4.1 Global analysis

In the conducted experiment, users issued a total of 155 different queries. User's first language, Portuguese, was used in 76% of the search sessions and English in the remaining sessions. Each user was involved in 8 search sessions, 4 for each information need. A deeper analysis shows us that all search sessions in Medline and WebMD were made in English. In Yahoo, 92% of the search sessions were made in Portuguese and, in the other search engines, Portuguese was the preferred language. This suggests that, in most cases, the use of the English language might not have been a user's choice but an imposition of the selected search engine. Only 17% of the queries used advanced or boolean operators and the average number of terms was 3.78 (SD = 2.01). The majority of the search sessions are associated with 2 (19%), 3 (37%) or 4 terms (19%). Only 3% of the queries used medico-scientific terminology. The proportion of structured queries found is similar to the one reported by Jansen and Pooch (2001) and the average number of terms is slightly superior.

7.4.2 Language

The global analysis of the language used in queries showed that users tend to search in Portuguese, only showing a different behavior when using search engines with contents in other languages. Yet, we decided to further analyze the influence of context features on the choice of language because, in Yahoo!, some users opted for the English language.

In the column *Language* of Table 7.1, we can see that female users have a higher proportion of search sessions in English than the male users. Through the information in Table 7.2, we conclude that users employ English queries when the topic is more familiar and the task is clearer.

We detected that, in users that use the Web more often to search for information (`ws_freq`), there is a tendency to use English more frequently but the difference between languages is not significant.

In the initial questionnaire users were inquired about their preferred language in web searches. Although this is a variable that is not explored in this study, we were curious to know if a systematic use of English leads to more successful web searches (`ws_success`). We found that every user that always find what he looks for (5 in `ws_success`), routinely use English in their web searches. However there was no statistical evidence of this (Mann-Whitney U=6912, p=0.06).

7.4.3 Advanced and boolean operators

There is statistical evidence to conclude older users employ advanced or boolean operators more often (Table 7.3). In Table 7.2 we can see that users that don't use advanced and boolean operators are associated with fewer web searches and a lower web search success rate. This suggests that, as the experience in web search increases, users apply more structured queries, an habit that is associated with a higher rate of success. The same habit also affects positively the success rate of health searches.

Users who have made previous searches on the topic use more advanced and boolean operators (Table 7.1). There is also evidence to state there is an association between the use of operators and the medical specialty. A higher

Table 7.1: Context effects of nominal variables: Chi-square test results. *p<.05; **p<.01. Question mark represents a Chi-square approximation that may be incorrect. Proportions as $p_{row}(column)$.

Var	Language	Operators	Terminology
gender	$p_f(en) > p_m(en)$ $\chi^2(1)$=12.68 p=0.00** (>)	$p_f(y) < p_m(y)$ $\chi^2(1)$= 0.26 p=0.31 (<)	$p_f(y) > p_m(y)$ $\chi^2(1)$= 1.19? p=0.14 (>)
hs_wuse	$p_n(en) < p_y(en)$ $\chi^2(1)$=1.05 p=0.15 (<)	$p_n(y) > p_y(y)$ $\chi^2(1)$=0.33 p=0.28 (>)	$p_n(y) > p_y(y)$ $\chi^2(1)$=4.78? p=0.01* (>)
prev_search	$p_n(en) < p_y(en)$ $\chi^2(1)$=2.46 p=0.06 (<)	$p_n(y) < p_y(y)$ $\chi^2(1)$=16.19 p=0.00** (<)	$p_n(y) < p_y(y)$ $\chi^2(1)$=13.98? p=0.00** (<)
qtype	$\chi^2(5)$=2.24? p=0.81	$\chi^2(5)$=10.95? p=0.05	$\chi^2(5)$=7.53? p=0.18
speciality	$\chi^2(3)$=5.17 p=0.16	$\chi^2(3)$=35.66 p=0.00**	$\chi^2(3)$=4.38? p=0.22

proportion of gynecology tasks (43%) are associated with structured queries and no urology tasks used advanced operators. In Table 7.2, we see that structured queries are associated with clearer and more difficult tasks.

7.4.4 Use of medico-scientific terminology

Since only five queries, formulated by two users, employed medico-scientific terminology, results reported in this section do not have the same statistical strength, particularly in the Chi-square tests where the high number of cells with expected values lower than 5 amplifies the test value. When compared with the familiarity and task's variables, user, web search and health search variables have even less statistical meaning. Although aware of this situation, we decided to present the results of our analysis because they may lead to new research hypothesis that may be studied later.

The reduced number of queries with medico-scientific terminology is, by itself, an indicator of its lack of use by health consumers. This reality might be explored in query expansion whenever users' literacy is adjusted to this type of terminology.

Contrary to our expectations, results show that the use of medico-scientific terminology might be related to a smaller frequency of health searches (`hs_freq`). Results also suggest the use of medico-scientific terminology might be associated with more successful health searches (`hs_success`). In Table 7.2 we can see that queries with medico-scientific terminology are associated with more familiar (`prev_search`, `familiarity`) and clear tasks.

Table 7.2: Context effects of ordinal variables: median and Mann-Whitney U test results. *p<.05; **p<.01. Signs > and < indicate one-tailed tests.

Var	Language	Operators	Terminology
clarity	EN: 5, PT: 4 U=10375.5 p=0.00**(>)	N: 4, Y: 5 U=2929.5 p=0.00**(<)	N: 4, Y: 5 U=488 p=0.03*(<)
easiness	EN: 3, PT: 3 U=8219.5 p=0.28(<)	N: 3, Y: 2 U=6307.5 p=0.00**(>)	N: 3, Y: 2 U=1086 p=0.13(>)
familiarity	EN: 3, PT: 3 U=10039.5 p=0.00**(>)	N: 3, Y: 3 U=4521.5 p=0.28(<)	N: 3, Y: 4 U=378 p=0.00**(<)
hs_freq	EN: 1, PT: 1 U=6311.5 p=0.25 (<)	N: 1, Y: 1 U=2983.5 p=0.11(<)	N: 1, Y: 0 U=1006 p=0.03*(>)
hs_success	EN: 3, PT: 4 U=6223.5 p=0.24(<)	N: 4, Y: 4 U=2335.5 p=0.00**(<)	N: 4, Y: 5 U=222 p=0.00**(<)
ws_freq	EN: 4, PT: 4 U=6831.5 p=0.06(>)	N:4, Y: 4 U=2971.5 p=0.04*(<)	N: 4, Y: 4 U=438 p=0.03*(<)
ws_success	EN: 4, PT: 4 U=6351.5 p=0.2(>)	N:4, Y: 4 U=2551.5 p=0.03*(<)	N: 4, Y: 4 U=630 p=0.21(<)

Table 7.3: Context effects of ratio variables: mean (sd) and t-test result. *p<.05; **p<.01

Var	Language	Operators	Terminology
age	EN: 27.52(9.26) PT: 27.19(10.14) t(138.01)=0.25 p=0.8	N: 26.13(8.76) Y: 35.03(13.21) t(36.28)=-3.74 p=0.00**	N: 27.3(9.98) Y: 26.67(5.16) t(5.93)=0.28 p=0.78
ws_years	EN: 8.01(3.17) PT: 8.51(2.93) t(118.54)=-1.15 p=0.25	N: 8.54(2.75) Y: 7.27(4.22) t(36.13)=1.67 p=0.10	N: 8.34(3.03) Y: 10(0) t(249)=-8.69 p=0.00**

Table 7.4: Context effects of nominal variables on the number of terms. *p<.05; **p<.01. KW stands for Kruskal-Wallis.

Var	Mean (sd)	Test	p-value
gender	Female: 3.90 (2.21) Male: 3.33(1.23)	t(224.30)=2.58	p=0.00** (>)
hs_webuse	No: 3.18 (0.85) Yes: 3.89(2.19)	t(229.22)=-3.72	p=0.00** (<)
prev_search	No: 3.31 (1.42) Yes: 4.5(2.81)	t(88.66)= -3.53	p=0.00** (<)
qtype	Overview: 2.82 (0.95) Disease Management: 3.94 (2.05) Treatment: 2.64 (0.77) Prevention/Screening: 4.82 (2.45) Prognosis/Outcome: 6.00 (1.15) Diagnosis/Symptoms: 3.98 (1.94)	KW $\chi^2(5)$= 75.20	p=0.00**
specialty	Psychiatry: 2.71 (0.93) Dermatology: 4.83 (2.65) Gynecology: 4.83 (2.07) Urology: 4.41 (1.53)	KW $\chi^2(3)$= 113.11	p=0.00**

7.4.5 Number of terms

To analyze the effects of users' age on the number of terms in the query, we calculated the Spearman correlation coefficient, obtaining a low correlation of ρ=0.16, p<0.01**. Although age does not have a great influence on the number of terms, the gender does. As can be seen in Table 7.4, females use more terms per query.

The Spearman correlation between years of experience in web search and number of terms used in a query (ρ= - 0.29, p<0.01**) points out an inverse relationship with low expression and suggests that, as the number of years of experience in web search increases, the number of terms gets smaller. The means presented in Table 7.5 show that users that search the Web more frequently have a tendency to formulate longer queries. However, the pairwise comparison presented in Table 7.6 shows that the only significant difference lays in the comparison of the 2nd (*Once a month*) and 3rd (*Once a week*) levels of frequency, in which the first has a lower median. The mean number of terms by web search success (`ws_success`) made us suspect the use of more terms per query could lead to higher success rates but we found no statistically significant differences.

There is statistical evidence to state that who has the habit of using the Web to search for health information employ more terms per query. In these users, the ones that conduct health searches more often tend to user more terms than occasional health searchers. In fact, after the pairwise comparison, we found that the highest frequency (5) of health searches in the Web has a statistically

Table 7.5: Context effects of ordinal variables on the number of terms. *p<.05; **p<.01.

Var	Mean (sd)	Kruskal-Wallis	p-value
clarity	1: 2.50 (0.58) 2: 2.75 (1.06) 3: 3.90 (1.72) 4: 2.86 (1.11) 5: 3.96 (2.42)	KW $\chi^2(4)$= 24.65	p=0.00**
easiness	1: 3.00 (1.41) 2: 3.56 (1.79) 3: 4.17 (2.21) 4: 2.41 (0.71) 5: 2.75 (1.07)	KW $\chi^2(4)$= 39.31	p=0.00**
familiarity	1: 3.25 (1.42) 2: 2.83 (0.93) 3: 4.23 (2.53) 4: 3.91 (1.99) 5: 3.47 (1.29)	KW $\chi^2(4)$= 18.93	p=0.00**
hs_freq	1: 3.44 (1.17) 2: 3.67 (1.40) 3: 3.10 (1.93) 5: 6.87 (3.55)	KW $\chi^2(3)$= 35.00	p=0.00**
hs_success	2: 3.12 (0.61) 3: 3.28 (1.37) 4: 4.01 (2.06) 5: 4.47 (3.17)	KW $\chi^2(3)$= 5.54	p=0.14
ws_freq	2: 2.83 (0.70) 3: 3.85 (1.37) 4: 3.84 (2.35)	KW $\chi^2(2)$= 9.06	p=0.01*
ws_success	3: 3.02 (0.89) 4: 3.69 (1.78) 5: 3.50 (1.77)	KW $\chi^2(2)$= 2.47	p=0.29

Table 7.6: Context effects on the number of query terms. Significant differences found in multiple comparisons. P-value divided by the number of tests performed. MW stands for Mann-Whitney.

Var	Difference	Test value	p-value
clarity	3>2	MW U=473.5	p<0.05/10
	3>4	t(161.98)=4.71	p<0.01/10
	5>4	t(167.77)=-4.13	p<0.01/10
easiness	2>4	t(125.67)=5.33	p<0.01/10
	3>4	t(149.06)=7.59	p<0.01/10
	3>5	MW U=2183.5	p<0.01/10
familiarity	2<3	t(118.24)=-4.89	p<0.01/10
	2<4	t(91.51)=-4.07	p<0.01/10
hs_freq	5>1	MW U=386	p<0.01/6
	5>2	MW U=515.5	p<0.01/6
	5>3	MW U=163.5	p<0.01/6
qtype	Overview<Prevention/Screening	t(98.31)=-6.6	p<0.01/15
	Overview<Prognosis/Outcome	MW U=4	p<0.01/15
	Overview<Diagnosis/Symptoms	t(159.31)=-5.25	p<0.01/15
	Treatment<Prevention/Screening	t(94.31)=-7.27	p<0.01/15
	Treatment<Prognosis/Outcome	MW U=0	p<0.01/15
	Treatment<Diagnosis/Symptoms	t(148.49)=-6.18	p<0.01/15
specialty	Psychiatry<Dermatology	t(64.35)=-6.08	p<0.01/6
	Psychiatry<Gynecology	t(67.86)=-7.69	p<0.01/6
	Psychiatry<Urology	t(51.75)=-7.04	p<0.01/6
ws_freq	2<3	MW U=488	p<0.01/3

higher median of terms than all the other frequencies. Just like what happens in web search success, the descriptive analysis of health search success make us hypothesize that longer queries have higher health success rates. However, these differences are not significant.

As can be seen in Table 7.4, users with previous searches on the topic use more terms per query. The same happens when users are more familiar with the topic. In fact, we found that the 2nd level of familiarity uses less terms than the 3rd and 4th levels (Table 7.6).

The distribution of query terms changes with medical specialties and also with query types (Table 7.4). Further analysis (Table 7.6), allowed us to conclude that the number of terms in psychiatry is lower than in every other specialty. In the query type, we found statistical evidence to say that *overview* and *treatment* questions have, in average, less terms than the *prevention/screening*, *prognosis/outcome* and *diagnosis/symptoms* questions.

If the 3rd level was excluded from the clarity variable, we would conclude that clearer tasks were associated with a higher number of terms. With statis-

tical meaning, we observe that level 3 uses more terms than level 2 and 4 and that level 5 uses more terms than level 4. Regarding the easiness of task, results show that more complex tasks are associated with longer queries. In fact, the highest levels of easiness have less terms than the 2nd and 3rd levels.

7.5 RELEVANCE JUDGMENTS ANALYSIS

In the analysis of the effects of context features on relevance judgments we have also considered the features that belong to the query and document dimensions presented in Table 4.3. Moreover, the hstatus, usual-engine and taskstat features from the user, health search and task dimensions were also considered.

The data analyzed in this section consists of 9572 relevance judgments. Most judgments classify documents as non-relevant (58%), 26% as partially relevant and 17% as totally relevant. Distinguishing levels of relevance 1 and 2 in a scale of 0 (non-relevant), 1 (partially relevant) and 2 (totally relevant) was one of the main difficulties felt, and explicitly pointed by users. The presence of the highest peak on the non-relevance side is in accordance with what Saracevic (2007b) reports.

Our analysis followed the strategy explicit in Figure 7.3. On nominal and ordinal variables (Tables 7.7 and 7.8) we compared the median of relevance in each group of the variable. In ratio variables (Table 7.9) we compared the mean of the variable (e.g. age) in the three levels of relevance.

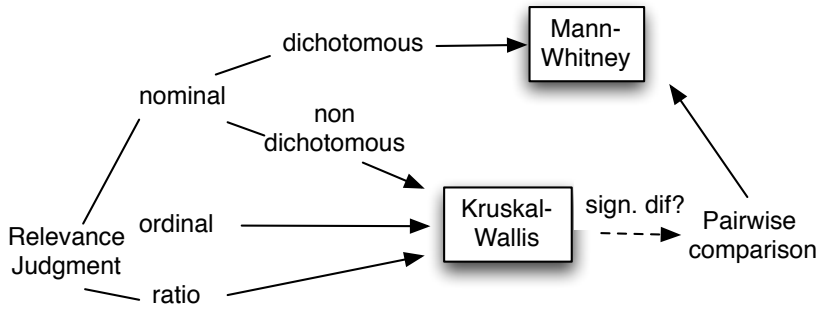

Figure 7.3: Relevance statistical analysis.

7.5.1 *User*

In Table 7.9 it is possible to see the average level of relevance increases with users' age. With further analysis we verified that the average age of users in relevance 0 is lower than in relevance 1 and 2 (Table 7.10). These results make us conclude that younger students tend to classify documents as non-relevant more often. This raises the following question: "Do older users find documents more relevant?". Or is health information more meaningful to older students who are more sensitive to health searches and, therefore, more careful in their analysis?

As seen in Table 7.7, male users judge documents with higher values of relevance.

Table 7.7: Context effects of nominal dichotomous variables on relevance. *p<.05; **p<.01. MW are the initials of Mann-Whitney. All medians are 0, except the one on *usualengine = yes* that is 1.

Var	Test	p-value
gender	MW U= 5504548	p=0.00** (F<M)
hs_webuse	MW U= 4420862	p=0.00** (no<yes)
medterms	MW U= 1068658	p=0.00** (no>yes)
prev_search	MW U= 7420967	p=0.00** (no>yes)
qadv	MW U= 5911562	p=0.00** (no<yes)
qlang	MW U= 8229611	p=0.03* (en<pt)
usualengine	MW U=6951394	p=0.00** (no<yes)

Table 7.8: Context effects of nominal and non-dichotomous variables and ordinal variables on relevance. *p<.05; **p<.01.

Var	Kruskal-Wallis	p-value
clarity	KW $\chi^2(4)$= 39.90	p=0.00**
docrank	KW $\chi^2(2)$= 286.46	p=0.01**
doctype	KW $\chi^2(4)$= 10.18	p=0.03*
easiness	KW $\chi^2(4)$= 25.82	p=0.00**
familiarity	KW $\chi^2(4)$=25.47	p=0.00**
hs_freq	KW $\chi^2(3)$= 48.85	p=0.00**
hs_success	KW $\chi^2(3)$= 105.52	p=0.00**
hstatus	KW $\chi^2(2)$= 14.12	p=0.00**
qtype	KW $\chi^2(5)$= 85.13	p=0.00**
taskstat	KW $\chi^2(4)$= 81.96	p=0.00**
specialty	KW $\chi^2(3)$= 70.31	p=0.00**
ws_freq	KW $\chi^2(3)$= 5.87	p=0.05
ws_success	KW $\chi^2(2)$=61.56	p=0.00**

In user's health status we detected significant differences on the relevance assessed by healthier users (5^{th} level in hstatus) and users with the 3^{rd} and 4^{th} levels: 5<3 and 5<4. This suggests that healthier people judge documents with lower relevance scores. In this question no one answered the 1^{st} and 2^{nd} option. This result agrees with the hypothesis we raised when analyzing the age. Are healthier students less prone to health searches and have less motivation to analyze the documents in depth?

7.5.2 Web search experience

We found that users with less years of web search experience tend to rate documents more often with 0 than with 1.

In the frequency of web searches (`ws_freq`) we found no significant differences but we detected differences in the web search success rate (`ws_success`).

Table 7.9: Context effects of ratio variables on relevance. *p<.05; **p<.01.

Var	Mean (sd)	Kruskal-Wallis	p-value
age	0: 26.8 (9.27) 1: 27.86 (10.79) 2: 27.88 (10.25)	KW $\chi^2(2)$= 30.44	p=0.00**
nterms	0: 3.80 (1.99) 1: 3.65 (1.88) 2: 3.52 (1.89)	KW $\chi^2(2)$= 28.17	p=0.00**
snippet	0:105.3 (278.62) 1: 102 (75.55) 2: 108 (85.83)	KW $\chi^2(2)$= 25.15	p=0.00**
title	0:77.21 (24.93) 1: 77.49 (23.47) 2: 73.93 (23.95)	KW $\chi^2(2)$= 28.83	p=0.00**
ws_years	0: 8.05 (2.90) 1: 8.76 (3.11) 2: 8.61 (3.05)	KW $\chi^2(2)$= 85.71	p=0.00**

Not surprisingly, users that feel they find everything (5 in `ws_success`) find documents more relevant: 5>3 and 5>4. Also, and not expected, we found that 3>4, this is, users that consider to have median success (3 in `ws_success`) rate relevance higher than users with 4 in `ws_success`. No user considered to have the lowest levels (1 and 2) of web search success.

7.5.3 Health search experience

As can be seen in Table 7.7, users who usually conduct health searches on the Web (`hs_webuse`) tend to rate relevance higher than the ones that don't use the Web for this purpose.

In Table 7.8 we can see there are significant relevance differences in the levels of health searches' frequency and health search success rate. In health search frequency, by Table 7.10, we conclude that the lowest frequency in health searches is associated with higher levels of relevance and the opposite happens with the highest levels of frequency in health web searches. This suggests that, as the frequency of health searches rises, the relevance criterion becomes stricter.

Regarding the health search success rate, nobody answered the option 1. Surprisingly, we found that the highest level of success (5 in `hs_success`) is associated with lowest levels of relevance and that the median level of success (3 in `hs_success`) is associated with the highest levels of relevance: 5<2, 5<3, 5< 4, 3>2 and 3>4.

We also concluded that relevance is significantly higher in search engines that users typically use in their own health searches. This suggests habit leads to trust the search engine.

Table 7.10: Relevance judgment analysis. Significant differences found in multiple comparisons - part I. P-value divided by the number of tests performed. Values in *Difference* regard relevance levels in ratio variables and variable's groups in the remaining cases.

Var	Difference	Mann-Whitney	p-value
age	0<1	U= 3471313	p<0.01/3
	0<2	U= 2558583	p<0.01/3
clarity	3>4	U= 2842290	p<0.01/10
	3>5	U= 5120817	p<0.01/10
docrank	0>1	U= 7769394	p<0.01/3
	0>2	U= 5555828	p<0.01/3
	1>2	U= 2174968	p<0.01/3
doctype	pdf>html	U= 1925281	p<0.05/10
easiness	1<3	U= 396414	p<0.05/10
	1<4	U= 96912.5	p<0.01/10
	1<5	U= 67709.5	p<0.01/10
	2<3	U= 5542021	p<0.05/10
	2<4	U= 1352704	p<0.01/10
familiarity	4<1	U= 723743	p<0.05/10
	4<2	U= 2475447	p<0.01/10
	4<3	U= 2827383	p<0.01/10
hs_freq	1>2	U= 3439556	p<0.01/6
	1>3	U= 1485523	p<0.01/6
	1>5	U= 941181	p<0.01/6
	5<2	U= 1049677	p<0.05/6
	5<3	U= 459028	p<0.05/6
hs_success	5<2	U= 255937.5	p<0.01/6
	5<3	U= 1589742	p<0.01/6
	5<4	U= 1774288	p<0.01/6
	3>2	U= 606714	p<0.05/6
	3>4	U= 4798980	p<0.05/6
hstatus	5<3	U= 1760352	p<0.05/3
	5<4	U= 3223237	p<0.01/3
nterms	0>1	U= 6975251	p<0.05/3
	0>2	U= 4805382	p<0.01/3
	2<1	U= 2065639	p<0.05/3

Table 7.11: Relevance judgment analysis. Significant differences found in multiple comparisons - part II. P-value divided by the number of tests performed. Values in *Difference* regard relevance levels in ratio variables and variable's groups in the remaining cases.

Var	Difference	Mann-Whitney	p-value
qtype	Prevention/Screening<Overview	U=2667458	p<0.01/15
	Prevention/Screening<Disease Management	U=536497	p<0.01/15
	Prevention/Screening<Treatment	U=1963428	p<0.01/15
	Prevention/Screening<Diagnosis/Symptoms	U=3125707	p<0.01/15
	Prognosis/Outcome<Overview	U=100631	p<0.01/15
	Prognosis/Outcome<Disease Management	U=20288.5	p<0.01/15
	Prognosis/Outcome<Treatment	U=74228.5	p<0.01/15
	Prognosis/Outcome<Diagnosis/Symptoms	U=85668	p<0.01/15
snippet	0>1	U=7138175	p<0.01/3
specialty	Psychiatry>Dermatology	U= 4671748	p<0.01/6
	Psychiatry>Gynecology	U= 4328250	p<0.05/6
	Psychiatry>Urology	U= 3250727	p<0.05/6
	Dermatology<Gynecology	U=1418530	p<0.01/6
	Dermatology<Urology	U=1067578	p<0.01/6
taskstat	1>2	U= 50724.5	p<0.01/10
	1>3	U= 226135	p<0.05/10
	1>5	U= 86700	p<0.01/10
	3>5	U= 2182096	p<0.01/10
	4>5	U= 4409888	p<0.01/10
	2<3	U= 1052153	p<0.01/10
	2<4	U=1004460	p<0.01/10
title	2<0	U=4809854	p<0.01/3
	2<1	U=2148352	p<0.01/3
ws_success	5>3	U= 100818.5	p<0.01/3
	5>4	U= 468468.5	p<0.01/3
	3>4	U= 3584470	p<0.01/3
ws_years	0<1	U= 2500464	p<0.01/3

7.5.4 Familiarity with the topic

The data in Table 7.7 let us see that users who have done previous searches on the topic (`prev_search`) tend to rate relevance lower than users who didn't. This might be explained by more demanding needs. In Table 7.8 we see there are significant differences between familiarity levels. Further analysis allowed us to conclude that the highest level of familiarity is associated with the lowest relevance. This corroborates our conjectures based on `prev_search`.

7.5.5 Task

Analyzing Tables 7.7 and 7.8 we see there are significant differences between the groups of all the variables in the task dimension: specialty, question type, clarity and easiness. The specific differences will be described next.

In terms of clarity we found that somehow clear tasks (3 in `clarity`) have higher relevance rates than more clear tasks (4 and 5 in `clarity`). In the clarity aspect, a clear pattern does not emerge.

More difficult tasks have lower relevance scores. As expected, we found that tasks with the lowest rate of easiness (1 and 2 in `easiness`) have lower relevance scores then tasks with `easiness` 3, 4 and 5.

Regarding the question type we found that the Prevention/Screening and the Prognosis/Outcome categories, compared with all other types of questions, have the lowest relevance scores.

We can also verify that the psychiatry specialty is associated with the highest levels of relevance. On the contrary, dermatology is associated with lowest levels of relevance.

7.5.6 Query

In the query dimension we noticed (Table 7.7) that relevance is significantly higher when queries use advanced operators, when they use Portuguese terms and when they use lay terms instead of medico-scientific ones. This latter result contradicts a finding presented in Section 7.4.4, where we concluded that, according to `hs_freq`, the use of medico-scientific terminology was associated with more successful health searches. This happens because the relevance evaluated by `hs_freq` is motivational and differs from the situational relevance that is being studied here. In fact, we have already noticed in Section 7.5.3 that motivational relevance is not consistent with the situational one. It is also important to note that the use of lay terms may result in a set of retrieved documents with a language more adjusted to the health consumer and, therefore, in a result set with greater situational relevance.

The means presented in Table 7.9 show the number of terms in a query decreases as relevance increases. We found significant differences in the number of terms' distributions by relevance levels. More precisely, we confirmed our suspicion, that is, relevance 0 has the largest median of terms and level 2 has the lowest median of terms: 0>1, 0>2 and 2<1.

7.5.7 Document

As expected, relevance decreases with the position of the document in the ranking. This tendency can be seen in the means presented in Table 7.9 and in the pairwise comparison.

We found differences in the relevance associated with different types of documents (Table 7.7) where pdf documents have higher relevance than html documents.

Analyzing if the title and snippet sizes had any influence on relevance judgments, we found the distributions of these variables differ by relevance level (Table 7.9). Further analysis let us conclude that non-relevant documents have longer snippets than partially relevant documents and that documents classified as totally relevant have shorter titles than non-relevant or partially relevant documents. Although title and snippet lengths may influence the decision of accessing a document, they don't seem to have impact on the assessment of relevance.

7.5.8 Situational versus motivational relevance

Besides exploring how some of the variables in Table 4.3 affect relevance judgments, we also wanted to study the relationship between the situational rele-

vance given by the relevance judgments and the motivational relevance given by the task completion status (`taskstat`) as perceived by the user.

In Table 7.8 we can see the situational relevance differs by levels of motivational relevance. With further analysis we conclude that users with a lower feeling of success rated higher relevance scores. The exception occurs in the comparison of the 2^{nd} level of success with the 3^{rd} and the 4^{th} level. Although not statistically significant, this is confirmed by a negative Spearman correlation between both types of relevance (ρ=- 0.02, p=0.14). This result was a surprise and is discussed in the following section.

7.6 DISCUSSION OF RESULTS

Based on the results presented in the previous sections, we will now discuss the main results and raise hypothesis.

Users express their queries in English less than we expected and they do it mainly because some search engines have their collections in English. They do it more frequently when the topic is more familiar or the task is clearer. Females tend to use more English terms or to select more often search engines with English content. Even though we found that Portuguese queries had better situational relevance, we think this conclusion was affected by the low English proficiency of the users.

Results suggest that, as the experience in web search increases, users apply more structured queries and this is associated with a higher rate of success, motivational and situational. Also, users with higher health search success rate and users with previous searches on the topic tend to use more advanced operators.

We confirmed our hypothesis and noticed that health consumers seldom use medico-scientific terminology. The small number of queries with medical terminology does not allow a reliable statistical analysis. However we found tendencies that should be explored in further studies. Does the use of medical terminology result in more successful searches? Or does it result in documents whose language is inaccessible to health consumers? Are these terms used more often in familiar tasks? Although we found contradictory results in Section 7.5.6, the rare use of this type of terminology by health consumers opens doors in its exploration on query expansion techniques, assuming the language of the documents retrieved is still accessible to the user.

We found that women, users that did previous searches on the topic and users who frequently use the Web to search for health and other types of information, use more terms per query. Are women more expansive in web search than men? The fact that users with greater familiarity express their information needs with longer queries agrees with some studies mentioned by Jansen and Pooch (2001). However, longer queries did not result in a higher situational relevance.

Queries associated with psychiatry information needs have fewer terms than other specialties. Is it because it is harder to express psychological symptoms than physical ones? The overview and treatment types are also associated with shorter queries. We suppose it might be motivated by a desire of a larger recall in this exploratory kind of questions.

A clearer task is associated with longer queries. We think it is because users have a more clear idea of what they want and therefore think of more terms to describe the information need. Results also suggest a tendency to use longer and structured queries in more complex tasks.

Younger and healthier users often classify documents as non-relevant. Is this type of users stricter in their criteria? Or does this happen because health searches are not so meaningful to this type of users and so they had less motivation to carefully evaluate the documents? We also found that male users judge relevance with higher scores.

Users with less years of web search experience tend to rate documents more often with 0 than with 1. Does this mean this type of users have less confidence in Web documents?

Users that usually conduct health searches on the Web tend to rate relevance higher. Yet, a frequent health searcher is more demanding than an occasional one, being associated with lower relevance scores. Interestingly, users find documents more relevant if they are using a familiar search engine. This suggests habit leads to trust.

We found that users with previous searches on the topic tend to rate relevance lower. This result is in agreement with Saracevic (2007b): "less subject expertise seems to lead to more lenient and relatively higher relevance ratings".

As expected, more difficult tasks have, in general, lower relevance scores. We also found that psychiatry has higher relevance when compared to other specialties. Has the Web more and better information on this topic? Is it because it is a topic easier to discuss online than in face-to-face conversations? About the question type, we found that the Prevention/Screening and the Prognosis/Outcome categories have the lowest relevance scores. The last result is not a surprise since it is hard to do a prognosis without a complete health profile.

Relevance is significantly higher when queries have advanced operators and use lay terms instead of medico-scientific ones. This last result contradicts another finding saying that the use of medico-scientific terminology is associated with a higher feeling of successful health searches (`hs_success`). With this result we see that situational and motivational relevance are not always in harmony. This is emphasized by another finding saying that users with a greater feeling of success are associated with lower relevance scores. This jeopardizes evaluations done with only one type of relevance and asks for evaluation models that incorporate several types of success measures.

As expected, relevance decreases with the position of the document in the ranking. This finding agrees with the concept of ranking that is supposed to be ordered by relevance and with what Saracevic (2007b) reports: "information objects presented early have a higher probability of being inferred as relevant". We also found that pdf documents have higher relevance than html documents.

7.7 CONCLUSION

We have conducted a user study to analyze the influence of user and task context features on query formulation. Moreover, we analyzed the influence of the above features and also of query and document features on relevance judg-

ments. We have reached findings that can foster new ideas to improve information retrieval and also ask for alternative measures of success.

Through the questionnaires we have asked users to evaluate their success rate in web search, in health search and in the completion of the tasks in which they were involved. Some of our findings based on these variables were contradictory to the findings we reached on top of relevance judgments. This suggests that traditional ways to evaluate IR systems can be improved through the incorporation of additional measures that can bring new insights.

Our findings show that the use of advanced operators is directly connected with web search experience and they lead to web and health search success. Similarly, the use of medico-scientific terminology is associated with a higher familiarity with the topic and also leads to higher rates of successful health searches. Along with the rare use of medico-scientific terminology by health consumers, these findings can be used to detect expertise and adjust the IR process, applying specific query expansion techniques or adjusting the result sets.

Results have also raised hypothesis that should be tested in new studies, ideally focusing on a smaller set of variables to avoid interdependencies. A first hypothesis is that questions of the Prognosis/Outcome type need more user context to be successful. The other is that the Web is rich in psychiatric information and Web's anonymity attracts health searches on this topic.

Although English queries led to lower relevance scores, we believe the translation of terms to their English synonym might be a good strategy to improve the result set to users that understand English. The larger number of English queries in tasks with more familiar topics or clearer definitions suggests this can also be a good strategy to users with good health literacy or users familiar with the topic. We think the results of this study were affected by the low English literacy of the users. Following this hypothesis, in the first study of Part III of this dissertation, described in Chapter 9, we analyze how changes in queries' language affect the outcome of a retrieval task in users with different English proficiencies. This is one of the studies that is based on the second experiment of this dissertation, described in Chapter 8. This experiment was also the basis of the work described in the previous chapter.

PART III

QUERY FORMULATION: CONTEXTUALIZATION BY ENGLISH PROFICIENCY, HEALTH LITERACY AND TOPIC FAMILIARITY

USER EXPERIMENT 2

8.1 INTRODUCTION

We conducted an interactive light IR experiment, run on the laboratory, with 40 participants having different levels of English proficiency, health literacy and topic familiarity. Each participant performed 8 tasks, associated with different information situations, query languages and query terminologies. In each task, users had to assess the relevance, comprehension and other characteristics of 30 documents. More details regarding this experiment are given in the following sections.

This experiment served as basis for two main studies and three smaller ones. The first main study is described on Chapter 9 and intends to analyze the impact of query language in health searches performed by users with different English proficiencies. The second main study is focused on exploring how the terminology of a query can affect the retrieval experience of users with different levels of health literacy and topic familiarity. This study is described in Chapter 10. With the data collected in this experiment we also study the impact of limiting the collection of a search engine to certified health documents, as described in Chapter 6. In addition, we explore how context features interact with each other, considering the terminology of the query, as described in Chapter 11. Finally, we study the query (re)formulation behavior of users with different health literacy and topic familiarity, described in Chapter 12.

8.2 INFORMATION SITUATIONS

Following the Borlund (2003b) Interactive IR evaluation model, we defined 8 health information situations that act as the platform against which relevance is judged. For each of them we defined 4 search queries, 2 in English and 2 in Portuguese, the users' native language. In each language, one of the queries was formulated using lay terminology and the other using medico-scientific terminology.

The information situations were defined based on questions submitted to the health category of the Yahoo! Answers service. From the list of open questions of this category and in a decreasing order of popularity, we selected the questions that satisfied the following requirements. Since most of the health Web searches are about diseases (Fox, 2006, 2011), we decided to focus on questions about treatments to a symptom/disease. Because we intend to study the effects of lay and medico-scientific queries, that is, queries containing only lay terms and queries containing at least one medico-scientific term, in users

with different characteristics, we had to ensure that, for each information situation, queries were different. For that reason, when choosing the questions, we made sure that the symptom/disease was associated with different syntaxes in both terminologies, as defined in the Multilingual Glossary of technical and popular medical terms in nine European Languages (Stichele, 1995), described in Chapter 2. For example, diabetes would be excluded because it is simultaneously a lay and a medico-scientific term. Moreover, we also guaranteed that each query had at least 30 results in both search systems used in this study. Every information situation is associated with only one clinical question type (treatment) to minimize the influence of this variable on the experiment. The defined health information situations are:

1. About 3 days ago, I started having a burning feeling every time I urinated. How should I treat this?

2. For the past 5 days my head has been very itchy and I don't have lice. What can I do to stop the itching?

3. I have high uric acid (8.0 mg/dL) with reference units 3.6 - 7.7. How can I lower my uric acid level?

4. I am suffering with an inflammation on my lips and mouth area for more than a year. I have difficulties eating. What can I do to treat it?

5. My father got bit by a dog and is in the hospital with a bone infection. How is this treated?

6. I frequently get heartburn even when I stay away from spicy stuff. What can I do to prevent it?

7. I have been noticing lots of hair coming out from my head. Usually I only comb my hair once a day. What can I do to stop losing my hair?

8. I'm on the computer all day so I type a lot and use the mouse. My right pointing finger is starting to give me some joint pain. How I can treat my finger?

Information situations were initially formulated in English and subsequently translated to Portuguese by the researchers. They were communicated to users in Portuguese. The researchers defined the queries for each information situation whereas the users only assessed the retrieved documents. Queries were built concatenating the symptom or disease with the word 'treatment'. Terms' translation between languages and terminologies was supported by the multilingual glossary mentioned above. Although smaller and less current than other existing consumer health vocabularies like the CHV, this glossary was the only one to simultaneously have multilingual (Portuguese and English included) and two types of terminology, lay and medico-scientific. This glossary did not restrict the selection of information needs. The queries defined for each information situation are presented in Table 8.1.

Table 8.1: Queries associated with the information situations (Sit).

Sit	PT/Lay	PT/MS	EN/Lay	EN/MS
1	dificuldade urinar tratamento	disúria tratamento	painful urination treatment	dysuria treatment
2	comichão cabeça tratamento	prurido cabeça tratamento	head itching treatment	head pruritus treatment
3	ácido úrico elevado tratamento	hiperuricemia tratamento	high uric acid treatment	hyperuricaemia treatment
4	inflamação boca tratamento	estomatite tratamento	mouth inflammation treatment	stomatitis treatment
5	infecção osso tratamento	osteomielite tratamento	bone infection treatment	osteomyelitis treatment
6	azia tratamento	pirose tratamento	heartburn treatment	pyrosis treatment
7	queda cabelo tratamento	alopecia tratamento	hair loss treatment	alopecia treatment
8	dor articulação tratamento	artralgia tratamento	joint pain treatment	arthralgia treatment

8.3 RETRIEVAL SYSTEMS

We used Google as a *black-box* search engine with two different collections. The first is associated with Google entire index and the second is the set of HONcode certified pages indexed by Google. To simplify, we consider that these two different collections lead to what we call two retrieval systems. As described in Chapter 2, the HONcode certification is proposed by the Health On the Net Foundation (HON) to help assess the accuracy of health content and the credibility of the publishers. We have used the Google custom search built by the HON to restrict the Google collection to HONcode certified sites. Currently, this collection contains more than 1 million pages and 52% of the sites are in English (Baujard et al., 2011). For each query, we automatically collected the top-30 results from each retrieval system. To reduce the risk of Google learning from the previous submitted queries, we ensured that returned links were never clicked. Additionally, to prevent changes in the search engine or in the HON collection, we submitted all queries within a very short time span.

8.4 ASSESSMENT TASKS

A query run on a retrieval system leads to an assessment task that a user can execute. Since we defined 8 information situations with 4 queries each, and we use 2 retrieval systems, a total of 64 assessment tasks exist in our user study.

Each user was assigned a set of 8 different assessment tasks. In the assignment of tasks to users we applied a Latin-square like procedure so that all users assessed the relevance: (1) of all information situations, but only once each; (2) of queries of both languages the same number of times; (3) of queries of both types of terminology the same number of times and (4) in all the retrieval systems the same number of times. We have also permuted the order of assessment tasks to avoid possible bias of relevance assessments owing to human behavior. We also guaranteed that each iteration of relevance assessments contained, in the same number of times, (5) queries of both languages, (6) queries of both terminologies and (7) tasks in both retrieval systems. Ad-

ditionally, to preempt users' fatigue, each task had to be performed in different days, that is, tasks had to be separated by an interval of, at least, 24 hours. Users did not have time limits to perform each task. In Table 8.2 we present the tasks assigned to a subset of 16 users where *e* stands for English, *p* for Portuguese, *l* for lay and *m* for medico-scientific.

Table 8.2: Latin square procedure followed in task assignment. Font style (regular and bold) defines the retrieval system, [1-8] defines the information situation, [e, p] the queries' language and [l, m] the queries' terminology.

	#Iteration							
User	1	2	3	4	5	6	7	8
1	**1el**	**2el**	**3pm**	**4pm**	5pl	6pl	8em	7em
2	**1em**	**2em**	**4pl**	**3pl**	5pm	6pm	7el	8el
3	**2pl**	**1pl**	**3em**	**4em**	6el	5el	8pm	7pm
4	**2pm**	**1pm**	**4el**	**3el**	6em	5em	7pl	8pl
5	3el	4el	2pm	1pm	**7pl**	**8pl**	**5em**	**6em**
6	3em	4em	1pl	2pl	**7pm**	**8pm**	**6el**	**5el**
7	4pl	3pl	2em	1em	**8el**	**7el**	**5pm**	**6pm**
8	4pm	3pm	1el	2el	**8em**	**7em**	**6pl**	**5pl**
9	5pl	6pl	8em	7em	**1el**	**2el**	**3pm**	**4pm**
10	5pm	6pm	7el	8el	**1em**	**2em**	**4pl**	**3pl**
11	6el	5el	8pm	7pm	**2pl**	**1pl**	**3em**	**4em**
12	6em	5em	7pl	8pl	**2pm**	**1pm**	**4el**	**3el**
13	**7pl**	**8pl**	**5em**	**6em**	3el	4el	2pm	1pm
14	**7pm**	**8pm**	**6el**	**5el**	3em	4em	1pl	2pl
15	**8el**	**7el**	**5pm**	**6pm**	4pl	3pl	2em	1em
16	**8em**	**7em**	**6pl**	**5pl**	4pm	3pm	1el	2el

8.5 SEARCH PROCEDURE

As represented in Figure 8.1, users began by answering a quiz to evaluate their English proficiency, described in Section 8.7, and a quiz to evaluate their health literacy described in Section 8.8. They then answered a questionnaire about demographic information, web search experience, health web search behavior and their familiarity with each medico-scientific term associated with every information situation (e.g.: osteomyelitis for information situation 5). This questionnaire is presented in Appendix D. Although users do not assess documents retrieved with their own queries, we also asked them to provide the query they would use for the information need triggered by the information situation, using a text-field similar to the ones used in search boxes. After this questionnaire, users enrolled in a sequence of 8 assessment tasks in which they judged the top-30 URL collected by the researchers for each task. Afterwards, users had to answer a post-search questionnaire, presented in Appendix E, about the performed task. Additionally, they defined two more queries and answered the information situation.

Using a web collaborative spreadsheet, as visible in Figure 8.2 for the first information situation, for each URL, users had to indicate the type of the document (webpage, pdf, ppt, doc or other), the language of the document (Por-

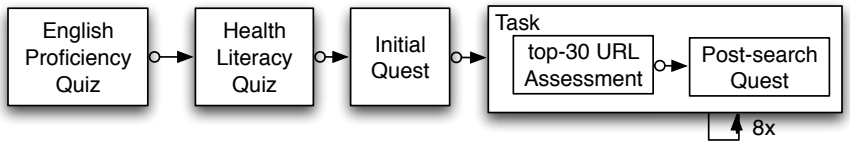

Figure 8.1: Procedure followed by the users.

tuguese, English or other), the relevance of the document to the information situation considering the user's own context, and the extent to which the document was understood. Each cell is associated with specific instructions and its content is validated against the allowed values. To maintain the coherence with the language of this document, the interface presented in Figure 8.2 was translated to English but users interacted with a Portuguese interface. Users were allowed to follow links to the internal pages of the provided URL, if they felt it made sense. Situations in which this was expected include pages with scientific papers' abstracts in which the access to the full-paper was only one-click away or pages in which content was deliberately separated in several pages with access through a 'previous-next' menu. Users were instructed to report the URL in which they followed hyperlinks. Users were also instructed to report the situations in which there was an error loading the URL and the situations in which the page loaded but it had no content (e.g.: restricted access).

About 3 days ago, I started having a burning feeling every time I urinated. How should I treat this?

Rank	URL	Doc type	Language	Relevance	Comprehension
1	http://www.drugs.com/condition/dysuria.html				
2	http://www.pregnancy-calendars.org/diseases/dysuria.html				
3	http://www.emedicinehealth.com/dysuria/article_em.htm				
4	http://www.emedicinehealth.com/dysuria/page5_em.htm				
5	http://www.ayushveda.com/health/dysuria.htm				
6	http://www.nativeremedies.com/ailment/dysuria-difficult-painful-urination-treatments.html				
7	http://www.womens-health-club.com/diseases/dysuria.htm				
8	http://www.merck.com/mmpe/sec17/ch226/ch226f.html				
9	http://www.articlegold.com/Article/Dysuria---Causes--Symptoms-and-Treatment/2130				

validation Please assess the document's relevance in a scale of 0 to 2. Use 0 for non-relevant documents, 1 for partially relevant documents and 2 to totally relevant documents.

Figure 8.2: Screenshot of the assessment interface for information situation 1.

Relevance and comprehension were assessed in a 3-value scale to convey more realism to the experiment (Borlund, 2003b). Since we wanted the relevance judgments to represent the value of the information objects for each particular user, we instructed them to judge relevance in accordance to the definition of situational relevance. According to Saracevic (1996), situational relevance is "the relation between the task at hand and the retrieved documents, being inferred by criteria like usefulness in decision making, appropriateness of information in resolution of a problem and reduction of uncertainty". To assess relevance, the three options pertaining the usefulness of the URL to the resolution of the problem, were 'not relevant', 'partially relevant' and 'totally relevant', denoted by 0, 1 and 2, respectively. For comprehension, the three options were 'I did not understand the document's content', 'I partially understood the document's content' and 'I understood the document', denoted by 0, 1 and 2, respectively.

In the post-search questionnaire users were asked to (1) evaluate the search task's completion status used to evaluate motivational relevance, (2) to indicate two additional queries for the information need triggered by the information

situation using text-fields similar to the ones used in search boxes, and (3) to indicate treatments for the condition mentioned in the information situation. In the post-search questionnaire users are asked (1) if they already searched for that topic, (2) to evaluate the task in terms of familiarity, (3) to evaluate their feeling of success with the task, and (4) to indicate treatments for the condition mentioned in the task.

8.6 READABILITY ASSESSMENT

Documents' readability was computed using the Simple Measure of Gobbledygook (SMOG) metric, following the Equation 8.1. Since this metric has been recommended as a measure of readability in consumer-oriented healthcare documents (Fitzsimmons et al., 2010), we decided to include it in our analysis.

$$SMOG = 1.043\sqrt{30\frac{\#polysyllables}{\#sentences}} + 3.1291 \tag{8.1}$$

As shown in Figure 8.3, the computation of the readability metric was performed in three stages. We started by extracting the main content of the documents, excluding components like menus, advertising, footers and headers. Then, we removed the HTML tags from the document generated in the first phase to obtain a text document with the main content of the original one. With this document we computed SMOG using a readability metrics API[1].

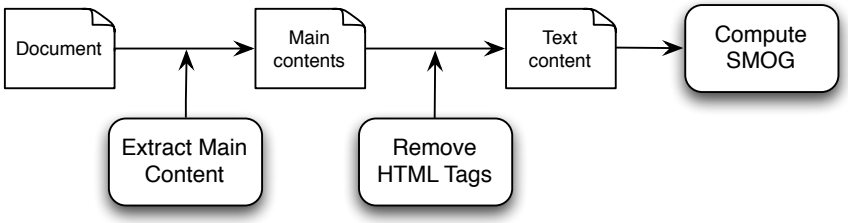

Figure 8.3: Computation of SMOG process.

We could not compute the SMOG metric in 8.7% of the URL for one of the following reasons: error loading the original URL (0.8%), restricted access to content in the original document (0.4%), main content with no text (6.5%) and errors during the extraction of the main content (1%).

Although the use of SMOG in the Portuguese language lacks statistical validation, we still decided to use it due to its applicability to health content and the lack of a validated tool to measure the readability of Portuguese documents.

8.7 ENGLISH PROFICIENCY ASSESSMENT

English proficiency was evaluated with a quiz with questions from a quiz available on the Web[2]. Users answered, in less than 20 minutes, to 26 random

[1] Available at: http://ipeirotis.appspot.com/readability-api.html

[2] Available at: http://www.transparent.com/learn-english/proficiency-test.html (Archived by WebCite at http://www.webcitation.org/5ym7JFqw1)

multiple-choice questions, 8 random questions from the *English Grammar I* category, 8 random questions from the *English Grammar II*, 5 random questions from the *English Vocabulary* and 5 random questions from the *English Reading comprehension*. The full set of questions, divided by category, is presented in Appendix F.

8.8 HEALTH LITERACY ASSESSMENT

Health Literacy can be assessed through several existing instruments like the Test of Functional Health Literacy in Adults (TOFHLA) (Parker et al., 1995) that takes up to 22 minutes to administer and the Short Test of Functional Health Literacy in Adults (STOFHLA) (Baker, 1999), a TOFHLA smaller version. The Rapid Estimate of Adult Literacy in Medicine (REALM) (Davis et al., 1993) is another option, easier and quicker to administer. In non-English languages there are other tools like the Short Assessment of Health Literacy for Spanish-speaking Adults (SAHLSA) (Lee et al., 2006) that was developed based on REALM and also incorporates a comprehension test using multiple-choice questions. In SAHLSA, 50 items have to be associated with one of two options.

Since there is not an instrument in Portuguese to assess health literacy, we decided to adapt SAHLSA because, when compared to English, Spanish is more similar to Portuguese. We translated the 50 medical concepts used in SAHLSA to Portuguese. In Appendix G we present the translated concepts along with the original Spanish term. We can see that the majority of the translations result in very similar terms. Users were asked to associate each concept to one of two terms, in less than 4 minutes. Users were instructed not to guess the answer. With SAHLSA, if users score less than 37, they are considered to have inadequate health literacy. We grouped users in three classes (Inadequate, Elementary, Good) based on the SAHLSA threshold and clusters obtained through hierarchical clustering.

8.9 TOPIC FAMILIARITY ASSESSMENT

To evaluate topic familiarity, users were asked if they previously searched for that topic. They also had to evaluate task familiarity in a 5-value scale and say if they knew the meaning of the medico-scientific concept behind the disease/condition associated with the information need (e.g.: osteomyelitis for information situation 5). To compute a single measure to assess topic familiarity, we combined the previous metrics, as stated in Equation 8.2, to obtain what we call of Combined Topic Familiarity (CTF).

$$CTF = TaskFam + 3 \times PreviousSearch + 2 \times KnewMSTerm \tag{8.2}$$

This formula considers that *TaskFam* is assessed in a 1 to 5 scale, *PreviousSearch* as 0 or 1 and *KnewMSTerm* as 0 or 1. The user's task familiarity assessment is considered the most important feature, followed by the existence of previous searches about the topic and the knowledge of the medical term. CTF is an integer that varies between 0 and 10. Since this is a discrete variable and 10 categories are not justifiable, we grouped CTF in three categories of

familiarity: unfamiliar (CTF<=3), somehow familiar (3<CTF<7) and familiar (CTF>=7).

To analyze the relationship between health literacy and topic familiarity, we applied the chi-squared test of independence and found we could not reject the null hypothesis that both variables are independent ($\chi^2(4)$=5.66, p=0.23). This helps to sustain the claim that both variables are different.

8.10 MEDICAL ACCURACY ASSESSMENT

In the post-search questionnaire, users had to write an answer to the information situation that triggered the assessment task. A medical doctor evaluated these answers in relation to their correct and incorrect content. Answer's correctness was evaluated in a scale of 0 (inappropriate answer) to 2 (appropriate answer) where 1 was used for answers with *some value*. In terms of incorrectness, answers were classified with 0 (all or almost all content is incorrect), 1 (some incorrect content) or 2 (no incorrect content). To exemplify the independence of these characteristics, suppose a user has given the following answer to the third information situation: "Reduce the ingestion of red meat. Increase weight.". This answer has some correct content but it's not complete and would be classified with 1 in terms of correctness. In addition, since it also has some incorrect content, it would be assessed with 1 in terms of incorrectness. If the second sentence was not included in the answer, its incorrectness assessment would be 2 instead of 1. On the other hand, if it had several other wrong suggestions, it could be classified with 0 in terms of incorrectness.

To evaluate the reliability of the medical assessments, a second medical doctor judged 30% of the answers and we estimated the inter-rater reliability using the weighted Cohen's Kappa, an adaptation of Cohen's Kappa to ordinal scales that treats disagreements differently. The measured weighted Cohen's Kappa, with squared weights, for the correctness ratings is 0.68 (95% CI: [0.54, 0.77]), indicating a substantial agreement. For the incorrectness ratings, this measure is 0.7 (95% CI: [0.48, 0.84]), also pointing a substantial agreement. These inter-rater reliability results assure the quality of the initial ratings.

From the answer's correctness and incorrectness we computed a third variable to which we called *medical accuracy* that corresponds to their sum. The *medical accuracy* varies therefore between 0 (lowest accuracy) and 4 (highest accuracy).

8.11 SUMMARY OF CONTEXT FEATURES

In Table 8.3 we summarize the context features used in this study. We group features into categories and, for each feature, we present its definition, measure scale and data collection methods. All the features have already been discussed in the previous sections.

8.12 USERS

Forty undergraduate students participated in this study (25 females; 15 males) with a mean age of 22.25 years (SD = 6.42). Although 4 students have non-Portuguese nationalities, all of them have Portuguese as their native language.

Table 8.3: Summary of context features used in this study.

Category	Context feature	Definition	Scale	Collection method
User	English Proficiency	Users' English skills.	Ordinal: low, elementary and good English proficiency.	Assessed through an English proficiency test graded from 0 to 100. Grouped later through hierarchical clustering.
	Health literacy	The "capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions" (Kutner et al., 2006).	SAHLSA score between 0 and 50. Grouped in Inadequate, Elementary and Good Health Literacy.	Grade obtained in the adapted SAHLSA health literacy assessment test.
User & Document	Relevance	It relates the task and the retrieved documents, "being inferred by criteria like usefulness in decision making, appropriateness of information in resolution of a problem and reduction of uncertainty" (Saracevic, 1996).	Ordinal: from 0 (not relevant) to 2 (totally relevant).	Users' judgment pertaining the usefulness of the document to the resolution of the problem. Part of their assessment task.
	Comprehension	Users' understanding of the document.	Ordinal: from 0 (not understood) to 2 (totally understood).	Users' judgment. Part of their assessment task.
Document	Readability	The degree to which the text contained in the document is easily read.	Rational.	Automatically computed through the Simple Measure of Gobbledygook readability measure.
	Type of document	File type of the main content found in the URL being assessed.	Nominal: webpage, pdf, ppt, doc or other.	Identified by users and manually validated by researchers when inconsistencies were found.
	HONcode certification	Is the document certified by the Health on the Net Foundation?	Nominal: yes or no.	Positive if it is in the set of retrieved documents of both systems. Automatic extraction.
	For consumers?	Positive if it belongs to the consumers' category in the HONCode classification of documents.	Nominal: yes or no.	Automatically computed.
	For professionals?	Positive if it belongs to the professionals' category in the HONCode classification of documents.	Nominal: yes or no.	Automatically computed.
User & Task	Answer's medical accuracy	The degree to which the answer that users give after each search task contains the adequate quantity of correct information and no incorrect information.	Varies between 0 (least accurate) and 4 (most accurate).	Computed using the medical evaluation of users' answers in terms of their correct and incorrect contents.
	Combined Topic Familiarity	User's general knowledge about the topic of a search task.	Varies between 0 and 10. Grouped in three classes: unfamiliar, somehow familiar and familiar.	Combined three user's assessments: task familiarity, previous searches on the topic and knowledge on the medico-scientific term behind the topic.
	Motivational relevance	It relates the user's goals and motivations with the information objects. It is expressed by users' feeling of success and satisfaction (Saracevic, 1996).	Ordinal scale of 1 (totally disagree) to 5 (totally agree).	Obtained through users' agreement with the following claim "I believe I succeeded in this search task", after the search task.

The evaluation of the English proficiency quiz was done in a 0 to 100 scale and students' average classification was 73.94 (SD=18.54). Hierarchical clustering was used to identify low English proficiency (EP1) (n=8), elementary English proficiency (EP2) (n=21), and good English proficiency (EP3) (n=11) groups.

In the health literacy test, evaluated in a 0 to 50 scale, users had in average 45.48 (sd=5.97). These results show that, globally, users have good health literacy. Users' distribution by health literacy classes is the following: Inadequate (9 users), Elementary (13 users) and Good Literacy (18 users).

Users' familiarity with a topic depends on the task's subject. A global analysis demonstrated that topic familiarity is mostly low. As said before, CTF varies between 0 and 10 and its mean value is 3.92 with a standard deviation of 2.18. Pairs "user, topic" are distributed by the proposed topic familiarity categories as follows: Unfamiliar (161 pairs), Somehow familiar (113 pairs) and Familiar (46 pairs). Through this distribution we can also see that the majority of tasks presents a topic unfamiliar to the user.

In a scale of 1 (Not healthy) to 5 (Very healthy), 77.5% answered 4 or 5 revealing a sample of healthy users. The mean number of years users have been searching the Web is 8.55 (SD = 2.17), most of the users (60%) do more than one search per day (5 in a 1-5 scale) and 70% say they find what they want almost all the time (4 in a 1-5 scale).

A small proportion of users (20%) never conducted a health search on the Web. The majority of the remaining users say they perform one health search per month (50%). Although the mode and median of the health search frequency decreases as health status improves, which might lead to the conclusion that less healthy people search more about health issues on the Web, the difference of medians of health search frequency between health status levels is not significant (KW $\chi^2(2) = 3.1$, p=0.22).

In health searches, users feel less successful than in general searches, being mostly divided between "I sometimes find what I am looking for" (3 in a 1-5 scale) - 41% and "I frequently find what I am looking for" (4 in a 1-5 scale) - 47%. A large number of users do their health searches always (63%) or almost always (31%) in Portuguese. The use of the English language to express health queries is less consensual: 22% answered they never use it, 34% almost never and 31% expressed they use it sometimes. A large majority of the users never use languages other than Portuguese and English (88%). Only a minority of the users routinely uses medico-scientific terminology in their health searches, 6.25% of the users use it in every health search and 9.4% use it frequently. The majority (59%) of the users say, sometimes, they use technical terminology and a quarter of the users never, or almost never, use this type of terminology.

In an open question, users were asked about their difficulties when performing health searches on the Web. Four the major issues identified in this set of answers are "finding medical terminology to formulate the query" (21% of the answers), "dealing with the quantity of medico-scientific terminology found in the retrieved documents" (21%), the small amount of documents in Portuguese (8%) and the difficulties of translating medical terms to English (4%).

8.13 CONCLUSION

In this chapter we described a second interactive light IR experiment in which 40 undergraduate students have participated. Although users also had to assess documents' comprehension, when compared with the first experiment, users had here a more focused participation. In this second experiment, users did not have to formulate queries and could not choose the search engine. In addition, the information situations were all associated with the *treatment* clinical question type. The collected user features are also different in the two experiments. Here, we also acquired information regarding users' English proficiency and health literacy.

This experiment was the basis of the studies described in Chapters 6, 9, 10, 11 and 12. The next chapter describes one of these studies and intends to analyze if and how English queries can be useful for non-English users having different English proficiencies.

MEASURING THE VALUE OF HEALTH QUERY TRANSLATION: AN ANALYSIS BY USER LANGUAGE PROFICIENCY

9.1 INTRODUCTION

An obstacle that users commonly face in consumer health retrieval is the lack of content in their native language (Cline and Haynes, 2001). In 2000, Grefenstette and Nioche (2000) estimated the ratio of web content in several non-English European languages in relation to English content. Through these values, lower than 7% in every language, we can see that English was at that time, by far, the most used language on the Web, compared with European languages. Presently, the reality has not changed much and Russian, the second most popular language, is only 11.1% of the English content (W3Techs, 2013). Health is no exception and, in this domain, English is even considered the *lingua franca* (Hersh, 2008a).

In the health domain, where information quality is critical, a larger quantity of information may mean an easier access to higher quality content. We think this can be explored by using web search engines to provide native speakers of languages with less presence on the Web a better service in health searches. This could be done through the suggestion of alternative English queries or, with less user involvement, through the inclusion of English content in the set of retrieved documents. We emphasize that, while the translation of commonsense terms may be simple for some users, with health matters, translations are often not obvious and frequently require domain knowledge to translate medical terms to other language. Moreover, since English documents are not accessible to every user having another primary language, the adopted strategy has to be personalized to users' English proficiency.

In this study we evaluate the effect of translating query terms from the users' native language to the English language, in users with different levels of English proficiency. Users' native language is Portuguese, a language with a smaller Web presence and a large number of native speakers. In 2000, Grefenstette and Nioche (2000) estimated that Portuguese content was only 2.4% of the English Web content. This proportion has raised but it is still only 4.2% (W3Techs, 2013). This accentuates the importance of promoting access to English content in Portuguese health searches. On the other hand, Portuguese is one of the most spoken languages in the world with about 178 million native speakers (Lewis, 2009) meaning that our results can be generalized to a substantial number of users.

The research presented here is different from previous work in several ways. First, unlike previous studies, it considers user characteristics, namely his language proficiency, while studying the impact of a query processing technique. Second, the evaluation is focused not only on users' relevance assessments but also on documents' type, language, comprehension and readability, users' motivational relevance and, more importantly, on the quality of the medical knowledge that emerges from the search session. As defined by Saracevic (1996), motivational relevance relates the user's goals and motivations with the information objects and it is expressed by the user's feeling of success and satisfaction. Third, it is one of the first works exploring the impact of query language translations in consumer health information retrieval.

The experiment behind this study was described in Chapter 8. The remainder of this chapter is structured as follows. In Section 9.2 we do a literature review on cross-language health IR. Sections 9.4, 9.5 and 9.6 report our findings and Section 9.7 discusses the results and presents their implications. We conclude in Section 9.8.

9.2 CROSS-LANGUAGE HEALTH IR

In Cross-Language Information Retrieval (CLIR) users can formulate a query in their most fluent language and retrieve documents in different languages. This can be accomplished through query or document translation. In the healthcare domain, the amount of work on CLIR has been small and most of the work focuses on the development of multilingual resources to assist it (Hersh, 2008a).

As we will point out, previous studies are focused on cross-language health IR but ignore the diversity of users and their inherent characteristics and specificities. The majority of the works explore the multilingual characteristics of the UMLS Metathesaurus described in Chapter 2. Only the works from Pirkola (1998) and Rosemblat et al. (2003) do not use this knowledge source. Pirkola (1998) studied the effects of query structure and three query translation methods using a general and a medical dictionary. The evaluation of the methods used TREC's health related topics and showed that structured queries translated with medical and general dictionaries are almost as good as the original English queries. Rosemblat et al. (2003) compared query and document translation for CLIR using machine translation and a subset of queries from the ClinicalTrials.gov website, concluding that query translation outperforms document translation in terms of retrieval performance.

To internationalize SAPHIRE, a system that extracts concepts from documents and queries, Hersh and Donohoe (1998) used the six languages available in the 1998 Metathesaurus after mapping the text in the documents to the concepts in the thesaurus. Its performance was evaluated on German terms and showed that additional work was still necessary to handle plural and suffix variants. Another experiment with German was made by Volk et al. (2002) who annotated documents and queries with linguistic information that included the identification of medical terms and semantic relations between them. This study's results showed that linguistic processing is essential to a good performance in the German language. Moreover, authors concluded that semantic information increases performance in retrieval.

Eichmann et al. (1998) used the UMLS to translate Spanish and French queries into English using several strategies (full match, partial match, dictionary based and simple addition of Spanish/French query words). To analyze the retrieval effectiveness of the translated queries, the authors used OHSUMED and concluded that, for Spanish, the results are similar to the ones reported on the CLIR literature and, for French, the results are worse. A French/English CLIR system was also proposed by Tran et al. (2004) to support the retrieval of English documents with French queries. Authors used a method to translate queries that mixes a hybrid machine translation method and a translation method based on the UMLS. Using Google and PubMed to predict the accuracy of their translations, authors showed that the hybrid method is better than the machine translation and thesaurus-based methods alone. Also in French and more focused on consumer health information retrieval, is the work from Névéol et al. (2006). Here, the UMLS is used to translate the lay terms of a French medical catalogue to MedlinePlus English topics. Authors suggest that this can be explored in the future to translate patient queries.

Lu et al. (2005) have worked on Chinese-English health CLIR. Using Web-based term translation, they proposed a semi-automatic approach to construct a Chinese-English MeSH.

The works described above evidence the lack of studies exploring the impact of translating queries in health information retrieval. This reality is even more extreme in consumer health information retrieval. Moreover, our study is the first to detail the analysis of the effect of query translation according to characteristics of users and documents.

9.3 RESEARCH QUESTIONS

One major research question drove our research and two secondary ones emerged during the study design. The prime question being investigated in this study is:

What is the impact of translating a health query to English for users with different levels of English proficiency?

The impact will be analyzed from six perspectives: documents' characteristics; precision; medical accuracy; documents' comprehension; documents' readability and motivational relevance. As far as the first of these is concerned, the analysis will only be done generally, that is, without considering users' English proficiency because it does not affect documents' characteristics.

The two other research questions are secondary because we can envision other experiment settings better suited to explore them. However, we feel the current study allows a superficial analysis that can help raise research hypothesis for future studies. These questions are:

1. Does access to English content affect users' query reformulation behavior? Is this similar for all English proficiency levels?

2. Is it possible to predict English proficiency through users' search habits?

To accomplish our research goals we conducted the laboratory user study described in Chapter 8. From the context features presented in Table 8.3, in this

study we have used the following ones: English proficiency, relevance, comprehension, readability, type of document, HONcode certification, answer's medical accuracy and motivational relevance.

9.4 QUERY TRANSLATION EFFECTS

The impact of queries' language translation is analyzed by documents' characteristics, including their comprehension and readability, by the medical accuracy of the answers and by users' motivational relevance. In the analysis that follows we consider the users' assessments made in both retrieval systems. We will use a * to mark significant results at $\alpha = 0.05$ and a ** to mark significant results at $\alpha = 0.01$.

9.4.1 Documents' characteristics

We manually checked all the URL in which there were discrepancies in users' assessments regarding the document type, its language and the cases in which users reported errors on HTTP loading or on access to content.

Portuguese queries led to more HTTP errors (1.4% of all URL retrieved through queries in this language) than English queries (0.2%). This difference is statistically significant at $\alpha = 0.01$ ($\chi^2(1) = 5.7$). In terms of URL without access to content, the proportion was similar in both languages (0.4%).

English queries returned 100% of documents in the English language while Portuguese queries returned 93% of Portuguese documents and 7% of Spanish documents. Spanish documents were retrieved due to the similarity of some terms between the two languages. They were retrieved in the queries *disúria tratamento* and *hiperuricemia tratamento*. The Spanish translations of these queries are *disuria tratamiento* and *hiperuricemia tratamiento* and the differences lay in the accentuation of one word and an additional character in the word *tratamento*. Search engines often ignore the first difference and the second may be considered a typographical error.

In terms of document type, as expected, both languages retrieved mostly webpages (96.3% in English queries and 85.8% in Portuguese ones) and the proportion of webpages in English queries is significantly higher than in Portuguese ones ($\chi^2(1) = 54.2$, $p < 0.01^{**}$). English queries retrieved less pdf documents (3.7% against 13.1%; $\chi^2(1) = 47.7$, $p < 0.01^{**}$) and did not retrieve PowerPoint (0.6% in Portuguese queries) or Word documents (0.5% in Portuguese queries). This seems to indicate that Portuguese content, when compared with English one, has less documents built specifically for the dissemination of health information on the Web.

9.4.2 Precision

We use Graded Average Precision (GAP) and Graded Precision (gP) with an equally balanced g_1 and g_2 to evaluate and compare precision. These measures are described in Section 5.2.1.

We start with a global analysis that does not consider users' English proficiency. In Table 9.1 we present, for each precision measure and language, its mean and standard deviation. As can be seen, English queries have a higher

precision and lower dispersion on all measures, differences that are statistically significant at $\alpha = 0.01$.

Table 9.1: Mean and standard deviation of GAP, gP10 and gP5 by language. Statistical differences between languages in each measure.

	EN		PT		$\mu_{en} > \mu_{pt}$?	
	$\overline{x}$	s	$\overline{x}$	s	test value	p value
GAP	0.73	0.16	0.61	0.24	$t(281.3) = 5.3$	p=0.00**
$gP10$	0.77	0.26	0.58	0.33	$t(307) = 6.5$	p=0.00**
$gP5$	0.69	0.26	0.48	0.31	$t(301) = 5.5$	p=0.00**

An analysis by level of English proficiency shows that English queries have, consistently, a higher GAP for all levels of English proficiency. Furthermore, English queries have lower dispersion for all English proficiency levels. From each level of English proficiency, we also tested whether the differences between languages were significant. In Table 9.2 we can see that, excluding one case, with the three measures used in this study, English queries had a significantly higher precision at $\alpha = 0.01$ in all levels of English proficiency. With the one exception, this difference is significant at $\alpha = 0.05$.

Table 9.2: GAP, gP5 and gP10 statistical differences in levels of English proficiency.

	Low EP	Elementary EP	Good EP
$\mu_{GAP_{en}} > \mu_{GAP_{pt}}$	t(53.74)=2.45	t(144.7)=3.82	t(78.72)=2.79
	p=0.01**	p=0.00**	p=0.00**
$\mu_{gP10_{en}} > \mu_{gP10_{pt}}$	t(59.12)=2.86	t(158.78)=4.83	t(84.51)=3.31
	p=0.00**	p=0.00**	p=0.00**
$\mu_{gP5_{en}} > \mu_{gP5_{pt}}$	t(58.22)=3.82	t(155.6)=3.61	t(82.65)=2.34
	p=0.00**	p=0.00**	p=0.01*

Using the Kruskal-Wallis test, we also investigated, in each language, if there were significant differences in the mean GAP/gP5/gP10 between levels of English proficiency. In both languages, we did not find significant differences between the three levels of proficiency.

9.4.3 Documents' comprehension

The comprehension median is 2 (in a scale of 0-2) and this indicates that, from a general perspective, users considered the documents easy to read. The median is the same for English and Portuguese queries.

An analysis by English proficiency (Figure 9.1) shows that, in users with low English proficiency, the proportion of documents in which the users understood the document (2 in the comprehension scale) is higher for documents retrieved with Portuguese queries (41.6%) than with English queries

(24.6%). For elementary English proficiency users, Portuguese queries still have a higher degree of comprehension, although the difference is smaller than in low proficiency users. In the two groups of users mentioned above, the median of comprehension of documents retrieved with English queries is significantly lower than with Portuguese queries at $\alpha = 0.01$ (W=387138 in low proficiency users and W=2872800 in elementary proficiency users). For good English proficiency users, English queries have slightly higher comprehension scores but this difference is not significant.

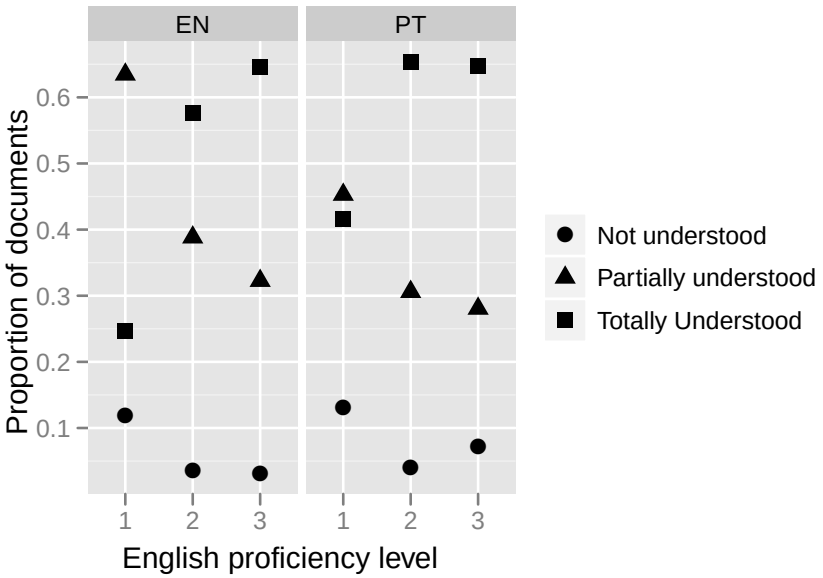

Figure 9.1: Proportion of documents by English proficiency level (low-1; elementary-2; good-3), query language and users' comprehension.

Since Portuguese queries retrieved both Portuguese and Spanish documents, we also analyzed how Spanish documents affected the reality described above. Spanish documents have lower comprehension and, in accordance with this, we noticed that these documents made the above differences smaller in the low and elementary proficiency users and higher in the good proficiency group. In other words, if we made the above comparisons considering only English and Portuguese documents, for low and elementary English proficiency users, the comprehension difference between Portuguese and English documents would be clearer and, in good English proficiency users, less clear.

In English queries, users with a higher level of English proficiency tend to evaluate the documents' comprehension higher than users in lower levels. We have applied the Kruskal-Wallis test and verified there are statistically significant differences in documents' comprehension between the three groups of users at $\alpha = 0.01$ (KW $\chi^2(2) = 431.4$). Further analysis with the Mann-Whitney test and the Bonferroni correction indicates that differences are significant, at $\alpha = 0.01$, between all levels of proficiency (Table 9.3). This agrees with what was expected.

Table 9.3: Differences of comprehension between levels of English Proficiency. R_i is the median of the comprehension in the proficiency level i.

	$R_1 < R_2$	$R_1 < R_3$	$R_2 < R_3$
EN	$W = 770864.5$	$W = 362368.5$	$W = 1517136$
	$p < 0.01/3^{**}$	$p < 0.01/3^{**}$	$p < 0.01/3^{**}$

9.4.4 Documents' readability

Documents readability was automatically evaluated using the SMOG metric and higher SMOG scores are associated with documents having lower readability. From the readability analysis, we excluded all documents for which we could not compute the SMOG readability metric; the assessments in which the user mentioned he could not access the content of the URL, although we did have access to it and did calculate the SMOG metric; and an English document, considered a severe outlier having a SMOG of 78.9 while the mean SMOG for English documents is 6.2. After a manual analysis we verified that this severe outlier was a document containing a set of sentences with the word inflammation with no logical sense. It looked like a work in progress document that was, accidentally, put online.

Different languages have different characteristics and therefore are associated with readability metrics of different magnitude. We empirically expect Portuguese documents to have a higher SMOG than English documents. Since readability measures are language dependent and Portuguese queries retrieved both Portuguese and Spanish documents, we will do this analysis by document language and not by the query's language. Also, and for the same reasons, we will only be able to compare the SMOG metric in documents of the same language. Our results show that Portuguese documents have a higher average SMOG (10% trimmed mean of 8.22) than English ones (10% trimmed mean of 6.19). Since SMOG is based on the number of polysyllables (words of 3 or more syllables), this indicates that Portuguese has more polysyllables than English, a statistically significant difference at $\alpha = 0.01$ (t(8210.7)=-35.12).

An analysis of the mean SMOG distribution by documents' comprehension shows that, in Portuguese and English documents, the degree of comprehension increases as SMOG decreases, this is, as the text becomes simpler. This is shown on Table 9.4. As expected, this difference is stronger with Portuguese documents where all users have similar competencies while, in English, different proficiencies may affect users' comprehension even in the simplest document. We only detected statistical differences in the mean SMOG between levels of comprehension in the Portuguese documents (KW $\chi^2(2) = 99.8$, $p < 0.01^{**}$). A pairwise comparison showed that only documents *fully understood* by users have a mean SMOG significantly lower than the other documents (Table 9.5).

In English documents, we did not find significant differences between levels of comprehension for each level of English proficiency.

In Figure 9.2, we present the distribution of the SMOG mean by level of relevance assessment and language of the document. As can be seen, Portuguese documents that are easier to read (lower SMOG) have higher relevance scores

Table 9.4: SMOG mean ($\overline{x}$) and standard deviation (s) by language and level of comprehension.

	Not understood		Partially understood		Totally understood	
	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s
EN	6.57	2.38	6.44	2.72	6.38	2.58
PT	8.65	2.24	8.40	1.89	8.10	2.20

Table 9.5: SMOG differences between comprehension levels. S_i is the SMOG mean for comprehension level i.

	$S_0 > S_1$	$S_0 > S_2$	$S_1 > S_2$
PT	$W = 193599$	$W = 446717$	$W = 1991804$
	$p = 0.02$	$p < 0.01/3^{**}$	$p < 0.01/3^{**}$

but this is only a tendency because the differences are not significant. Surprisingly, in English documents, there is an opposite trend that is significant (KW $\chi^2(2) = 28.4$, $p < 0.01^{**}$). A pairwise comparison showed that totally relevant documents have a higher SMOG mean than *non-relevant* ($W = 857138.5$, $p < 0.01/3^{**}$) or *partially relevant* ($W = 1060116$, $p < 0.01/3^{**}$) documents. This indicates that, in English documents, readability is not a major factor affecting relevance assessments.

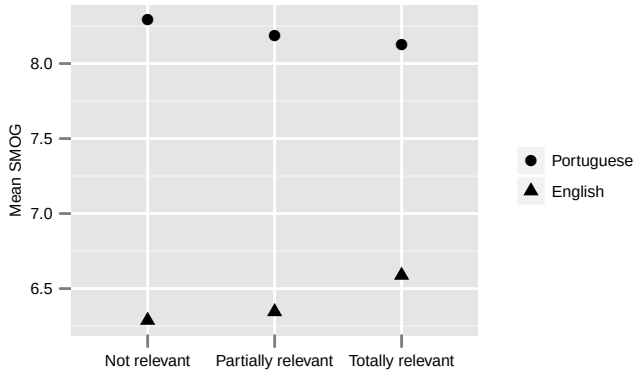

Figure 9.2: Mean SMOG by documents' relevance in each language.

Since we suspect this reality may differ for users with different proficiencies, we checked if there were significant differences in the SMOG mean between documents with different relevance scores, in each level of English proficiency. For users with low English proficiency, we found no significant differences ($F(2) = 2.12$, $p = 0.12$). For users with different English proficiencies, where significant differences were found (KW $\chi^2(2) = 13.86$, $p < 0.01^{**}$ in elementary proficiency and KW $\chi^2(2) = 12.1$, $p < 0.01^{**}$ in good proficiency), we run a set of pairwise comparison tests (Table 9.6). For elementary and good English proficiency users, the trend depicted in Figure 9.2 still applies, i.e.,

documents classified as totally relevant have a higher SMOG mean than documents assessed as *non-relevant* or *partially relevant*. This allows us to update our previous conclusion as follows: For users with elementary or good English proficiency, English documents' readability is not a major factor affecting relevance assessments.

Table 9.6: Differences of the mean SMOG between relevance scores in different levels of English Proficiency in English documents.

English Proficiency	$Rel_0 \neq Rel_1$	$Rel_0 < Rel_2$	$Rel_1 < Rel_2$
Elementary	$W = 244990.5$	$W = 241057.5$	$W = 302445.5$
	$p = 0.42$ (>)	$p < 0.01/3^{**}$	$p < 0.01/3^{**}$
Good	$W = 75703.5$	$W = 67203.5$	$W = 68625.5$
	$p = 0.13$ (<)	$p < 0.01/3^{**}$	$p < 0.05/3^{**}$

9.4.5 Medical accuracy

Not considering English proficiency, we can see in Figure 9.3 and Figure 9.4 that the query language does not affect the distribution of *medical accuracy* and *answer correctness*. In fact, in terms of these two variables, the proportion of answers in each level of classification is very similar in both languages. In terms of incorrectness, as shown in Figure 9.5, Portuguese queries show a better performance with a larger number of answers with *no incorrect content*. Yet, there are no statistical differences between the medians for incorrect content in both languages ($W = 12124.5$, p=0.19). Comparing the answers in terms of correctness and incorrectness, we can see that it is more probable to find an answer with *no incorrect content* than one with *appropriate content*.

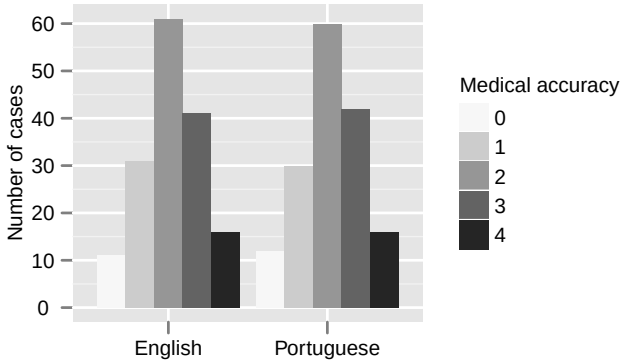

Figure 9.3: Answers' medical accuracy by query language.

English queries performed best for elementary English proficiency users where only 35% had an inappropriate answer in terms of answers' correctness. Good English proficiency users follow this group having 52% of the answers *with some value* in terms of correct contents. In low proficiency users, the proportion of inappropriate answers was high (56% with Portuguese queries

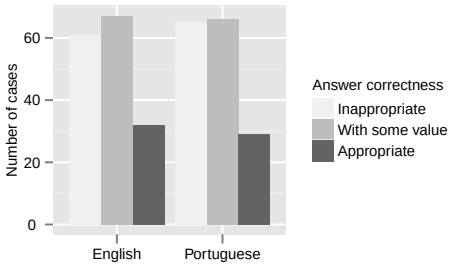

Figure 9.4: Answers' correctness by query language.

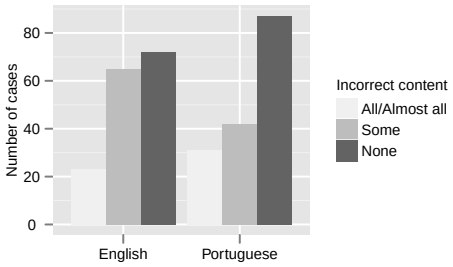

Figure 9.5: Answers' incorrectness by query language.

and 47% with English queries). Contrary to our expectations, in this group of users, Portuguese queries resulted in more inappropriate answers that English queries. All these differences show only a general tendency since the differences are not statistically significant, neither between levels of proficiency in each language, neither between languages at each level of proficiency.

In terms of answer incorrectness, the distribution is similar for all levels of English proficiency and is also similar to what was described when the English proficiency of the users was not being considered. For all levels of English proficiency, Portuguese queries result, in median, in answers with less incorrect content than English queries. Similarly to what happens in answer correctness, the best scenario, i.e., the scenario with less incorrect medical content, happens with English queries in the elementary English users.

Regarding the medical accuracy variable, for low English proficiency users, Portuguese queries result in more accurate knowledge than English queries (Figure 9.6). Since their comprehension in English content may be limited by their English proficiency, this matches our initial expectations. English queries result in more accurate answers for elementary and good proficiency users with less dispersion in elementary proficiency. Like in answer correctness, these differences are not statistically significant.

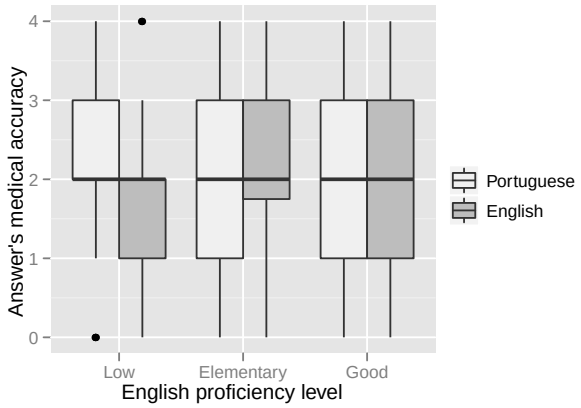

Figure 9.6: Medical accuracy boxplots by English proficiency and query language.

We also noticed that English queries led to more HONcode certified pages (60.8%) than Portuguese ones (52.5%), a significant difference ($\chi^2(1) = 89.7$,

$p<0.01^{**}$). An analysis by document language instead of query language shows that 61.1% of the retrieved English documents are HONcode certified, against 52% of the Portuguese documents, a statistically significant difference ($\chi^2(1) = 77.6$, $p<0.01/3^{**}$). Note that the reported percentages are above 50% because half of the documents are retrieved by the HON retrieval system and are certainly HONcode certified. Moreover, the Latin square procedure used in task assignment and the equal number of documents assessed in each task assure no biases are introduced in this analysis. In fact, the number of documents assessed with English queries in one system is equal to the number of documents assessed with English queries in the other which is equal to the number of documents assessed with Portuguese queries in either system.

9.4.6 Motivational relevance

To evaluate motivational relevance, we asked users to classify the search task completion status in a scale of 1 (completely unsatisfied) to 5 (completely satisfied) in the post-search questionnaire. Without considering the English proficiency of the users, we verified that the distributions of motivational relevance in English and Portuguese queries are very similar in terms of central tendency and dispersion. Both distributions have 4 as median. The main difference between them lays in the number of low outliers in the first level of motivational relevance which is greater in Portuguese queries. In other words, with Portuguese queries, users feel *completely unsatisfied* more frequently. For low English proficiency users, the median level of search task's satisfaction is lower (3) than in users with higher proficiency (4 in both elementary and good proficiency levels). Only the former type of users is *completely unsatisfied* in tasks with English queries. Moreover, low English proficiency users constitute the only group that is never *completely satisfied* with English queries.

Hypothesis testing allowed us to conclude that differences between languages for each level of English proficiency are not significant. However, significant differences were found between the motivational relevance median for the three levels of English proficiency ($KW\chi^2(2) = 9.93$, $p=0.00^{**}$ in English queries and $KW\chi^2(2) = 6.41$, $p=0.04^{*}$ in Portuguese queries). To determine the exact location of the differences, we did a pairwise comparison with the Bonferroni correction (Table 9.7). We found evidence to conclude, at different significance levels, that low proficiency users feel less satisfied than elementary and good English proficiency users with English queries.

Table 9.7: Differences of Motivational Relevance (MR) between levels of English Proficiency.

	$MR_1 < MR_2$	$MR_1 < MR_3$	$MR_2 \neq MR_3$
EN	$W = 884$	$Z = -2.34$	$W = 1928$
	$p < 0.01/3^{**}$	$p < 0.05/3^{*}$	$p = 0.66$
PT	$W = 968$	$Z = -1.41$	$W = 2056$
	$p < 0.05/3^{*}$	$p = 0.08$	$p = 0.26$

9.5 QUERY FORMULATION BEHAVIOR

In the initial questionnaire, users were asked to introduce the query they would use for the information needs triggered by the information situations. Then, after each assessment task, users were asked to write two additional queries. Each of these queries was manually analyzed in terms of number of terms, language and the existence of misspelled terms.

Regardless of their English proficiency, all users formulated an initial query in Portuguese. The mean number of terms was 4.13 (SD = 1.83). In this initial query, we found 4 queries (1.25%) with misspelled terms.

When users have completed an assessment task having a Portuguese query, the subsequent queries were mainly in Portuguese. In users with low and good English proficiency, 100% of the subsequent queries were Portuguese and, in elementary proficiency users, this proportion downs to 99% in the 2nd query and 98% in the 3rd query.

After English tasks, as expected, users more frequently formulate English queries. From a global perspective, 3.8% of the English tasks had both subsequent queries in English and, in 56% of the tasks, one of the queries was in English. In Table 9.8 we present an analysis by English proficiency in which we cannot detect an association between English user proficiency and the use of English to formulate queries. Regarding errors, 3.9% of subsequent Portuguese queries had misspelled terms against 4.3% of the English queries.

Table 9.8: Post-search queries in English after an English assessment task by user proficiency.

#English queries	Low EP	Elemen. EP	Good EP
2	6%	3%	4%
1	55%	57%	55%
0	39%	40%	41%

9.6 ENGLISH PROFICIENCY PREDICTION

In this section, we explore the associations between users' habits with respect to searches in English and their proficiency in this language. If a relationship is found, hypothesis can emerge and be tested in future studies using the search logs to predict language proficiency through past queries in English.

In the initial questionnaire, users were asked how often they conducted their health searches in Portuguese and in English. Since only users that had previously conducted a health search have answered this question, we regret we did not ask about their general behavior instead of focusing only on health searches. This answer was given in a scale of 1 (never) to 5 (always) in each language. Excluding users that did not answer, almost all said they search *always* or *almost always* in Portuguese. For low and good English proficiency levels, 100% of the users chose one of these two options and, in elementary proficiency users, 4.8% also answered *sometimes*.

When asked about their rate of English health searches, a large proportion of answers were concentrated on the opposite side of the scale. The majority

(62.5%) of the low English proficiency users said that *never* or *almost never* did it, while in elementary proficiency users this proportion downs to 56.3% and in the good English proficiency it downs even further to 50%. This shows a general tendency but there are no significant differences in the median between users of different English proficiencies (KW $\chi^2(2) = 1.2$, p=0.5).

9.7 DISCUSSION AND IMPLICATIONS

Through a user study we have investigated, from several perspectives, the impact of translating queries to English, in health IR search tasks done by users with different levels of English proficiency. We have also analyzed query formulation and reformulation behavior according to users' English proficiency and the main language of the previously assessed documents. Finally, we have explored the existing relations between search habits and English proficiency.

As a result of the analysis of the documents in the search results, we can conclude that the quality of web health information is better in English than in Portuguese. First, English queries retrieve a smaller proportion of pages with enduring HTTP errors. Second, when compared with Portuguese queries, English queries retrieve a larger proportion of webpages (96.3%) and a smaller diversity of document types (only webpages and *pdf*). Since we expect content built for the dissemination of consumer health information to be in a webpage format, we conclude that English queries retrieve more documents built specifically for the dissemination of health information on the Web. Finally, we also found that English pages have a significantly higher proportion of HONcode certified pages. This shows that the use of an English query is associated with a higher probability of retrieving certified content. This conclusion is independent of user features and reinforces one of the assumptions of this study: a larger quantity of information in English may mean an easier access to higher quality content.

For all levels of English proficiency, English queries have a significantly higher precision, independently of the used measure. Since we also found that low proficiency users feel less satisfied than elementary and good users with English queries, we conjecture that low proficiency users assess topic relevance, that is, "the relation between the query's topic and the documents' topic" (Saracevic, 1996), instead of situational relevance. We think the proficiency of these users is sufficient for them to identify if a document is about a certain topic but is not enough for them to understand the main message, what explains the lower satisfaction rates. This is also consistent with the results we have found in terms of medical accuracy.

Since all users have Portuguese as their native language we expected users to rate the comprehension of Portuguese documents higher than the comprehension of English documents, and this was confirmed in the low and elementary proficiency levels. In good proficiency users, the difference between both languages is not significant. This makes sense since their English proficiency is closer to the overall Portuguese proficiency. As expected, in English documents, comprehension increases with English proficiency.

In terms of readability we found that comprehension increases as the documents become easier to read. Since we only detected significant comprehension differences by readability in Portuguese documents, we suspect this is a

factor that only comes into play if language proficiency is guaranteed. The relation between relevance assessments and documents' readability shows that, for English documents and elementary and good English proficiency users, totally relevant documents have a significantly lower readability than partially and non-relevant documents. Either this means that, in these users and language, the presence of more scientific terminology boosts relevance or that readability is not one of the major factors determining the documents' relevance.

English queries tend to result in more accurate answers in elementary and good proficiency users, with less dispersion in elementary proficiency. For low English proficiency users, Portuguese queries tend to result in more accurate knowledge than English queries. This probably happens because their comprehension of English documents is limited by their proficiency.

None of the users formulated an initial query in English and few formulated the subsequent queries in this language. After performing English tasks, users formulate subsequent queries in English more frequently, regardless of their English proficiency level. This indicates that suggesting alternative English queries or even incorporating English documents in the answer set may also be a good way to trigger ideas on how to express the information need. The detection of more misspelled terms in English than in Portuguese may indicate that English terms for health concepts may not always be known or recalled. This adds value to suggestions of English translations of health concepts.

We believe English proficiency may be inferred from past search behaviors. For users with higher proficiency, we found an increased tendency to use English queries but, since differences are not significant, this is not conclusive. A specific study has to be done with this goal.

As shown in the previous paragraphs, English queries consistently have better results for users with at least elementary English proficiency. On the other hand, Portuguese queries behave better for users with low English proficiency, resulting in more accurate answers and a higher overall satisfaction with the search task. Together with the higher quality of English health web content, these findings confirmed our initial expectations and show that English content may and should be used to help users with other native languages and enough English proficiency. Existing Web search engines may use these conclusions to define personalized strategies that help users access English content when they formulate queries in their native language. These strategies may involve the user in the process or can be totally automatic. In the first case, alternative English queries can be suggested to the user who determines if he wants to use them or not. The alternative query may simply be a translation of the original query to English or may also include other variants of the English query through the inclusion/replacement of synonyms. Totally automatic strategies may be implemented through the inclusion of English content in the result set of the query in the users' native language. Merging the results set of the original query and of its English translation may be a good strategy. Although more important in the totally automatic strategies, personalization is also essential in the other strategies to avoid unnecessary distractions, users' waste of time and the overload of the search interface. Even though this study does not allow generalization of these results to languages other than Portuguese, we believe the conclusions of this study could still apply to several

languages with small presence on the Web and we would like to test it in future studies.

9.8 CONCLUSION

English is by far the most used language on the Web which has, therefore, a larger proportion of English health content. We observed that English health content has a larger proportion of health-certified documents, is more suited to disseminate health information and is associated with less HTTP errors. For these reasons, we are convinced this can be explored to provide a better service to non-English speaking users. Difficulties expressed by users on health searches strengthen our conviction. Yet, we are aware that the approach has to be personalized to users' English proficiency. Results suggest that translation approaches should be used only on users with at least elementary English proficiency. As revealed by this study, despite the higher precision of English queries in low English proficiency users, these users have a lower degree of comprehension of English documents, obtain less accurate knowledge through English queries and feel less satisfied in the tasks with this type of queries. Although some of these results are not surprising, we consider important to have an empirical demonstration of these facts.

We also found that the readability of documents should be a criterion for ranking, especially if the user is proficient in the documents' language. Moreover, we found that a more complex terminology may inspire confidence in the retrieved documents but this conclusion has to be further explored. These findings suggest that a cross-lingual assistance personalized to users' English proficiency could improve non-English consumer health retrieval and could be helpful in an educational sense, enabling non-English speaking users to learn English medical terminology. Moreover, it may also be helpful to trigger new search strategies and to help the user construct queries that give access to documents that may not be reached otherwise.

The next chapter describes the other main study that derives from the experiment reported in Chapter 8. In that study, we explore the impact of queries' terminology, namely the lay and medico-scientific ones, in users with different health literacy and topic familiarity.

EFFECTS OF QUERY TERMINOLOGY ON HEALTH SEARCHES: AN ANALYSIS BY USER'S HEALTH LITERACY AND TOPIC FAMILIARITY

10.1 INTRODUCTION

Although most users are satisfied with their health searches, some get frustrated or confused (Fox, 2006; Petrock, 2010). This happens more in individuals with less education as showed by the Pew Internet report (Fox, 2006): 22% feel frustrated by the inability to find what they want (27% in those without a college degree and 18% in those with a college degree) and 18% feel confused with what they find online (24% in those without a college degree and 15% in those with a college degree). Since the educational level has a strong impact on health literacy, this is not surprising.

The widespread use of the Web to retrieve health information implies a large diversity of users performing this type of task. One characteristic that is expected to differ between users is their health literacy, a differentiation that can be caused by differences in age or education. A study that assessed the usability of 125 websites offering health resources reported that about one third of these sites "required a college education to comprehend extracted health information" (Becker, 2004).

The mismatch in the languages used by health consumers and health professionals also poses a barrier to an effective access to relevant information (Zielstorff, 2003). Since information may be presented at a high reading level and include medical jargon (Cline and Haynes, 2001), the ability to understand the retrieved information may fail and, if so, user's satisfaction may be at risk. In fact, this is one of the typical problems felt by consumers when performing health information searches (Kogan et al., 2001). Other popular problems are: the difficulty or even inability to formulate a health query from an information need due to the lack of proper medical terms (Zhang, 2010; Toms and Latter, 2007) and the difficulty to formulate it without misspellings or use of wrong medical terms (Kogan et al., 2001; McCray and Tse, 2003).

In this research we study the effect of translating query terms between lay and medico-scientific terminologies, in users with different characteristics, namely, health literacy and topic familiarity. We believe that a user model that considers the above context features may be used to improve health information retrieval through, for example, the suggestion of alternative queries or by re-ranking results. The work presented here is the first to consider user

context features while studying the impact of a query processing technique in several aspects of the retrieval process. Moreover, the evaluation considers not only users' relevance assessments as several existing works but also the quality of the medical knowledge that emerges from the search session.

The remainder of this chapter is structured as follows. We start to present the existing related research and then the research questions that guided this investigation. The following sections have a detailed description of our findings, which are discussed afterwards.

10.2 RELATED WORK

In this section we start by describing works that explore medico-scientific terminologies with the goal of improving health IR. Next, we focus on IR works that explore the two main context features used in this work, namely health literacy and topic familiarity.

10.2.1 Exploration of medico-scientific terminologies in Health Information Retrieval

It is known that there are mismatches between consumers' terminology and the one used in health documents and standard medical vocabularies (Zeng et al., 2002; Eerola and Vakkari, 2008). To evaluate the impact of these mismatches, Plovnick and Zeng (2004) compared the performance of consumer queries with the performance of the same queries reformulated with terminology from the UMLS. Each query was submitted to Google and MedlinePlus and the relevance was assessed comparing results with a gold standard answer. The authors used P@30 to compare both type of queries and, through descriptive analysis, concluded this type of reformulation may be a promising strategy to improve consumer health-information searches. Previous studies (Patrick et al., 2001; Zeng et al., 2002) have reached similar conclusions. Patrick et al. (2001) compared the performance of lay and medico-scientific queries on the retrieval of diabetes-related web information. The evaluation was based on the number of sites maintained by non-profit healthcare professional organizations, academic organizations, or governmental organizations that appeared in the top-20 results. Authors found the number of such sites was lower with lay queries. While studying the characteristics of consumer terminology for health IR, Zeng et al. (2002) concluded that 51% of the lay queries returned no information, though matching information existed in the database.

Considering the poorer results of lay queries and the fact that non-experts use medico-scientific terminology less often than experts (White et al., 2009, 2008), it is expectable that a comprehensive terminology support improves health IR (Zeng et al., 2002). Some works therefore propose and evaluate strategies to translate lay terms to medico-scientific ones (Lu et al., 2006). Others go further and present query suggestion systems (Zeng et al., 2006; Luo et al., 2008; Luo, 2009) and others come up with ways to identify the mixture of terminologies in order to minimize the language gap and improve health IR (Crain et al., 2010). These works are briefly described next.

Lu et al. (2006) translated query terms from lay to professional ones in the context of cross language health IR (CLHIR). If the lay term appears in

the Medical Subject Headings (MeSH) thesaurus, an immediate translation is made. If not, the authors propose an approximate string matching of the non-professional terms to the professional ones. In the other cases, they propose to use Web resources with the argument that an increasing number of sites contain laypersons' terms and their corresponding professional terms. Their evaluation showed improvements on the performance of MeSH concept mapping and CLHIR.

Luo et al. (2008) and Luo (2009) propose and evaluate two similar search engines for health IR: MedSearch and iMed. Both search engines accept long queries and transform them to shorter ones by extracting the most representative terms. Moreover, they suggest medical phrases to help the user digest the retrieved documents and refine the query. These phrases are extracted and ranked based on the MeSH, the collection of crawled webpages and the query. In addition, to help users provide information about their medical situation, iMed uses a questionnaire-based query interface. MedSearch was evaluated with questions posted on medical discussion forums and assessments from five non-medical persons. iMed was evaluated with real medical case records from the Family Medicine Online Database (FMOD) and medical exam questions that had answers available as the ground truth. In both cases, the experiments showed that the search engines handle medical queries effectively and efficiently. The Health Information Query Assistant (HIQuA) system, developed by Zeng et al. (2006), suggests alternative query terms, selected according to their semantic distance to the user's initial query terms. Queries are first mapped to one or more concepts of the UMLS and then the semantic distance between concepts is calculated based on co-occurrences in medical literature, log data and on UMLS semantic relations. Authors found statistically significant higher rates of successful queries, that is, queries with at least one relevant result on the top-10, but no statistical differences on user satisfaction or users' ability to complete the task.

Crain et al. (2010) propose to overcome the language gap between lay and medico-scientific terminology with a Bayesian model that, given a document, can infer the mixture of topics and dialects (slang, common and technical) and the most likely topic and dialect of each word. Authors found a 25% improvement in nDCG@5 when using this model to support health IR.

The interplay between user contextual features and the terminological aspects of health IR is less explored in the existing literature. From the works mentioned above, only the health search engine described by Luo (2009) collects and uses information about the user through the questionnaire-based interface. Another study investigates the effect of user factors on the familiarity with health terms and uses gender as a proxy for background knowledge about gender-specific illnesses (Keselman et al., 2006). Authors recruited a convenience sample of 50 users and designed an instrument to test users' familiarity with 27 health terms of different "familiarity likelihood scores" and three categories: "male", "female" and "neutral". This study's findings support the idea that background knowledge and experience affect users' familiarity with health terms. Moreover, authors conclude that health literacy is another variable expected to influence familiarity. A more recent article (Zeng-Treitler et al., 2008) uses context to estimate consumer familiarity with health terminology but the explored features are not related to the user. In the proposed method, the authors use a network in which each node represents a term, and

each term is connected with other terms that co-occur with it. The context of a term can be a query session, a sentence, a paragraph, or a document. The method was applied to query logs and was validated using results from previous consumer surveys. Authors concluded that this method is a good alternative to existing term familiarity assessment methods.

10.2.2 Health literacy in Information Retrieval

On the health domain and according to the USA Department of Health and Human Services (2000), health literacy is "the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions". The 2003 USA assessment of adult literacy (Kutner et al., 2006) found that 36% of adults in the United States have basic or below basic health literacy skills. McCray (2005) does a good review of the literature about health literacy. According to this author, a substantial portion of the literature addresses the mismatch between the health literacy of the patient and the readability of the documents.

To the best of our knowledge, few IR studies consider user's health literacy. In the Health Information Query Assistant study previously described, Zeng et al. (2006) empirically concluded that query recommendations are not adequate for inadequate health literacy users. A work from Wang and Liu (2005) describes a personalized health IR system that adjusts results to users' health literacy level, but no evaluation was performed. Summers and Summers (2005), with the goal of making web health contents more usable and accessible for users with low health literacy, concluded that this type of users often avoid search due to its spelling and typing requirements. They also concluded that these users have difficulties processing search results pages.

10.2.3 The influence of topic familiarity in Information Retrieval

Topic familiarity, or domain knowledge as it is also frequently referred to, can be defined as the user's general knowledge about the topic of a search task. It is acknowledged that topic familiarity can be an important factor in IR (Capra and Pérez-Quiñones, 2006) and there are several research works that explore this feature. These works can be grouped in 4 major categories: studies that analyze how topic familiarity influences information search behavior; works that analyze how it influences query formulation, with or without system' suggestions; studies that analyze its relation with information retrieval performance and, finally, research that uses it to rank retrieved documents.

As can be seen in Table 10.1, studies investigating the relationship between topic familiarity and information search behavior are based on user studies and all evaluate the familiarity with the topic through users' self assessment. They differ on the type of analyzed behaviors and, typically, these behaviors are acquired through log records of the user study. The conclusions of these studies allow us to state that, as the familiarity with the topic increases, so does the search efficacy. Moreover, the resources the user values become more specialized, the user's effort (task completion time and number of queries) decreases and the importance given to certain relevance criteria change. As can be seen through the studies described in the rest of this section, the performance conclusions are not always consensual.

Table 10.1: Research relating search behavior with topic familiarity.

Study	Behaviors	Methodology	Conclusions
(Kelly and Cool, 2002)	Reading time and efficacy.	User study with 36 subjects; familiarity acquired from post-search questionnaires; reading time obtained from the logs; efficacy is #docs saved/#docs viewed.	As topic familiarity increases, search efficacy increases.
(Wen et al., 2006)	Resources and relevance criteria.	User study with 18 subjects; 2 search tasks, one familiar and the other non-familiar; familiarity assessed through a questionnaire; important relevance criteria identified by the users from a list of 12.	Unfamiliar topic leads to more generic and fewer specialized resources. Different relevance criteria on less familiar topics: accuracy and accessibility less valued; consistency with other information more important.
(Liu and Belkin, 2010)	Decision Time - the duration from opening a document to the user first starting to use, save, or leave it.	3-session lab experiment with 24 users; familiarity assessed by the user; time data acquired from log files.	For documents with a certain degree of usefulness, users with different levels of topic knowledge had different Decision Times.
(Qu et al., 2010)	Completion time, number of queries and website used to start the search.	User study with 30 subjects and 2 search tasks; familiarity assessed by the users.	Lower topic familiarity leads to longer completion time and more queries but does not affect the search entrance.

Research dedicated to query formulation behavior can be split in two large groups, one that analyzes users' query formulation habits without the system's interference and another that explores users' behaviors during query expansion. In a longitudinal study, Wildemuth (2004) analyzes the search terms used by medical students on six clinical problems in three occasions, one before students received any instruction on the topic, the second just after the course on the topic and the third occurred six months after the end of the course. Wildemuth (2004) concluded that, when domain knowledge was very low (first assessment), users made more moves, i.e., additions and deletions of concepts to the query. This is probably due to their initial inability to choose the appropriate terms and is in accordance with the conclusions of the study from Qu et al. (2010). Finally, Wildemuth (2004) also concluded that, although it improved performance in all occasions, system assistance during query formulation is more useful when users have less knowledge on the topic. This work has also a good research review on the effects of domain knowledge in information retrieval.

Two studies explore the influence of topic knowledge on the use of a thesaurus for query expansion. In the first, Sihvonen and Vakkari (2004) conducted a user study with 15 users with knowledge on the topic and 15 users without it. Results were acquired through search logs and interviews with the subjects. Authors concluded that the use of the thesaurus was helpful for the experts but not for the novices to improve search effectiveness. Search success was measured as the number of references retrieved that were judged relevant by external experts. This conclusion contradicts Wildemuth (2004) conclusions. In the other study, Shiri (2005) analyzed how topic's familiarity affected users' behavior on thesaurus' use and concluded that "searches involving mod-

erately and very familiar topics were associated with browsing around twice as many thesaurus terms as was the case for unfamiliar topic".

Studies analyzing the influence of topic familiarity on IR performance focus on different aspects. Liu and Belkin (2010) focused on documents' usefulness, wanting to know if topic knowledge could be used to predict it. Kelly and Cool (2002), also on Table 10.1, considered efficacy as the ratio between the number of documents saved and the number of documents viewed. Al-Maskari and Sanderson (2010) investigated factors influencing user satisfaction and they found no relationship between familiarity and satisfaction. They also found no significant differences between familiar and unfamiliar users in the number of relevant documents identified by the users, the number of TREC relevant documents and the time taken by the user to locate the first relevant documents. The same authors conducted a user study (Al-Maskari and Sanderson, 2011) with 56 subjects and 56 topics from the TREC collection to analyze the influence of users' cognitive skills on user effectiveness. They asked users to assess familiarity after completing the search for each topic and found no significant correlation between familiarity and users' perceptual speed. Muresan et al. (2006) used the TREC HARD track (Allan et al., 2003) to examine the impact of document characteristics like readability and concreteness/abstractness on document relevance assessments by users with different levels of familiarity with the topic. Authors concluded that a higher readability has positive effects on retrieval performance, regardless of user's familiarity with the topic.

In a slightly different type of work, Kumaran et al. (2005) define a model of topic familiarity required to read a document that is used to classify pages as introductory or advanced ones. The classifier uses features like the document reading level, the distribution of stop-words and non-text features like the average line-length. This classifier was used to re-rank the result set showing introductory documents higher in the rank. This work does not consider the user's familiarity with the topic but shows it is possible to distinguish documents by the familiarity required to read them in order to re-rank results according to users' characteristics.

That we know of, no studies consider users' topic familiarity in health IR.

10.3 RESEARCH QUESTIONS

The following research questions (RQ) drove our research. They are similar in their aim but differ in the object of analysis.

- What is the impact on the characteristics of the retrieved documents of replacing lay query terms by medico-scientific ones? (RQ1)
- What is the impact on search task precision (RQ2), users' comprehension of documents (RQ3), accuracy of the medical knowledge (RQ4), task completion status (RQ5) of replacing lay query terms by medico-scientific ones, in users with different levels of health literacy and topic familiarity?

RQ1 does not consider HL or TF, because the characteristics of the retrieved documents are the only surveyed feature that does not depend on the user.

To answer our research questions we conducted the laboratory user study described in Chapter 8. From the context features presented in Table 8.3, in this study we have used the following ones: health literacy, relevance, comprehension, readability, type of document, HONcode certification, for consumers?, for professionals?, answer's medical accuracy, combined topic familiarity and motivational relevance.

10.4 DATA ANALYSIS

In this section we use descriptive and inferential statistics. To visualize differences between populations we use boxplots that graphically describe variables and their dispersion depicting the 25th percentile (Q_1) subtracted of 1.5 of the interquartile range, Q_1, the median (Q_2), the 75th percentile (Q_3), Q_3 plus 1.5 of the interquartile range and outliers.

In terms of inferential statistics we followed the strategy presented in Figure 10.1. Whenever possible we applied a parametric test instead of a non-parametric due to the former's greater statistical power. The selection of the hypothesis test depends on the number of groups to be compared and on the scale of the variable that is being compared. Whenever a nominal variable is involved, as happens in almost all documents' characteristics, we use the test of equal proportions with the chi-squared value. Note that, when comparing two samples, the chi-squared test for equality of two proportions is the same thing as a z-test since the chi-squared distribution with one degree of freedom is the square of a normal deviate one. In situations where ordinal variables are involved, we employed the Mann-Whitney test and used the W letter to indicate the test value. In variables with a ratio scale, whenever it was possible we applied the t-test. In the other situations we applied the Mann-Whitney test. The only exception occurs in the SMOG analysis where the differences in the variance of both groups made us apply the Welch t-test, an adaptation of the t-test intended for use with two samples with unequal variances. When more than two groups are being compared we initially applied the one way ANOVA or the Kruskal-Wallis test (KW) to verify if there were significant differences between the groups and, if so, we either applied the Tukey's test or we did a pairwise comparison. In the pairwise comparison we applied the Bonferroni correction, dividing α by the total number of comparisons to minimize the type I error. These comparisons allowed us to detect where are the differences are located.

10.4.1 Document Characteristics Analysis

This study involved the evaluation of 1652 URL. From these, 879 were retrieved through queries with lay terminology and 886 through queries with medico-scientific terms, with 113 URL being retrieved through both types of queries. As can be seen in Table 10.2, queries with medico-scientific terminology led to more HTTP Errors and more "no content" errors than queries without it. However, none of these differences is statistically significant.

In terms of document type, both types of queries retrieved mostly webpages (Table 10.2). This proportion is significantly higher in the first type of queries. The Portable Document Format (pdf) is the second most common type of document in both types of queries and its proportion is significantly

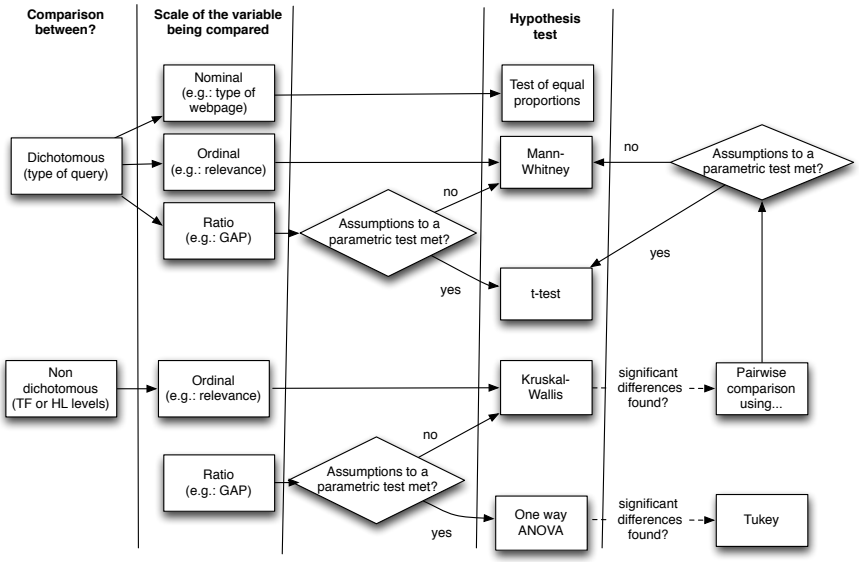

Figure 10.1: Inferential statistical strategy.

Table 10.2: Differences in documents' features for both types of queries.

	Feature	Lay queries	Medico-scientific queries	Significant differences?
Errors	HTTP	0.57%	1.02%	No
	No content	0.23%	0.68%	No
Document Type	Webpages	95.8%	86.22%	$\chi^2(1) = 47.17$, $p = 3.25 \times 10^{-12}$
	PDF	4.13%	12.7%	$\chi^2(1) = 40.78$, $p < 8.54 \times 10^{-11}$
	PowerPoint	0.11%	0.57%	No
	Word	0%	0.46%	No
HON Certification	Certified pages	53.01%	52.71%	No
	Consumer pages	33.9%	26.64%	$\chi^2(1) = 10.7$, $p = 5 \times 10^{-4}$
	Professional pages	5.12%	14.67%	$\chi^2(1) = 44.02$, $p = 1.62 \times 10^{-11}$
Readability	SMOG (mean)	7.17	7.69	Welch's t(8751.1)=-9.06, $p < 2.2 \times 10^{-16}$

higher in queries with medico-scientific terminology. Just like pdf documents, PowerPoint and Word documents are more frequent in queries with medico-scientific terminology, yet neither of these proportions is significantly different between types of queries. The larger proportion of webpages retrieved by lay queries indicates that this type of queries retrieves more documents specifically built for the dissemination of health information on the Web.

The proportion of HONcode certified pages is very similar in both types of queries (Table 10.2). As expected, queries with lay terminology retrieved more consumer-oriented documents, a difference that is statistically significant. In terms of medico-scientific documents the opposite happens, that is, queries with medico-scientific terminology retrieve more medico-scientific documents, also a statistically significant difference.

As explained before, documents' readability was assessed through the SMOG

metric. Overall, SMOG ranged from 3.71 to 33.09 with a mean of 7.94 (sd=2.35). As expected, documents retrieved with queries containing medico-scientific terminology are more difficult to read (Table 10.2). Since the two samples are not homogeneous in variance, we used the Welch's t test and verified that the difference between both medians is statistically significant.

Discussion

Our findings show that, replacing lay query terms by medico-scientific ones results in retrieving a smaller proportion of webpages, a large proportion of pdf, less consumer-oriented documents, more professional-oriented documents and less readable documents. The smaller proportion of webpages indicates medico-scientific queries retrieve more documents not specifically built for the dissemination of health information on the Web. The HON categories more associated with the documents retrieved with medico-scientific queries show that the user has to be better prepared to access their contents. This is confirmed by these documents' lower readability.

10.4.2 Precision Analysis

We use Graded Average Precision (GAP) and Graded Precision (gP) with an equally balanced g_1 and g_2 to evaluate and compare precision. These measures are described in Section 5.2.1.

The precision measures mentioned above are based on the participants' relevance assessments. Like Borlund (2003b), we assume these assessments represent the value of the documents for a particular user at a particular moment and thus can only be made by the user at that time. Additionally, while the current practice in IR involves the use of a gold standard to compute precision, we intentionally decided to do it this way because we are not interested in topical relevance as classic works usually are. Instead, we are interested in situational relevance that encompasses cognitive relevance and can only be assessed through user judgments. Saracevic (1996) distinguishes these types of relevance as:

- Topical relevance: the relation between the query's topic and the documents' topic.
- Cognitive relevance: the relation between the state of knowledge and cognitive information need of a user, and the retrieved documents. It can be inferred from criteria like cognitive correspondence and informativeness.
- Situational relevance: the relation between the task at hand and the retrieved documents, being inferred by criteria like usefulness in decision making, appropriateness of information in resolution of a problem and reduction of uncertainty.

Only studying situational relevance we can fully explore the influence of health literacy and topic familiarity. For example, documents about the topic that are not understood by the user should not be considered useful for the situation at hand.

We start by a global analysis that does not consider user context features. In Figure 10.2 we present six boxplots. For each of the three precision measures, a boxplot is presented for each type of query, with and without medico-scientific terms. We can see that queries containing medico-scientific terms tend to have higher precision with every measure. However, the only significant difference was found with gP10 at α=0.05 (t(317.6)=-1.70, p=0.045). This means that, in the top-10 results, medico-scientific queries retrieve a higher proportion of relevant documents.

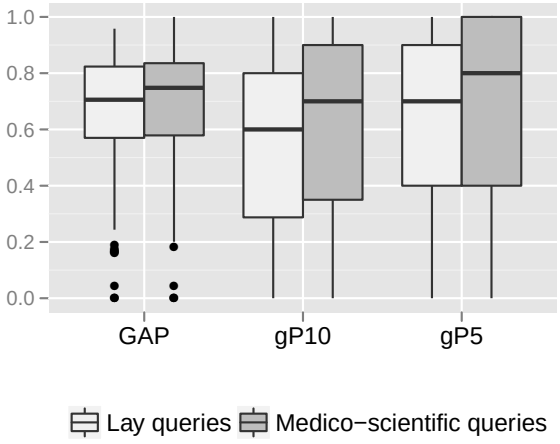

Figure 10.2: GAP, gP10 and gP5 boxplots by type of query.

GAP distribution by health literacy and query type can be visualized in Figure 10.3. Similarly to the global tendency, GAP tends to be higher in queries with medico-scientific terminology in every level of health literacy. However, we found no significant differences in GAP between types of query in each level of health literacy. We also tested if the mean GAP of each type of query had differences between levels of health literacy and did not find any significant ones. Although non significant, the higher GAP of medico-scientific queries in the lowest level of health literacy surprised us. Comparably to what happens with GAP, we found no statistically significant differences with gP5 and gP10 between types of query in each level of health literacy and between health literacy levels in each type of query.

In Figure 10.4 we present GAP distributions per query type and topic familiarity. As with health literacy, there is a tendency to have higher GAP in medico-scientific queries in all levels of topic familiarity. However this is just a tendency since none of the differences is statistically significant. Similarly, there are no significant differences in the mean GAP between levels of topic familiarity in each type of query. In terms of gP5 and gP10, the only statistically significant difference was found on the "somehow familiar" level using gP5. We found that users of this level have sessions with higher gP5 mean with queries using medico-scientific terminology than without it (t(111)=-2.1, p=0.019).

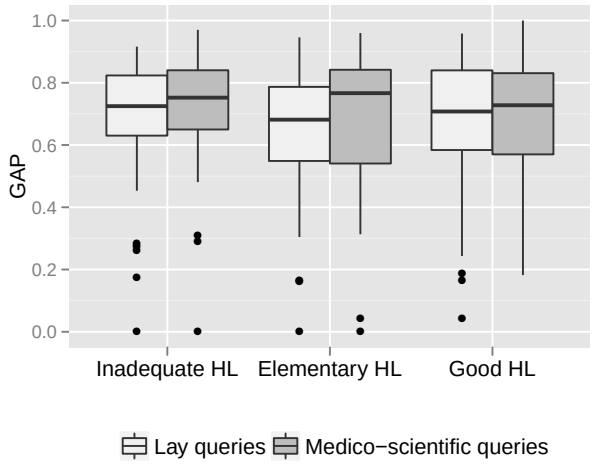

Figure 10.3: GAP by type of query and health literacy level.

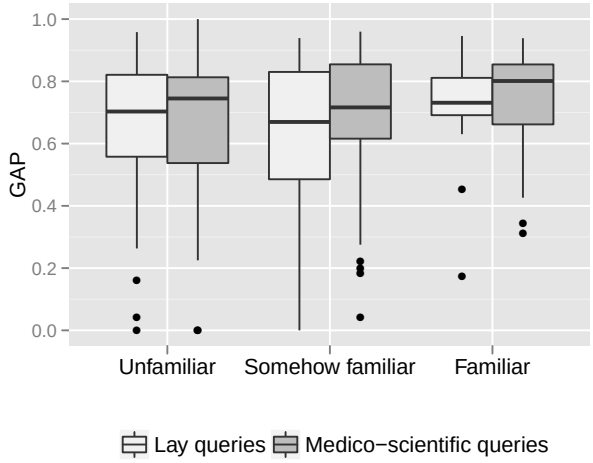

Figure 10.4: GAP boxplots by type of query and topic familiarity level.

Discussion

Medico-scientific queries show a higher precision in the top-10 retrieved results. This agrees with previous studies (Plovnick and Zeng, 2004; Patrick et al., 2001; Zeng et al., 2002) that, through descriptive statistics, concluded that this type of queries leads to better results. The analysis by users' health literacy revealed no significant differences in all the comparisons made. This partially surprised us because we expected medico-scientific queries to have lower precision than lay queries in users with inadequate health literacy levels. Regarding the topic familiarity analysis we found that medico-scientific queries have a higher precision in the top-5 retrieved results than lay queries, on users "somehow familiar" with the topic. We found no significant differences in the mean GAP between levels of topic familiarity in each type of query and this agrees with Al-Maskari and Sanderson (2010) who found no significant differences between familiar and unfamiliar users in the number of relevant documents.

10.4.3 *Comprehension Analysis*

We can say that, in general, users understand documents well because the comprehension median is 2 (Totally understood) in a scale of 0 to 2. However, if we repeat this analysis by query type, we see that in lay queries the median is still the same but in medico-scientific queries it drops to 1. These medians are significantly different (W = 13025482, p < 2.2x10-16).

In Figure 10.5 we present the proportion of documents by level of health literacy, query type and comprehension level. In this figure we can see that, when compared with medico-scientific queries, comprehension is higher in documents retrieved with lay queries in every level of health literacy. Not only "totally understood" appears more often in lay queries but "not understood" documents also appear less. As can be seen in Table 10.3, all these differences are statistically significant.

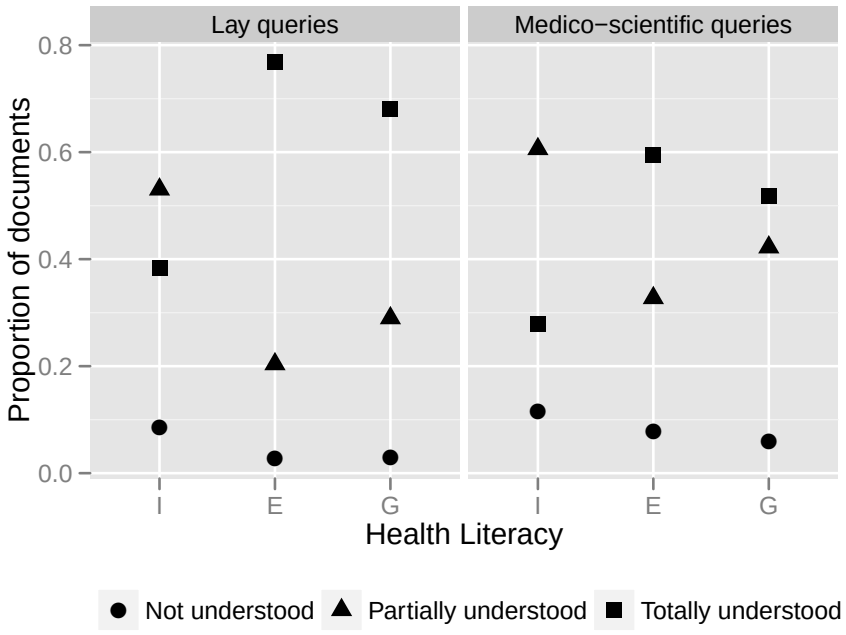

Figure 10.5: Proportion of documents by health literacy (I-Inadequate, E-Elementary, G-Good), query type and comprehension level.

Table 10.3: Significant differences between the median of comprehension in both types of queries, by health literacy level.

	$Comp_{Lay} > Comp_{MS}$	
	W	**p-value**
Inadequate HL	W = 636653	$p = 1.104 \times 10^{-7}$
Elementary HL	W = 1398119	$p < 2.2 \times 10^{-16}$
Good HL	W = 2645866	$p < 2.2 \times 10^{-16}$

Moreover, Figure 10.5 also shows that users with higher health literacy "to-

tally understand" more documents than users with inadequate health literacy, in both types of queries. The opposite happens with "not understood" documents. Using the Kruskal-Wallis test, we verified there are statistically significant differences in document's comprehension between levels of health literacy ($KW\chi^2(2) = 440.36$, $p < 2.2\times10^{-16}$ in lay queries and $KW\chi^2(2) = 247.96$, $p < 2.2\times10^{-16}$ in medico-scientific queries). In a pairwise comparison (Table 10.4), we found that comprehension in users with inadequate health literacy is lower than comprehension in users with elementary or good health literacy. Moreover, and contrary to our expectations, we found that the comprehension of elementary health literate users is higher than comprehension in users with good literacy.

Table 10.4: Significant differences between medians of comprehension between levels of health literacy (I-Inadequate, E-Elementary, G-Good), by query type.

	Lay queries	**Medico-scientific queries**
$Comp_{hl=I} < Comp_{hl=E}$	W = 505318 $p < 2.2 \times 10^{-16} < 0.01/3$	W = 566613 $p < 2.2 \times 10^{-16} < 0.01/3$
$Comp_{hl=I} < Comp_{hl=G}$	W = 791604 $p < 2.2 \times 10^{-16} < 0.01/3$	W = 845470.5 $p < 2.2 \times 10^{-16} < 0.01/3$
$Comp_{hl=E} > Comp_{hl=G}$	W = 1803308 $p = 6.37 \times 10^{-9} < 0.01/3$	W = 1715570 $p = 13.06 \times 10^{-5} < 0.01/3$

In Figure 10.6 we present the proportion of documents by level of topic familiarity, comprehension level and query type. As can be seen, the comprehension of documents by users with different topic familiarities changes with the type of query. In line with the previous results, the comprehension of the documents is always higher in sessions with lay queries. As can be seen in Table 10.5 these differences are statistically significant.

Table 10.5: Significant differences between the median of comprehension in both types of queries, by topic familiarity level.

	$Comp_{Lay} > Comp_{MS}$	
	W	**p-value**
Unfamiliar	W = 3038410	$p = 1.58 \times 10^{-6}$
Somehow familiar	W = 1785294	$p < 2.2 \times 10^{-16}$
Familiar	W = 279186.5	$p < 9.08 \times 10^{-16}$

In lay queries, it happens what we expected, i.e., the comprehension of documents tends to increase with topic familiarity. In terms of significant differences, as can be seen in Table 10.6, we found that unfamiliar users, when compared with other users, understand worse the documents retrieved with this type of queries.

In medico-scientific queries, we were surprised to find that users "somehow familiar" with the topic find documents harder to understand than users

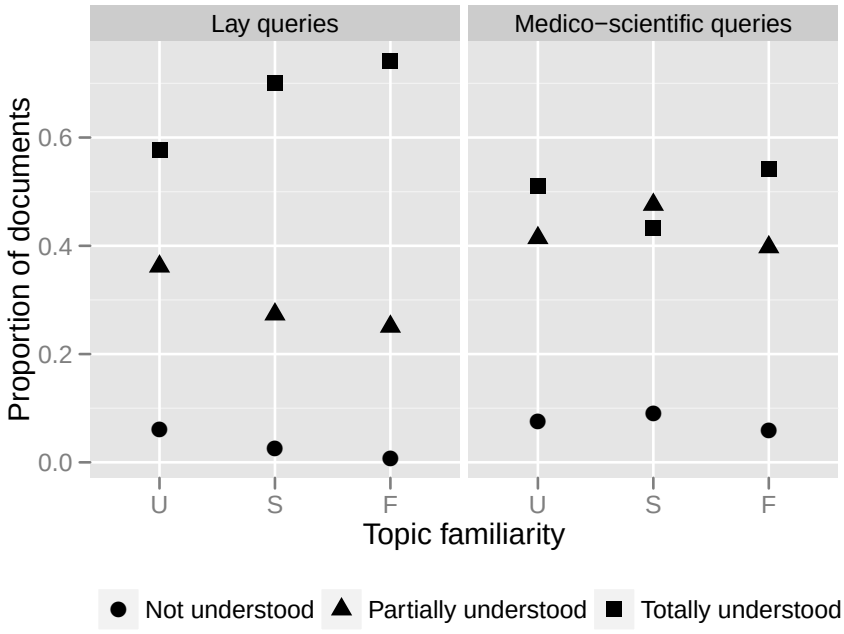

Figure 10.6: Proportion of documents by topic familiarity (U-Unfamiliar, S-Somehow familiar, F-Familiar), query type and comprehension level.

Table 10.6: Statistical differences between medians of comprehension between levels of topic familiarity (U-Unfamiliar, S-Somehow familiar, F-Familiar), by query type.

	Lay queries		**Medico-scientific queries**
$Comp_{tf=U} <$ $Comp_{tf=S}$	W = 1796246 $p < 2.2 \times 10^{-16} < 0.01/3$	$Comp_{tf=U} >$ $Comp_{tf=S}$	W = 2068034 $p < 1.08 \times 10^{-6} < 0.01/3$
$Comp_{tf=U} <$ $Comp_{tf=F}$	W = 628375 $p = 8.99 \times 10^{-16} < 0.01/3$	$Comp_{tf=U} <$ $Comp_{tf=F}$	W = 827630.5 $p = 0.04$
$Comp_{tf=S} <$ $Comp_{tf=F}$	W = 507711.5 $p = 0.02$	$Comp_{tf=S} <$ $Comp_{tf=F}$	W = 5345930 $p < 1.76 \times 10^{-7} < 0.01/3$

"unfamiliar" with the topic. Also, but now as expected, users "somehow familiar" with the topic understand documents worse than familiar users. As seen in Table 10.6 both differences are statistically significant. In medico-scientific queries, we found no significant differences between unfamiliar and familiar users.

Discussion

We found that users comprehend better documents retrieved with lay queries than documents retrieved with medico-scientific queries. This happens in general and at every level of health literacy and topic familiarity. In terms of health literacy we also found that users with inadequate health literacy understand documents worse than users with higher health literacy using both types of queries. This is in agreement with the definition of health literacy.

Birru et al. (2004), who studied information literacy instead of health literacy, reached a more drastic but similar conclusion, concluding that low literacy users were unable to interpret the retrieved information. Surprisingly, we also found that users with good health literacy understand documents worse than users with elementary health literacy. In terms of topic familiarity, we found that, when compared with other users, users unfamiliar with the topic understand worse the documents retrieved with lay queries. With medico-scientific queries, we found that users "somehow familiar" with the topic understand documents worse than users familiar with the topic but also worse than users unfamiliar with it. The latter result let us conclude that topic familiarity is not a necessary condition to comprehend a medico-scientific document. Characteristics like health literacy or knowledge about medico-scientific terminology may be more preponderant.

10.5 MEDICAL ACCURACY ANALYSIS

As previously explained, users had to provide an answer to the information need that led to the task. A medical doctor later evaluated the answers in terms of correct and incorrect contents. These two assessments were then combined into what we call medical accuracy.

In Figures 10.7, 10.8 and 10.9 we present the distributions of the medical accuracy, correct contents and incorrect contents of the answer in each type of query. In these figures we can see that query terminology does not strongly affect these variables. We observe that the proportion of answers in each level of classification is similar and no significant differences were found in the median of each variable between types of queries.

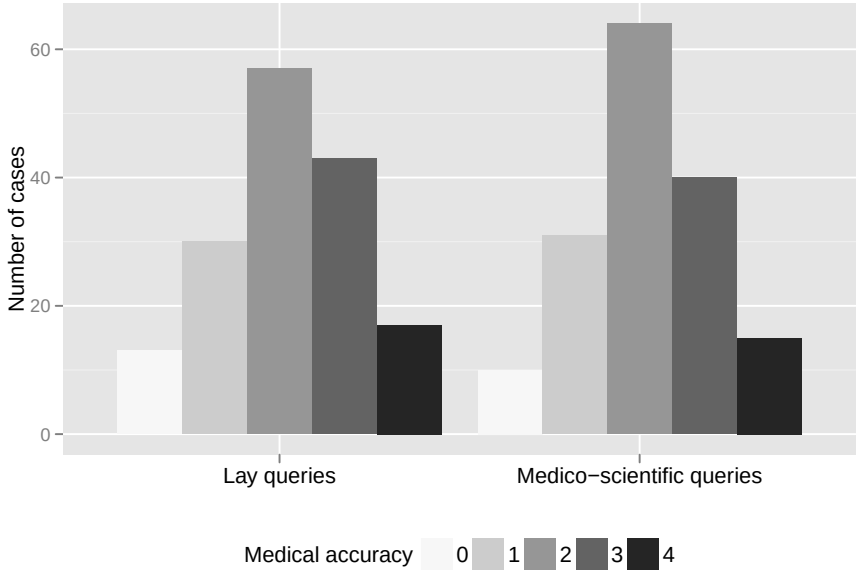

Figure 10.7: Answer's medical accuracy by query type.

As can be seen in Figure 10.10, despite a slight improvement of medical accuracy with the level of health literacy in both types of queries, the median

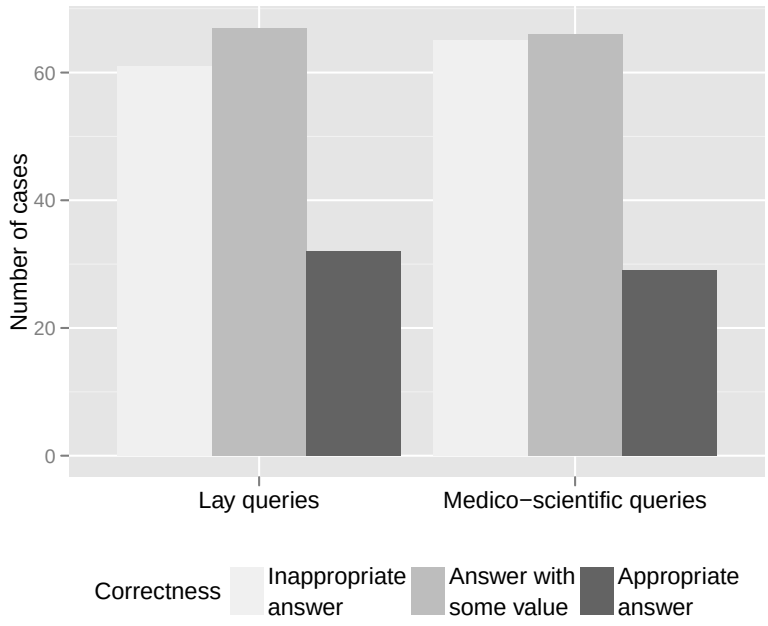

Figure 10.8: Answer's correctness by query type.

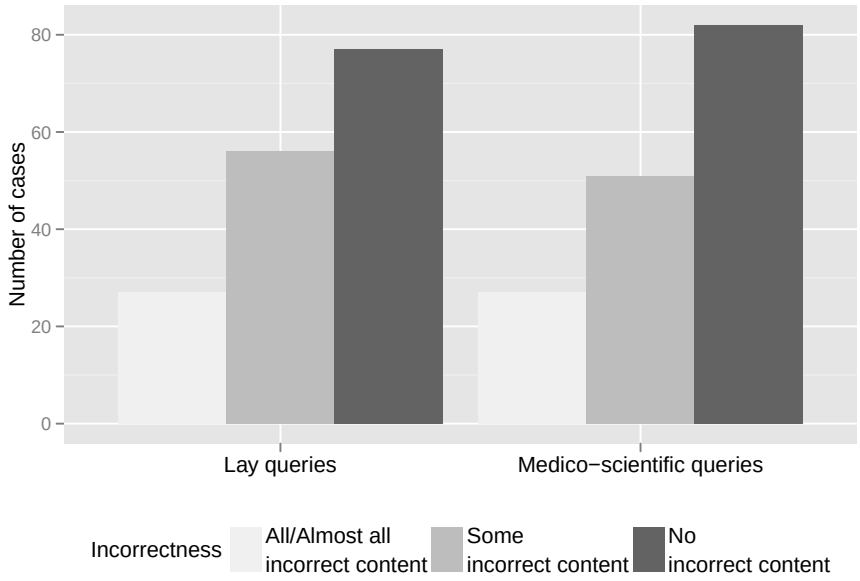

Figure 10.9: Answer's incorrectness by query type.

of this variable is always 2. However, just as with answer's correctness and incorrectness, differences are not significant. We also did not find significant differences between levels of health literacy in each type of query.

The median of answer's correctness is always 1 ("answer with some value") except in users familiar with the topic using medico-scientific queries, in which case it is 0 ("inappropriate answer"). In users familiar with the topic, the median of answer's correctness is significantly lower in medico-scientific queries

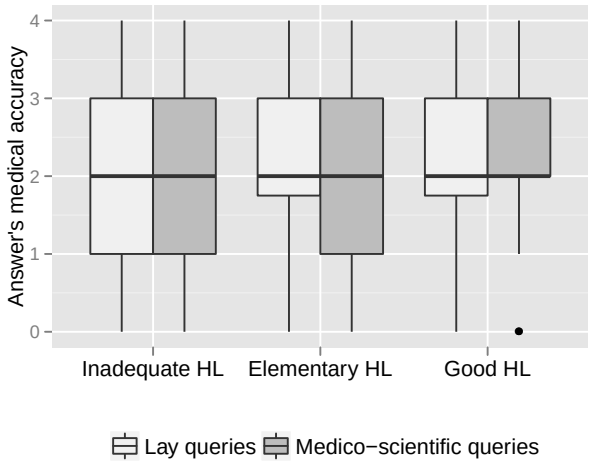

Figure 10.10: Medical accuracy by health literacy and query type.

(W=337.5, p = 0.03) when compared with lay queries. In users with different familiarity levels, no significant differences were found. In medico-scientific queries we also found differences in answers' correctness between levels of topic familiarity ($KW\chi^2(2) = 11.72, p = 0.003$). Further analysis led us to the conclusion that, with this type of queries, users unfamiliar with the topic give answers more accurate than those familiar with the topic ($W = 1540$, $p = 7.45 \times 10^{-12} < 0.01/3$). In lay queries, no significant differences were found. These results surprised us because we expected users familiar with the topic to become better prepared with medico-scientific queries and also to give better answers than other users, independently of the query type.

In terms of incorrect contents, the tendency is symmetric to the one described above, i.e., with medico-scientific queries there is a slight tendency to have answers with less incorrect content as the familiarity with the topic increases. In non-familiar users the median is 1, in those who are somehow familiar it is 1.5 and in familiar users it is 2. In spite of this tendency, we found no significant differences between query types in each level of topic familiarity. In both query types we also did not find significant differences between levels of topic familiarity.

The distributions of medical accuracy by topic familiarity and query type can be seen in Figure 10.11. The median of this variable is always 2 and, against our expectations, medico-scientific queries seem to result in more accurate answers in users who are not familiar with the topic. We found no significant differences either between query types in each level of familiarity, or between levels of familiarity in each type of query.

Discussion

We found that the type of query does not affect answer's correctness, incorrectness and global accuracy, neither in the general user nor in users with specific levels of health literacy. In terms of topic familiarity and with respect to answers' correctness, we detected that familiar users give answers with less correct content with medico-scientific queries than with lay queries.

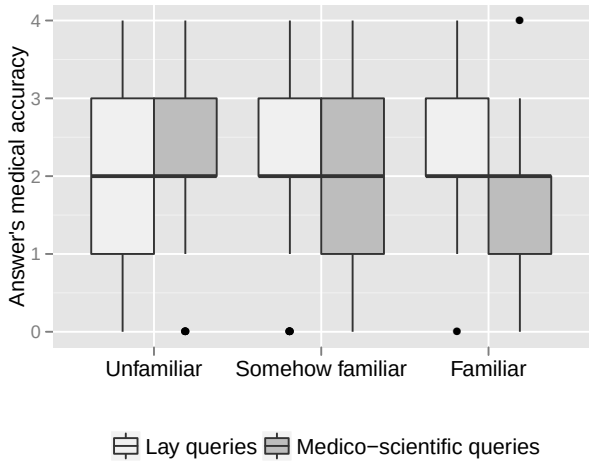

Figure 10.11: Medical accuracy by topic familiarity and query type.

Moreover, these users give answers with less correct content than non-familiar users with medico-scientific queries. Concerning answers' incorrectness, in medico-scientific queries, we found a tendency to have answers with less incorrect content as the familiarity with the topic increases, yet a non-significant difference. In medical accuracy, that combines the above measures, no significant differences were found. These results probably mean that familiar users were more restrained, less verbose when giving their answers what leads to answers with simultaneously less correct and less incorrect contents.

10.6 MOTIVATIONAL RELEVANCE ANALYSIS

Motivational relevance was evaluated through users' assessment of the task completion status in a scale of 1 (completely unsatisfied) to 5 (completely satisfied).

Since the median of the task completion status is 4 with both types of queries, globally we can say that users were satisfied with the search sessions. The distributions in both types of queries are very similar denoting that the type of query does not interfere with users' feeling of success.

An analysis of the motivational relevance by health literacy (Figure 10.12) reveals that users with inadequate health literacy feel less satisfied than elementary or good health literacy users. We found significant differences between health literacy levels in both types of queries (lay – $\chi^2(2) = 8.18, p = 0.02$; medico-scientific – $\chi^2(2) = 6.26, p = 0.04$). At $\alpha = 0.05$, we found that, with lay queries, users with inadequate health literacy feel less satisfied than users with elementary health literacy (W= 647.5, p=0.005<0.05/3) and good health literacy (W=961, p=0.0086<0.05/3). In medico-scientific queries, users with inadequate health literacy feel less satisfied than elementary literate users (W=680, p=0.01<0.05/3). As can be seen in Figure 10.12 there are no visible differences between types of queries in each level of health literacy. Through hypothesis tests we reached the same conclusion, i.e., there are no significant differences in the median of the task completion status between query types in each level of health literacy.

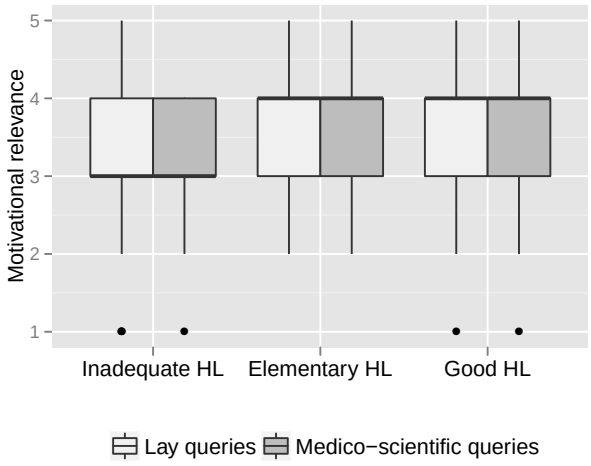

Figure 10.12: Motivational Relevance by health literacy level and query type.

Although the median of motivational relevance is always 4 (Figure 10.13), this variable slightly increases with topic familiarity, independently of the query type. However, this is only a tendency since we found no statistical significant differences between levels of topic familiarity in each type of query. Users familiar with the topic tend to be more satisfied with medico-scientific queries than with lay ones, but this difference is not significant.

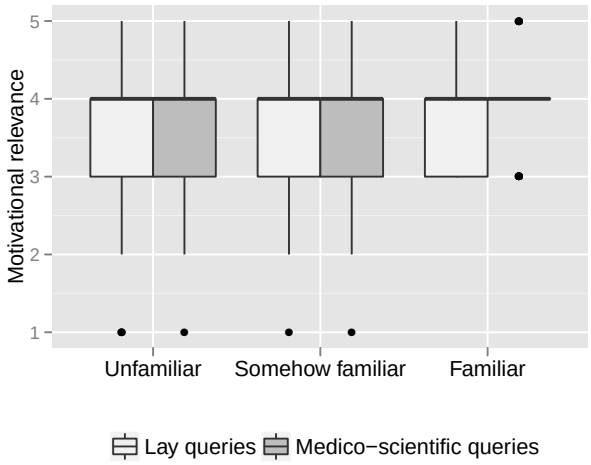

Figure 10.13: Motivational Relevance by topic familiarity and query type.

Discussion

In general, the type of query does not affect motivational relevance. The analysis by health literacy revealed that inadequate health literate users feel less satisfied than elementary and good health literate users with lay queries and less satisfied than elementary health literate users in medico-scientific sessions. Through these results we can see that low health literacy has a negative impact

on users' feeling of success in health search sessions. In terms of familiarity, users tend to be more satisfied with medico-scientific queries than with lay ones, but this difference is not significant.

10.7 FINAL DISCUSSION AND IMPLICATIONS

As demonstrated, medico-scientific queries demand more from users. Documents retrieved with these queries are mostly aimed at health professionals thus requiring users to be better prepared in health subjects. Moreover, we have shown these documents are less readable and are worse understood than documents retrieved with lay queries. When compared with lay queries, we found that medico-scientific queries have higher precision at the top-10 retrieved documents. Although non-significant, we also found medico-scientific queries surpass lay queries in gP5 and GAP. With GAP the same happens in all levels of health literacy (non significant differences). We were surprised to see the same trend in inadequate health literate users but, since these users have a mean GAP higher than other users on both types of queries, we suspect this happens because "less subject expertise seems to lead to more lenient and relatively higher relevance ratings" (Saracevic, 2007b). This means inadequate health literate users may give higher relevance scores than users with more health literacy to documents that are less helpful to them. Regarding medical accuracy, we found no significant differences but noticed that medico-scientific sessions slightly tend to generate knowledge with less incorrect contents and equal correct contents than lay sessions.

Comparing users with different levels of health literacy we found that users with inadequate health literacy understand documents worse and have less task success than users with higher health-literacy, with both types of queries. This corroborates our previous explanation that the former type of users assigns higher relevance scores to documents that are not helpful to them. Although not significant, we found that answers' medical accuracy tends to increase with user's health literacy. This is due to the presence of less incorrect knowledge in users with more health literacy, another trend we found. This is true on both types of queries but is stronger in medico-scientific sessions, showing that users with higher levels of health literacy are more apt to assimilate medico-scientific documents. These findings indicate that search engines should detect inadequate health literate users and return documents with contents adequate to them.

Concerning topic familiarity, we found that users not familiar with a topic, when compared with other users, understand worse the documents retrieved with lay queries. In medico-scientific queries we found that "somehow familiar" users understand documents worse than non-familiar users. This means that, in medico-scientific documents, health literacy may be more important to document's comprehension than topic familiarity. In terms of non-significant differences, we found that the quantity of incorrect contents in the knowledge that emerges from a medico-scientific session tends to decrease with topic familiarity. Moreover, users who are familiar with the topic tend to have higher motivational relevance and a higher GAP with medico-scientific queries when compared to lay queries.

Our findings suggest that a personalized query suggestion system would

improve the IR experience in the health domain. Toms and Latter (2007) reached the same conclusion while examining consumers searching for health information. These authors concluded that systems that provide assistance to query development are more helpful than specialized medical search engines. They infer that the key to successful queries, one of the major challenges in this type of search, is in the underlying infrastructure that supports the search process, which should be responsive to both consumers and experts. However, we argue that personalization should not be bipolar and distinguish only health consumers from health professionals. The personalization of the query suggestion system should be made by level of health literacy and level of familiarity with the health topic, which change with the topic and the health consumer. According to our results, users who have inadequate health literacy or are unfamiliar with the topic should be provided with recommendations of lay queries. On the other hand, users with higher health literacy or topic familiarity should be given alternative queries with medico-scientific terminology.

A previous study suggests that non-expert domain expertise is dynamic and may be developing over time (White et al., 2009). Our approach, when compared to the bipolar personalization strategy mentioned previously, does not have the drawback of hindering learning over time for health consumers. In fact, in users that are not unfamiliar with a topic, the system, through the queries it suggests and the documents it might give access to, supports and encourages people to learn more about the topic. Yet, in users unfamiliar with the topic, the suggestion of lay queries reinforces behavior. To address this gap, we suggest that either the query suggestion system, or the system that predicts the familiarity with the topic, take into account the number of previous searches on the topic. This information might help assess if the user is prepared to receive medico-scientific queries or even if he can raise one level in the scale of topic familiarity. This should, however, be carefully studied as further work. Moreover, this approach can only be effective if the system has access to all previous searches the user has made on the topic.

10.8 CONCLUSIONS

We have conducted a user study to analyze how changes in query terminology affect the health retrieval experience of users with different levels of health literacy and topic familiarity. We studied several aspects related to the information retrieval experience, namely documents' readability, documents' comprehension, sessions' precision, sessions' medical accuracy and motivational relevance.

Many results suggest that a personalized query suggestion system would improve the information retrieval experience in the health domain. Depending on the user, namely on his health literacy and topic familiarity, the system should provide medico-scientific or lay alternative suggestions to the query inserted by the user. This would not only give access to new types of documents but would also foster the learning of terminology that can be used in future queries. Our results suggest that users with inadequate health literacy and users who are unfamiliar with the topic should be provided with recommendations of lay queries. On the other hand, users with higher health literacy or higher topic familiarity should be given alternative queries with medico-

scientific terminology.

We have also concluded that search engines should detect users with inadequate health literacy and return documents with contents adequate to them, either with pictorial contents or with higher levels of readability. Moreover, since readability is important to all health consumers using both types of queries, it should be incorporated in search engines' ranking algorithms. In fact, we found that the relevance of a document highly depends on its comprehension. Health websites who want to provide information to consumers should also be aware that, if they need to use medico-scientific terminology, they should, at least, simplify the remaining contents.

In addition to the analysis made in this study, in the next chapter we explore how some context features affect one another. More specifically, we analyze how readability, comprehension, precision, medical accuracy and motivational relevance are related when different terminologies are used in the query.

INTERPLAY OF CONTEXT FEATURES CONSIDERING THE TERMINOLOGY OF THE QUERY

11.1 INTRODUCTION

In Chapter 10 we analyzed how changes in query's terminology affect the experience of users with different health literacy and topic familiarity. To complement that analysis, in this chapter we study the interplay between aspects related to the information retrieval experience, namely, documents' readability, documents' comprehension, sessions' precision, sessions' medical accuracy and tasks' motivational relevance, considering the terminology of the query.

This study is based on the user experiment described in Chapter 8 and the statistical assumptions mentioned in Section 10.4. Three research questions drove this study:

1. How does documents' readability affects the comprehension of the documents and the precision, medical accuracy and motivational relevance of the session, with lay and medico-scientific queries?

2. How does documents' comprehension affect the precision, medical accuracy and motivational relevance of the session, with lay and medico-scientific queries?

3. How are precision, medical accuracy and motivational relevance related, with lay and medico-scientific queries?

The presentation of the results is organized by research question, that is, three major sections will follow, one for the readability impact, other for the comprehension impact and the third for the relation between precision, medical accuracy and motivational relevance. After these sections, we discuss the results and summarize our conclusions.

11.2 READABILITY IMPACT

11.2.1 On Comprehension

As can be seen in Figure 11.1, in medico-scientific queries the comprehension increases with documents' readability, i.e., as SMOG gets lower. In fact, in these queries, we found statistically significant differences in readability between all levels of comprehension (Table 11.1). In lay queries we found the

readability of the documents does not affect the comprehension so much. Surprisingly, as seen in Table 11.1, we even found that documents "totally understood" are significantly harder to read than documents "partially understood".

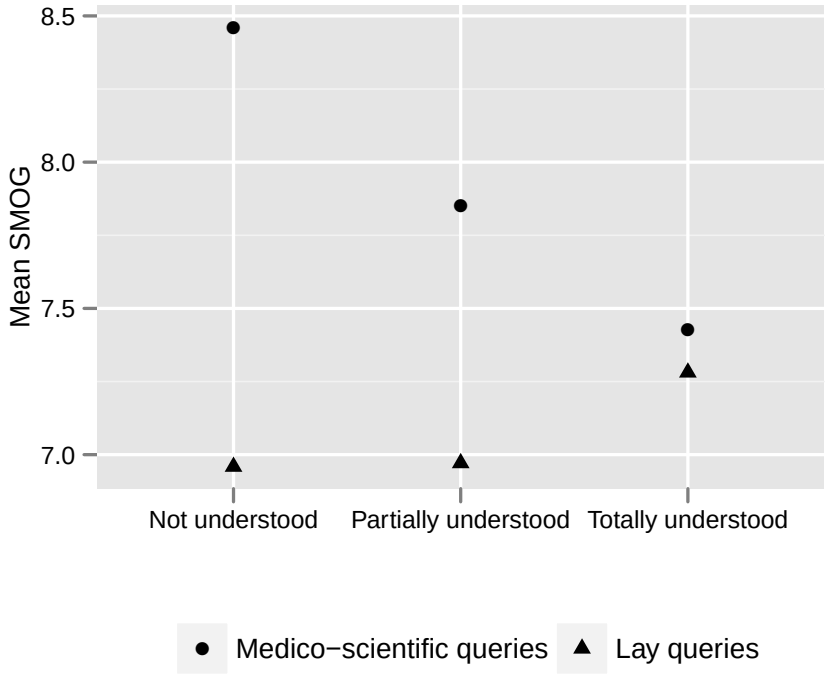

Figure 11.1: Mean SMOG by comprehension level and type of query.

Table 11.1: Significant differences in the mean SMOG between comprehension levels. $SMOG_n$ is the SMOG at comprehension level n. ** signs a p<0.01/3.

	Lay queries	**Medico-scientific queries**
$SMOG_0 > SMOG_1$	-	W=353147.5, p = 8.71e-08**
$SMOG_0 > SMOG_2$	-	W = 449795, p < 2.2e-16**
$SMOG_1 < SMOG_2$	W=1890011, p =1.1e-04**	-
$SMOG_1 > SMOG_2$	-	W=2177938, p = 3.08e-08**

11.2.2 On Precision

To analyze how the readability of the documents affects precision we do not use graded precision but compare the readability score of the documents with their relevance assessments. The mean SMOG per relevance level and type of query is plotted in Figure 11.2. As can be seen, in both types of queries, documents classified as "not relevant" are harder to read than the other documents. These are the only significant differences found in the mean SMOG between levels of relevance in each type of query. The tests' values and the associated p-values are presented in Table 11.2. Since we are performing multiple compar-

isons, we applied the Bonferroni correction, dividing α by the number of tests performed in each type of query, i.e., 3 (relevance level 0 with relevance levels 1 and 2 and relevance level 1 with relevance level 2). These results show that a low readability (that corresponds to a higher SMOG) can be considered a serious obstacle for a document's relevance, disregarding the type of terminology.

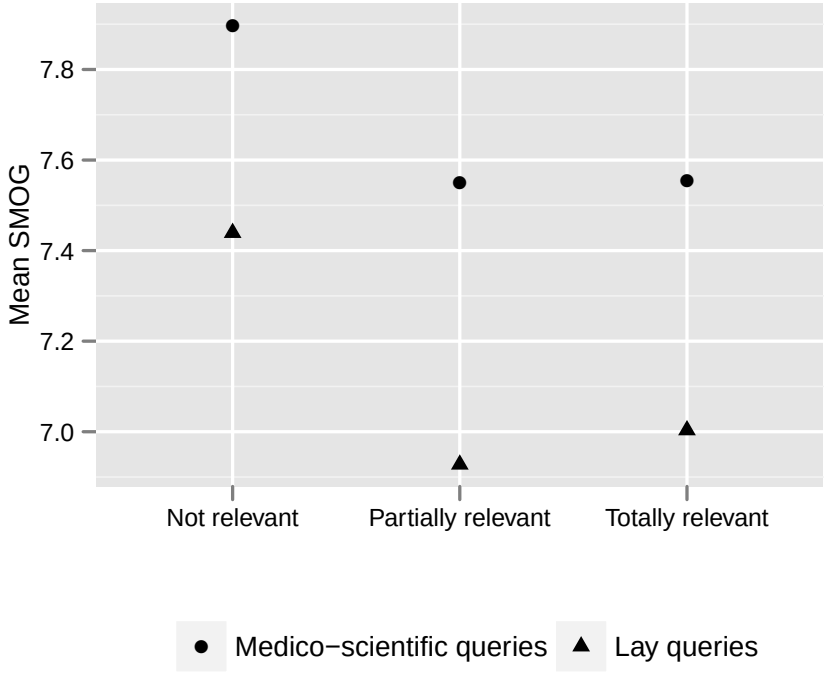

Figure 11.2: Mean SMOG per relevance level and type of query.

Table 11.2: Significant differences in the mean SMOG (SG) between levels of relevance. ** signs a p<0.01/3.

	Lay queries	**Medico-scientific queries**
$SG_0 > SG_1$	W = 1492732, p = 3.967e-15**	W = 1206654, p = 4.522e-08**
$SG_0 > SG_2$	W = 1300120, p = 3.489e-08**	W = 1266217, p = 5.693e-09**

11.2.3 On Medical accuracy

As shown in Figure 11.3, although with some exceptions, answers with higher medical accuracy tend to be associated with documents that are harder to read (higher SMOG) in both types of queries. The highest level of medical accuracy (4) is the big exception where the mean SMOG is lower (mean=7.39, sd=0.09) than the SMOG for answers with a medical accuracy of 3 (mean=7.69, sd=0.06). We found significant statistical differences in the mean SMOG between levels of medical accuracy in each query type (for lay queries, F(4)= 4.264, p=0.0019; for medico-scientific queries, KW χ^2 = 49.65 , p=4.27e-10).

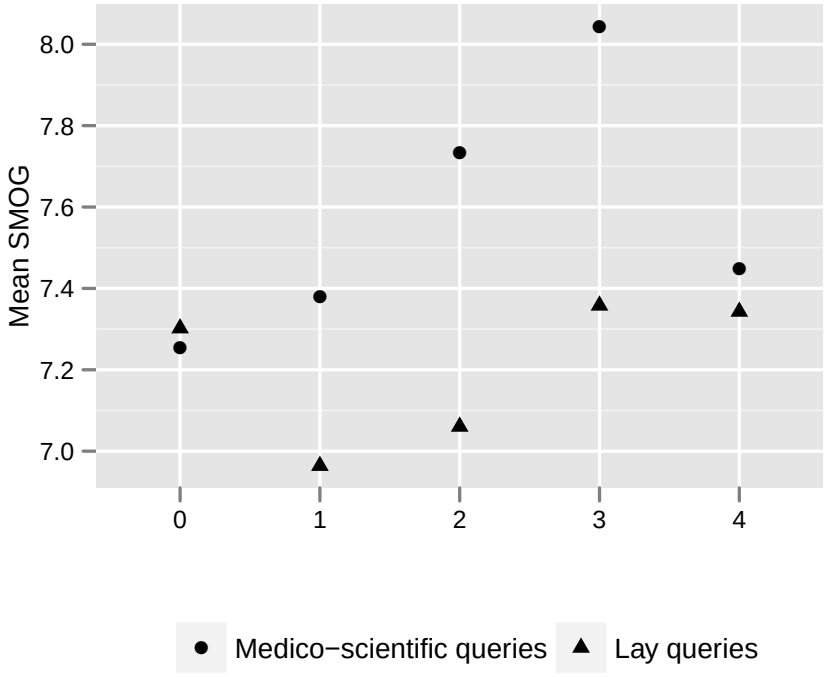

Figure 11.3: Mean SMOG by answer's medical accuracy and type of query.

In Table 11.3 we present the significant results of the pairwise comparisons in each type of query. In lay queries we used Tukey's HSD test and in medico-scientific queries we used several Wilcoxon tests with the Bonferroni correction. We found that, with both types of queries, documents associated with sessions with a medical accuracy of 3 are more complex (higher SMOG mean) than documents of sessions with a medical accuracy of 1 and 2. With medico-scientific queries, they are also more complex than documents of sessions with a medical accuracy of 0. Moreover, sessions with a medical accuracy of 2 have documents less readable than documents pertaining to sessions with a medical accuracy of 0 and 1.

Table 11.3: Mean SMOG significant differences between levels of medical accuracy by types of query. SG_n is the SMOG at medical accuracy of n. * signs a p<0.05/3 and ** signs a p<0.01/3.

	$SG_0 < SG_2$	$SG_0 < SG_3$	$SG_1 < SG_2$	$SG_1 < SG_3$	$SG_2 < SG_3$
Lay	-	-	-	(0.071; 0.717)	(0.023; 0.572)
	-	-	-	p=0.0078	p=0.026
MS	W = 213191.5	W = 123912	W=656522	W=382476.5	W=873451.5
	p = 0.0021*	p = 4.964e-06**	p = 1.28e-05**	p = 2.282e-10**	p = 0.0011*

11.2.4 On Motivational Relevance

In Figure 11.4 we plotted the mean SMOG by motivational relevance and query type. As can be seen, in medico-scientific queries the sessions with lowest

satisfaction rates (1 and 2) are associated with less readable documents. In lay queries this only happens with the first level of motivational relevance. In sessions where users feel "completely satisfied", documents' readability is lower than the one in previous levels of satisfaction. In medico-scientific queries we detected no significant differences in the mean SMOG between levels of satisfaction with the task but we found them in lay queries (F(4)=5.91, p=9.7e-05).

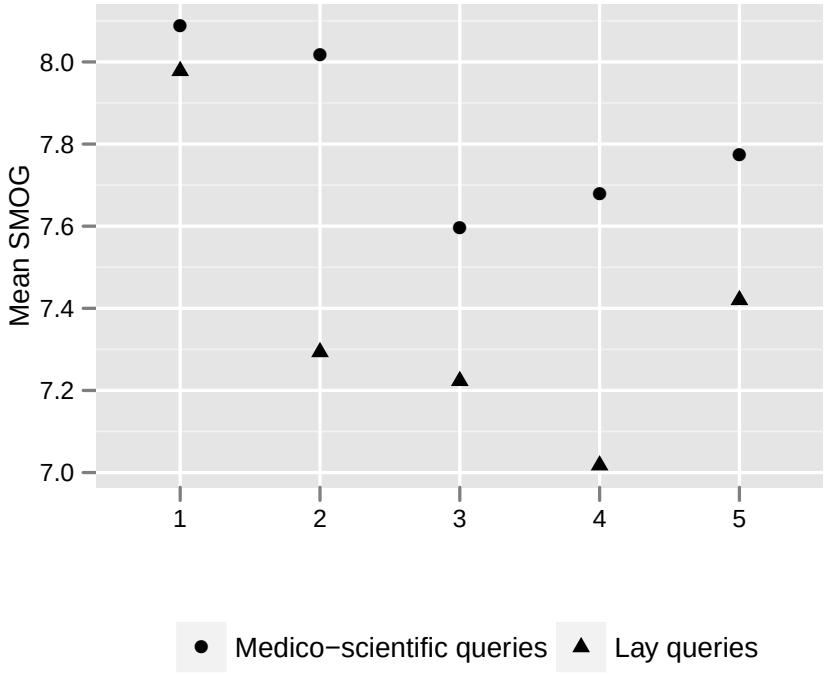

Figure 11.4: Mean SMOG by motivational relevance and type of query.

As can be seen in Table 11.4, we found that documents associated with sessions where users feel "completely unsatisfied" are less readable than the ones of sessions associated with levels 3 and 4 of the motivational relevance. Moreover, this also happens with documents of sessions where the users feel "completely satisfied" when compared with documents of sessions where users feel "satisfied" (4).

Table 11.4: Significant differences in the SMOG metric between levels of motivational relevance in lay queries. $SMOG_n$ - SMOG at motivational relevance n.

$SMOG_1 > SMOG_3$	$SMOG_1 > SMOG_4$	$SMOG_5 > SMOG_4$
(-1.459; -0.051)	(-1.657; -0.264)	(0.055; 0.751)
p=0.029	p=0.001	p=0.014

11.3.1 *On Precision*

We can see, in Figure 11.5, that relevance increases as comprehension increases in both types of queries. With lay queries we found that "totally relevant" documents have a comprehension median higher than the other documents (Table 11.5). With medico-scientific queries, as the relevance of the documents increases, so does their comprehension. As can be seen in Table 11.5 all these differences are statistically significant. In line with the previous results, comprehension is higher in lay queries in every level of relevance: 0 (W = 1991875, p<2.2e-16), 1 (W = 987763, p<2.2e-16) and 2 (W = 908917, p = 5.161e-14).

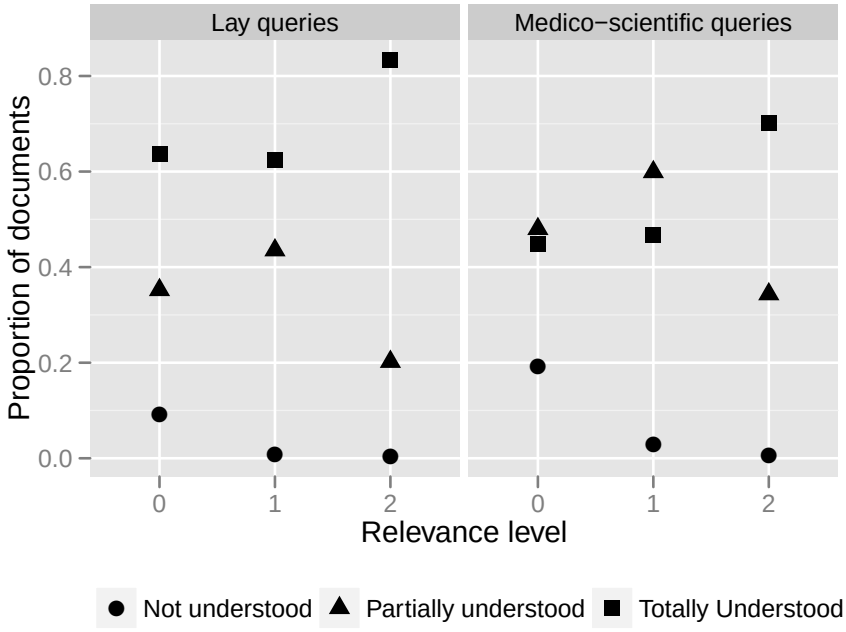

Figure 11.5: Proportion of documents by comprehension, relevance level and query type.

Table 11.5: Significant differences in the comprehension median between relevance levels. $Comp_n$ – Comprehension at relevance level n. ** signs a p<0.01/3.

	Lay queries	**Medico-scientific queries**
$Comp_0 < Comp_1$	W=1260380, p = 0.1219	W = 965872, p = 1.023e-08**
$Comp_0 < Comp_2$	W=906232.5, p < 2.2e-16**	W=765754.5, p < 2.2e-16**
$Comp_1 < Comp_2$	W=614633.5, p < 2.2e-16**	W=654833.5, p < 2.2e-16**

11.3.2 *On Medical accuracy*

Relating comprehension with medical accuracy we observe different behaviors in lay and medico-scientific queries. In lay queries the tendency shows

that, as the accuracy of the answer improves, the global comprehension of the documents of that session also increases. As can be seen in Figure 11.6, in this type of queries, higher levels of medical accuracy have more "totally understood" and less "partially understood" documents. Using the Kruskal-Wallis test we detected significant differences on the median of comprehension between levels of medical accuracy (KW $\chi^2(4) = 41.59$, p=2.026e-08).

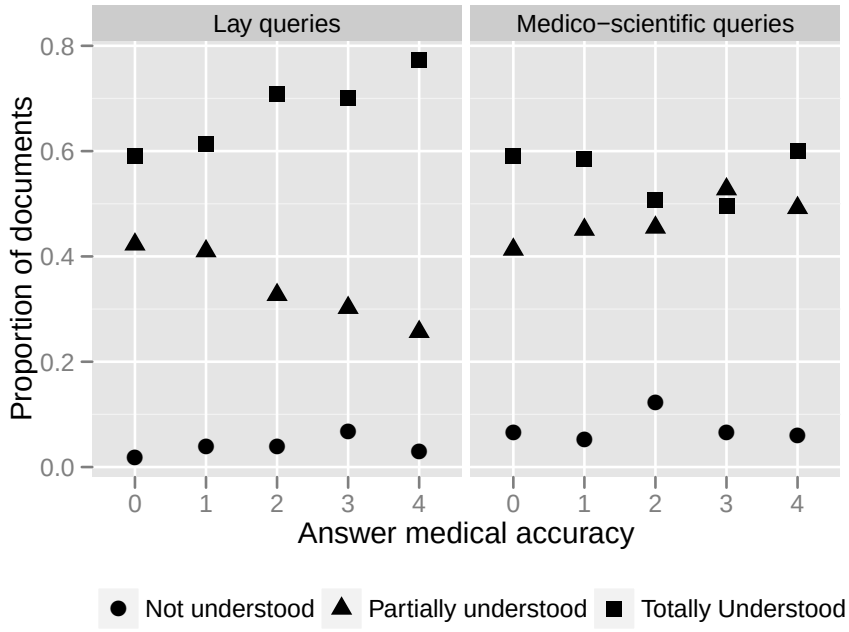

Figure 11.6: Proportion of documents by comprehension, level of medical accuracy and query type.

With further analysis (Table 11.6) we concluded that the comprehension in the lowest levels of medical accuracy (0 and 1) is lower than the comprehension in the levels 2 and 4 of medical accuracy. This means that, when medical terminology is not an issue, comprehension may be a decisive factor to the accuracy of the knowledge obtained from a session.

Table 11.6: Significant differences in the comprehension median between medical accuracy levels in lay queries. ** signs a p<0.01/3.

$Comp_0 < Comp_2$	$Comp_0 < Comp_4$	$Comp_1 < Comp_2$	$Comp_1 < Comp_4$
W = 269071.5	W = 75189.5	W = 610154.5	W = 170947.5
p = 0.0002384**	p = 2.511e-07**	p = 3.858e-05**	p = 3.684e-08**

In medico-scientific queries the reality is different and we found that some of the lower levels of medical accuracy, namely 0 and 1, are associated with documents that are comprehended better than the ones of higher levels of medical accuracy, namely 2 and 3. These differences are statistically significant as can be seen in Table 11.7. This makes us suspect that, when medico-scientific terminology is used in documents, comprehension is not so preponderant to the

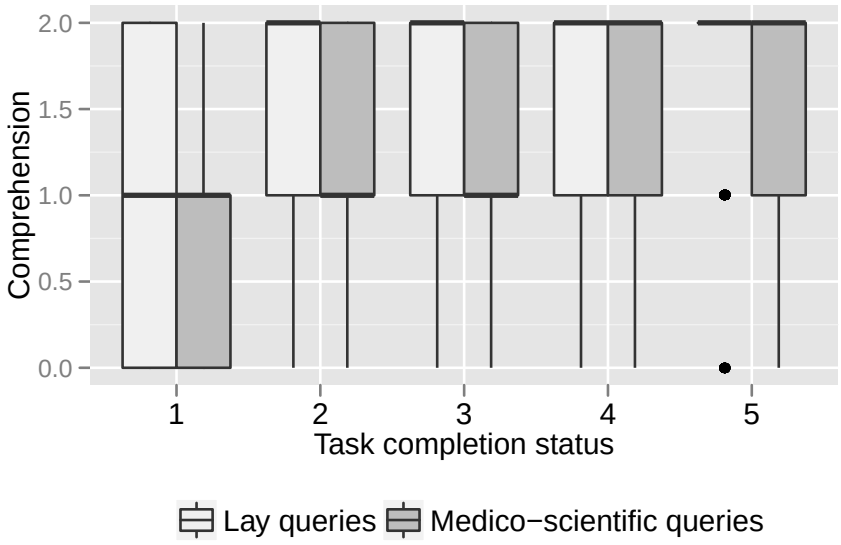

Figure 11.7: Distributions of comprehension by motivational relevance and query type.

answers' medical accuracy. Users may understand documents a little worse but still be able to assimilate the main message.

Table 11.7: Significant differences in the comprehension median between medical accuracy levels in medico-scientific queries. Compn – Comprehension at medical accuracy n. ** signs a p<0.01/3.

$Comp_0 > Comp_2$	$Comp_0 > Comp_3$	$Comp_1 > Comp_2$	$Comp_1 > Comp_3$
W = 267363	W = 166383	W = 794181.5	W = 493273
p = 0.00019**	p = 0.0006**	p = 3.452e-05**	p = 0.00053**

11.3.3 On Motivational Relevance

As seen in Figure 11.7, the comprehension tends to increase with the feeling of success in the search task. In both types of queries we detected significant differences in the median of comprehension between levels of motivational relevance (lay - KW $\chi^2(4)$ = 121.47, p<2.2e-16; medico-scientific - KW $\chi^2(4)$ = 164.95, p<2.2e-16). A set of Wilcoxon tests with the Bonferroni correction allowed us to conclude that, in medico-scientific queries, the comprehension significantly grows as we move from lower levels to higher levels in the motivational relevance scale. In lay queries, this also happens with the exception of the comparison of the 2nd level with the 3rd and the 4th levels. As we can be see in Table 11.8, all the differences are significant at α=0.01, since all the p-values are inferior to 0.01/10.

Table 11.8: Statistically significant differences in the median of comprehension between levels of motivational relevance.

	Lay queries	Medico-scientific queries
$Comp_1 < Comp_2$	W = 11676.5, p=7.391e-07	W = 5119.5, p=2.106e-06
$Comp_1 < Comp_3$	W = 66599.5, p=2.279e-09	W = 26407.5, p=1.526e-10
$Comp_1 < Comp_4$	W = 89145.5, p=1.23e-14	W = 32663.5, p=3.76e-14
$Comp_1 < Comp_5$	W= 20048, p<2.2e-16	W = 6652.5, p<2.2e-16
$Comp_2 < Comp_3$	-	W = 187861, p=0.0007877
$Comp_2 < Comp_4$	-	W = 239695.5, p=1.406e-09
$Comp_2 < Comp_5$	W= 63345.5, p=3.12e-07	W = 49552.5, p<2.2e-16
$Comp_3 < Comp_4$	W = 1600860, p=1.384e-07	W = 1622274, p=6.437e-08
$Comp_3 < Comp_5$	W = 356269.5, p=4.38e-15	W = 340579.5, p<2.2e-16
$Comp_4 < Comp_5$	W= 573607, p=6.649e-07	W = 535750.5, p=2.475e-09

11.4 Relation between Precision, Medical Accuracy and Motivational Relevance

11.4.1 Relation of Precision with Medical accuracy

Figure 11.8 shows the distributions of GAP by answer's medical accuracy and query type. We cannot detect a clear pattern of association between answer's medical accuracy and GAP. Within medico-scientific queries, we could not detect significant differences in the mean GAP between levels of medical accuracy. On the other hand, we detected them in sessions with lay queries (F(4)=2.767, p=0.029). Tukey's HSD test allowed us to conclude that answers with the lowest medical accuracy (0) are associated with sessions with lower mean GAP than sessions of answers with a medical accuracy of 1 ((0.025, 0.41), p=0.017) and 3 ((0.009, 0.372), p=0.033).

11.4.2 Relation of Precision with Motivational Relevance

Globally, and as expected, the median of GAP tends to increase with the degree of satisfaction with the session (Figure 11.9). In lay queries we found significant differences between levels of motivational relevance (F(4)=6.65, p=5.78e-05). In Table 11.9 we present the significant differences found, i.e., sessions where users feel "unsatisfied" (2) have a lower mean GAP than sessions with higher feeling of success. In medico-scientific queries we found no significant differences between levels of motivational relevance.

Table 11.9: Significant differences in GAP between levels of motivational relevance in lay queries. GAP_n - GAP at motivational relevance n.

$GAP_2 < GAP_3$	$GAP_2 < GAP_4$	$GAP_2 < GAP_5$
(0.0198; 0.4155)	(0.118; 0.505)	(0.1278; 0.5713)
p=0.023	p=0.00016	p=0.0002

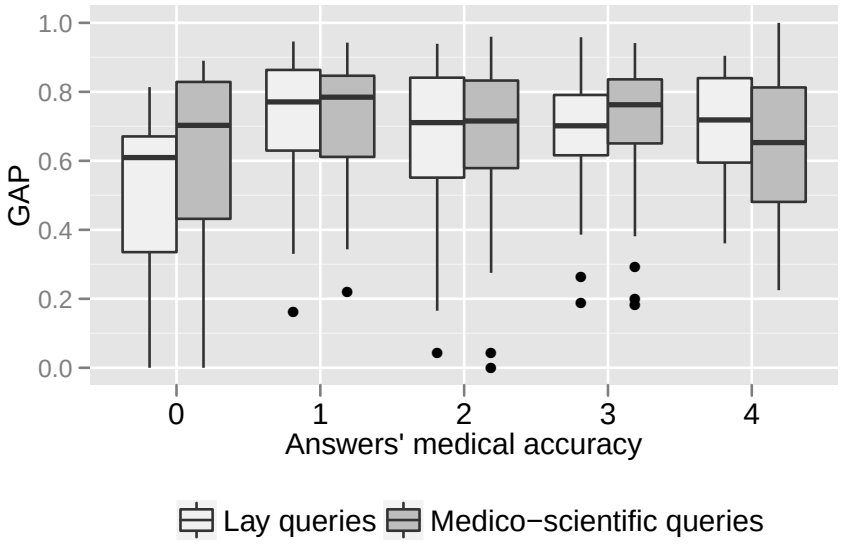

Figure 11.8: Distributions of GAP by answer's medical accuracy and query type.

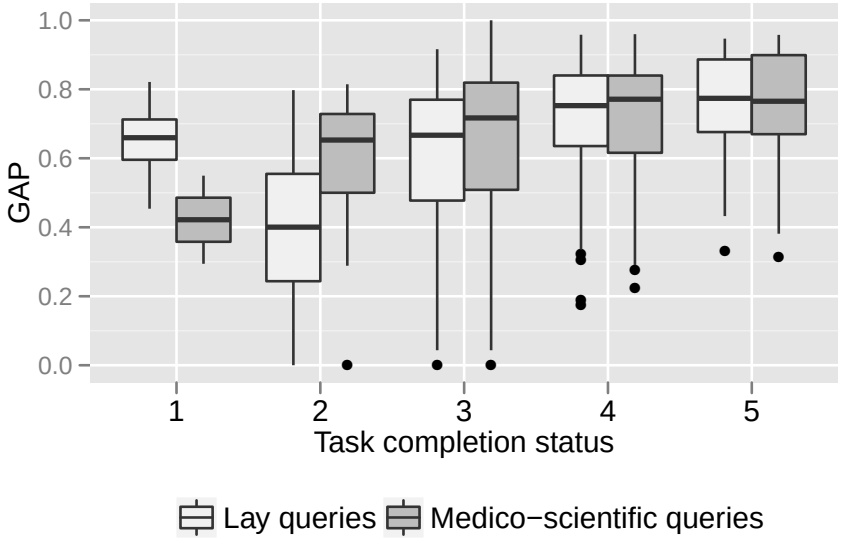

Figure 11.9: Distributions of GAP by answer's medical accuracy and query type.

11.4.3 Relation of Medical Accuracy with Motivational Relevance

The median of the medical accuracy is almost always 2, independently of the motivational relevance and query type. The only exception lays in the most successful search tasks with lay queries where the median of medical accuracy is 3 (Figure 11.10). We did not find any significant differences between the medians of medical accuracy between levels of motivational relevance in any of the types of query.

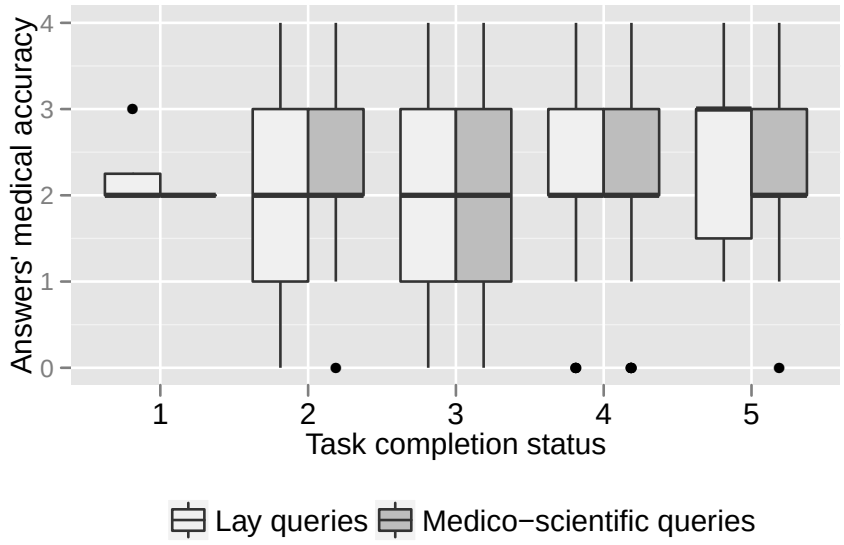

Figure 11.10: Distributions of medical accuracy by motivational relevance and query type.

11.5 DISCUSSION

An overview of the existing relations between analyzed dimensions is given in Table 11.10. The relation of readability with precision shows that low readability can be a serious obstacle to document's relevance, agreeing with Muresan et al. (2006) who concluded that a higher readability has positive effects on retrieval performance.

In terms of comprehension we concluded that readability is more important in documents with medico-scientific terminology. In documents retrieved by lay queries this is not so important since we even found that documents with lower readability were better understood than documents easier to read.

About medical accuracy, lay sessions one level below the accuracy scale's maximum (3), are associated with documents less readable than documents of sessions with medical accuracy of 1 and 2. In medico-scientific queries, this relation is more systematic. These results suggest that documents with more accurate medical contents are significantly harder to read which creates an obstacle to low literacy users.

Regarding the motivational relevance, with lay queries, we found that sessions where users feel "completely unsatisfied" have less readable documents than sessions where they are "neutral" or "satisfied". Although documents' lack of readability may be an obstacle to, at least, some success in the task, this is not true in the highest level of motivational relevance. In fact, we found that, with the same type of queries, sessions where users feel "completely satisfied" have documents harder to read than sessions where they feel "satisfied".

Through the relation between precision and comprehension we can see the relevance of a document highly depends on its comprehension. The relationship of precision with medical accuracy shows that, in lay queries, users relevance assessments are related to the medical accuracy of their answers. The number of relevant documents or the degree of documents' relevance affects

Table 11.10: Statistical significant relationships between analyzed dimensions. L = lay; MS = medico-scientific.

	Precision	**Comprehension**	**Med. accuracy**	**Motivational relevance**
Readability	Not relevant documents are less readable than partially and totally relevant ones [L&MS].	Readability not so crucial to comprehension as in professional queries [L]. Comprehension increases with readability [MS].	Highest levels of med. accuracy associated with less readable documents [L]. As the medical accuracy of the session increases, the readability of their documents decreases [MS].	Lack of readability is an obstacle to the initial levels of motivational relevance but not to the highest levels [L].
Precision	-	Totally relevant documents are better understood than less relevant documents [L]. Relevance systematically increases with comprehension [MS].	Sessions with the lowest medical accuracy have lower GAP than sessions with a medical accuracy of 1 and 3 [L].	Sessions where users feel "unsatisfied" have lower precision than sessions with higher feeling of success [L].
Comprehension	-	-	Comprehension in the lowest levels of medical accuracy is lower than comprehension in higher levels of medical accuracy [L]. Lowest levels of medical accuracy are associated with higher comprehension than higher levels of medical accuracy [MS].	Comprehension grows as we move from lower levels to higher levels in the motivational relevance scale [L&MS].
Med. Accuracy	-	-	-	No relationship found [L&MS].

users' answers. Since we found no significant differences in medico-scientific queries, we suspect that, with this type of queries, either relevance assessments are less related with the accuracy of the documents' contents, or users assess documents by an estimate of their relevance to others. For example, this can happen if a user assesses a document as relevant when the content seems related to the topic but he cannot understand it or use it for his own benefit. Finally, in the last cell of the "precision" row of Table 11.10, we can see that in lay sessions, users' feeling of success is related with session's precision. The explanations given for the absence of relationship between precision and medical accuracy may also explain why no relation was found between precision and motivational relevance in medico-scientific queries.

The relation between comprehension and medical accuracy in lay queries indicate that, when medical terminology is not a question, comprehension may be an important factor to the accuracy of the knowledge obtained from a session. In medico-scientific queries, the opposite finding let us say that, when medico-scientific terminology is used in documents, comprehension is not so preponderant to the answers' medical accuracy. Either the users understand documents worse but are still able to assimilate at least part of the contents, or the higher accuracy of medico-scientific documents, when compared with lay documents, compensate users' lower comprehension of these documents. With respect to motivational relevance, comprehension is found to be an important and influent factor.

11.6 CONCLUSION

In Chapter 10 we described how changes in query's terminology affect the experience of users with different health literacy and topic familiarity. Following that analysis, in this chapter we study the interplay between aspects related to the information retrieval experience, namely, documents' readability, documents' comprehension, sessions' precision, sessions' medical accuracy and tasks' motivational relevance, considering the terminology of the query.

This analysis allows us to conclude that readability is essential for a document to be at least partially relevant; that it becomes even more important if the document has medico-scientific terminology and that it is crucial in the lower satisfaction levels but not in the higher ones. This information can be used not only by search engines but also by health websites. Search engines can explore this, incorporating readability in their ranking algorithm, not only in searches associated with lay queries but also in search performed by consumers using professional queries. Through past behaviors it should not be difficult to predict if the user is a professional or a consumer. Search engines should be aware that the relevance of a document highly depends on its comprehension and that unsuccessful tasks have lower precision than more successful tasks, two findings of this study. Health websites who want to provide information to consumers should also be aware that, if they need to use medico-scientific terminology, they should, at least, simplify the remaining contents. It is important to note that we suspect the more accurate documents are, the harder they are to read.

In lay queries we found the medical accuracy of users' answers is related to the session's relevance assessments. This shows that users can, at least in part,

relate their relevance assessments with the medical accuracy of the documents. On the other hand, in medico-scientific queries we suspect this relationship is weaker. In lay queries, comprehension is more crucial to the accuracy of the resulting knowledge than in medico-scientific queries. In the latter, either the user understands documents worse but is still able to assimilate at least part of the content or the higher accuracy of medico-scientific documents, when compared with lay documents, compensates users' lower comprehension of these documents.

The last study based on the experiment described in Chapter 8 is described in the following chapter. In that study we analyze how health literacy and topic familiarity affect query formulation. Moreover we study how retrieval sessions having queries with different terminologies affect the query reformulation behavior of users with different health literacy and topic familiarity.

QUERY BEHAVIOR: THE IMPACT OF HEALTH LITERACY, TOPIC FAMILIARITY AND TERMINOLOGY

12.1 INTRODUCTION

There are mismatches between the terminology used by health consumers and the one used in standard medical vocabularies and health documents (Zeng et al., 2002), and this may be an obstacle to successful health searches. The development of techniques to improve the communication between health professionals and consumers and the proposal of initiatives to help consumers understand health information are receiving a large attention nowadays. The first was recently discussed in a workshop of the 2013's Conference on Human Factors in Computing Systems entitled "Patient-Clinician Communication – The Roadmap for Human-Computer Interaction" and the second was discussed in a panel of the Association for Information Science and Technology 2010 annual meeting (Souden and Rubenstein, 2010).

Two user characteristics influence the amplitude of this terminology gap. One is the health literacy, that is, "the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions" (USA Department of Health and Human Services, 2000). The other is topic familiarity, i.e., user's general knowledge about the topic of a search task (e.g.: diabetes). Note that these two features are distinct. A health consumer with good health literacy is expected to be unfamiliar with several health topics.

We are convinced that higher levels of health literacy (HL) and topic familiarity (TF) give users the ability to formulate medico-scientific queries in addition to lay queries and therefore a higher probability of finding the necessary information. Moreover, we also think the above characteristics influence the query reformulation behavior after an initial iteration where technical documents, i.e., documents containing medico-scientific terminology, are accessed. The characterization of these behaviors may help search engines decide if and how search assistance mechanisms like query suggestion or ranking algorithms can be personalized.

12.2 RELATED WORK

In the following subsection we describe the main work regarding the influence of topic familiarity on query formulation behavior. The lack of studies consid-

ering health information literacy made us describe, in the other subsection, studies that explore users' information literacy on IR behavior.

12.2.1 The influence of topic familiarity on query formulation

Several works explore the influence of topic familiarity in IR. In the health domain, Wildemuth (2004) examines the search behavior of medical students that were observed on three different occasions: at the beginning of a course, at the end of the course, and six months after the course. The author concluded that individuals with less domain knowledge were less efficient in selecting concepts to include in search queries and performed worse in search modification. Moreover, although it improved performance in all occasions, system assistance during query formulation was considered more useful on users with less knowledge on the topic.

Two different studies explore the influence of topic familiarity on the use of a thesaurus for query expansion. Sihvonen and Vakkari (2004) conducted a study with 15 users having knowledge on the topic and 15 users without this knowledge, concluding that the use of the thesaurus was helpful for the experts but not for the novices. This conclusion contradicts Wildemuth (2004) conclusions. In the other study, Shiri (2005) analyzed how topic's familiarity affected users' behavior on thesaurus' use and concluded that "searches involving moderately and very familiar topics were associated with browsing around twice as many thesaurus terms as was the case for unfamiliar topic".

12.2.2 The influence of information literacy on IR behavior

Birru et al. (2004) observed low literacy adults search for health information. The search terms used to find health information were one of the analyzed items. Authors concluded that, without guidance, users had difficulty "generating original search terms that would yield specific results", which constitutes a barrier to getting specific and targeted web health information. Note that this study explores users' information literacy and not users' health literacy. As defined by the National Forum on Information Literacy, information literacy is "the ability to know when there is a need for information, to be able to identify, locate, evaluate, and effectively use that information for the issue or problem at hand" (NFIL, 2013).

Another work focused on information literacy is the one conducted by Kodagoda and Wong (2008) to understand how low literacy users search for information. In their study, authors compared the retrieval performance of high and low literacy users and concluded that low literacy users take more time to complete the search task, are less accurate, spend more time on each web page, are less informed by webpages, have less focused search strategies, have a greater tendency to re-visit web pages and more likely get lost than high literacy users. In agreement with Summers and Summers (2005), Kodagoda and Wong (2008) concluded that low literacy users often prematurely abandon their searches, judging they reached their goal. In domains like health, where inappropriately interpreted information may have impact on the life of the user or someone they care, this can be problematic. The consequences on users' life, the importance of successful health searches for an informed health consumer and the prevalence of health web searches distinguish the health do-

main from others, where query behavior may have been studied in the light of users' familiarity or literacy.

To learn how to make web health contents more usable and accessible for users with low health literacy, Summers and Summers (2005) compared the reading and navigational strategies of users with different health literacy skills. Among several conclusions, they found that users with low literacy often avoid search because it requires proper spelling and typing capabilities and because they have difficulties processing search results pages. Considering users' information literacy, Kodagoda et al. (2012) proposed Invisque (INteractive VIsual Search and Query Environment), a system that allows users to create queries and search for information in a visual manner. The system was evaluated using three measures: search outcome (successful, unsuccessful or abandon), time spent and number of pages visited. Authors concluded that low-literacy users benefit from the system in terms of time spent and number of pages visited. However, users with higher literacy have a slightly worse performance with this system.

To the best of our knowledge, there are no works exploring the influence of health information literacy in web searches.

12.3 RESEARCH QUESTIONS

The experiment described in Chapter 8 served as basis for this study. Two main research questions guided this research:

1. How is health query formulation behavior affected by health literacy and familiarity with the topic?

2. How does the access to lay and medico-scientific content affect query reformulation in general and at different levels of health literacy and topic familiarity?

Besides these two main questions, we also defined a third, secondary research question. It is secondary because the research settings are not the ideal to analyze it. Even so, we think it is still possible to do a superficial analysis and raise hypothesis that can be analyzed in further studies. The question is:

3. Is it possible to predict users' health literacy and topic familiarity through their past terminology in health queries?

12.4 DATA ANALYSIS

To address the first research question, we characterize the queries initially formulated by users pertaining the presence of medico-scientific terminology, advanced operators, spelling errors and also the format of the query. We do it in a general way and also by health literacy and topic familiarity. We consider the query has medico-scientific terminology if it contains the disease/condition technical term as defined in the glossary of technical and popular medical terms described in Chapter 2. For example, for the information situation "About 3 days ago, I started having a burning feeling every time I urinated.

How should I treat this?", the query had to include the term *dysuria*. As advanced operators we consider the OR operator, phrase search (""), exclusion of terms (-) and fill the blanks (*). A query is considered to contain spelling errors if it includes at least one misspelled term. This is particularly important in health queries because medical terminology, mostly the scientific one, is hard to spell by users that are not health professionals. If the query begins by question words like 'how', 'what', 'when', 'where', 'who', 'why' or ends with a question mark, it is considered to be in a question format. To address the second research question we analyze how the access to content with lay and medico-scientific terminology affects the subsequent queries with respect to terminology.

To compare the number of terms employed by users with different health literacy and topic familiarity levels, we have used the ANOVA test. In all the other comparisons, we have used the test of equal proportions between pairs of samples. For example, to compare the inadequate HL group with the elementary HL group regarding the use of medico-scientific terminology, we compare the proportion of queries that include this type of terminology in the first group with the proportion of queries that use it in the second group. Although we present the chi-squared value for the proportion tests, note that, when comparing two samples, the chi-squared test for equality of two proportions is the same as a z-test. In fact, the chi-squared distribution with one degree of freedom is the square of a normal deviate one. Since we are performing multiple comparisons, we applied the Bonferroni correction in these tests, dividing α by 3, the number of tests performed. We use a ** to represent significant results at 0.01 and * for significant results at 0.05. To compute the Confidence Intervals (CI) we use the t-student statistic in the mean number of terms and the chi-squared distribution in the remaining ones.

12.4.1 Query formulation behavior

The mean number of terms in the initial query was 4.1 (95% CI: [3.9, 4.3]) with a standard deviation of 1.8. From the initial queries, 7.2% (95% CI: [4.7%, 10.7%]) included medico-scientific terminology, 26.6% (95% CI: [21.9%, 31.8%]) included advanced operators, 12.5% (95% CI: [9.2%, 16.7%]) were formulated in a question format and solely 1.2% (95% CI: [0.4%, 3.4%]) contained spelling errors. As expected, the proportion of initial queries with medico-scientific terminology is significantly higher in users who already knew the term: 22.1% (95% CI: [14.5%, 32%]). Still, most of these users formulate an initial query without medico-scientific terminology. The proportion of spelling errors is higher in queries formulated with medico-scientific terminology (4.3%) than in queries without it (1%), yet this proportion difference is not statistically significant.

An analysis by health literacy shows no significant differences in the number of terms and spelling errors by health literacy level. Regarding medico-scientific terminology, we found that good HL users use it significantly more than elementary HL users ($\chi^2(1)$=10.6, p=5.7e-04**). We also found that inadequate HL users employ advanced operators less often than elementary HL users ($\chi^2(1)$=8.3, p=2e-03**) and good HL users ($\chi^2(1)$=9.4, p=1e-03**) and design their query in a question format more often than good HL users ($\chi^2(1)$=10.7, p=5e-04**).

Results regarding the use of advances operators make us suspect that health literacy and web search expertise may be related. To verify this, we decided to analyze the relation between users' health literacy and the degree of success they think they have in general web search (Figure 12.1) and in health web search. In general web search, evaluated in a 5-value scale where 1 corresponds to the lowest success rate and 5 to the highest success rate, the median of the web search success is 4 in all levels of health literacy. However, the proportion of answers beneath 4 is higher in the inadequate health literacy level. Through the Chi-Squared test of independence, we found that web search success and health literacy are related ($\chi^2(4)$=54.3, p=4.6e-11**) having a weak positive association with a Spearman correlation of 0.34. In terms of health web search success, assessed in the same scale as web search success, its median is lower (2) in the inadequate HL level than in the other levels (3). Plus, we found that these variables are related ($\chi^2(6)$=32.3, p=1.4e-05**) with a positive, but low, Spearman correlation (0.19).

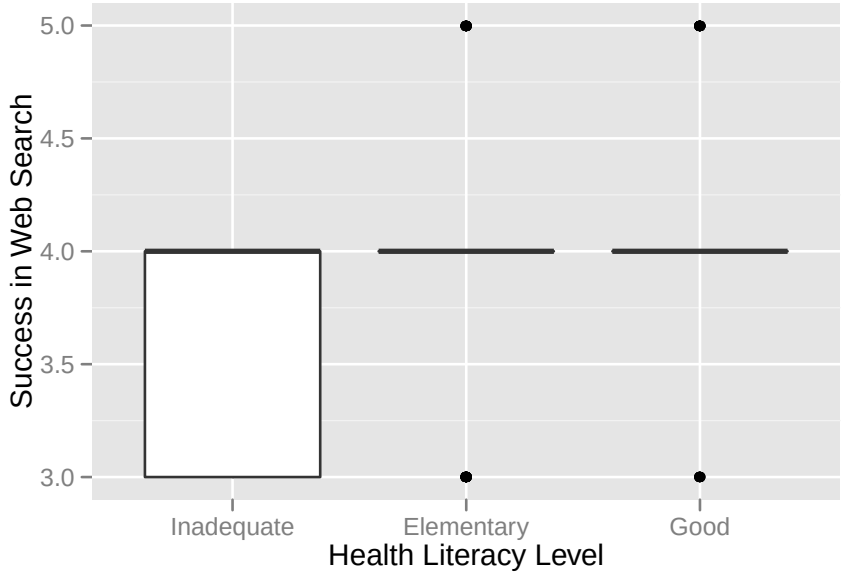

Figure 12.1: Success in web search by level of health literacy.

In terms of topic familiarity, we found that the number of terms does not significantly differ between levels of familiarity with the topic. We found that users who are not familiar with the topic use medico-scientific terminology less often than somehow familiar users ($\chi^2(1)$=16, p=3.16e-05**) and familiar users ($\chi^2(1)$=7.4, p=3e-03*).

12.4.2 Query reformulation behavior

After assessing documents retrieved with medico-scientific queries, the proportion of subsequent queries using medico-scientific terminology is 19.4%, while in lay sessions this proportion downs to 7.8%, a significant difference ($\chi^2(1)$=17.2, p=1.65 e-05**). After search tasks without medico-scientific terminology the proportion of queries in question format is 13.4%, higher than in tasks with it (12.5%), but not significantly different. Since queries in the format

of a question indicate user's difficulties in the search task (Aula et al., 2010), this may that indicate medico-scientific content helps in query reformulation and is probably richer in alternative terms.

Table 12.1 shows the proportion of queries with medico-scientific terms in query reformulation. Similarly to the initial queries, users that already know the scientific term, use medico-scientific terminology significantly more than the other users. In a global perspective, 24.7% of the post-search queries, formulated by users who knew the scientific term before the search session, use medico-scientific terminology. In contrast, in users who did not know the scientific term, only 8.9% of the post-search queries include this type of terminology. This difference is statistically significant, in general, and also after lay and medico-scientific sessions, what shows the importance of knowing the scientific term to the use of this type of terminology in future queries. However, lay sessions discourage the use of medico-scientific terminology even in users who already know the scientific term. In these users the proportions lowers from 34% in medico-scientific sessions to 14.4% in lay sessions.

Table 12.1: Proportion of reformulated queries with medico-scientific terminology by type of session in users who previously knew/knew not the scientific term.

Type of session	Know	Know not	Know > Know not?
All sessions	24.7%	8.9%	$\chi^2(1)$=27.2, p=9.02e-08**
Technical	34.0%	12.7%	$\chi^2(1)$=18.6, p=8.16e-06**
Lay	14.4%	5.2%	$\chi^2(1)$=6.4, p=0.006**

In reformulations including medico-scientific terminology, we also analyzed the reasons for this change. This type of terminology could have been excluded in the first query because it is part of users passive vocabulary and not of its active vocabulary. On the other hand, the documents assessed in the first iteration might have introduced new terminology to the user. Since, for each user, we are aware of his prior knowledge about the scientific term, we can use this information to distinguish both cases. As can be seen in Table 12.2, users who previously knew the scientific term used it in 44.3% of the post-search queries. Consequently, this is the proportion of cases where the scientific term was not included in the first query because it is part of users' passive vocabulary. From the 95% confidence interval, it is not possible to conclude that this proportion is significantly different from 50% and, therefore, significantly different from the proportion of cases due to terminology learning.

In Table 12.2 it is also possible to see that the first post-search queries including scientific terminology were mostly formulated by users who had the scientific term in their passive vocabulary. The opposite happens with the second post-search queries, that is, the majority of the users using medico-scientific terminology in the second query have just learned the term in the search session. Similarly to what happens with the global post-search query analysis, through the confidence intervals, we cannot conclude these proportions are significantly different from 50%.

In terms of health literacy, after medico-scientific tasks, good HL users are

Table 12.2: Proportion of post-search queries with medico-scientific terminology formulated by users that knew the scientific term and did not use this terminology in the pre-search query.

Post-search query	Scientific term known	95% CI
First	57.9%	[34%, 79%]
Second	39.2%	[26%, 54%]
Either first or second	44.3%	[33%, 57%]

more likely to use medico-scientific terminology (22.2%) when compared with elementary HL (16.4%) and inadequate HL (18.1%) users. None of these differences is significant. We also found that, in all levels of health literacy, the majority of the users (proportions between 62.5% and 69.4%) formulated one of the subsequent queries with medico-scientific terminology. Moreover, although in a very low proportion, only users with elementary (1.9%) and good health literacy (1.4%) formulated both queries with medico-scientific terminology.

An analysis by topic familiarity shows that the use of medico-scientific terminology after medico-scientific tasks increases with the topic familiarity (9.5% for not familiar, 26.8% for somehow familiar and 34% for familiar users). In terms of significant differences, we found that users who are not familiar with the topic are less prone to use medico-scientific terminology after medico-scientific sessions than somehow familiar ($\chi^2(1)$=12.9, p=1.6e-04**) and familiar ($\chi^2(1)$=15.7, p= 3.7e-05**) users.

Analyzing the number of subsequent queries with medico-scientific terminology, we found that the proportion of users with 2 medico-scientific queries increases with topic familiarity (from 0.6% to 0.9% to 4.4%) and the opposite happens with the proportion of users with 0 medico-scientific queries (from 40.4% to 23.9% to 21.7%). The majority of users in each level of topic familiarity wrote 1 medico-scientific query but, in users not familiar with the topic, this proportion is much lower than the proportions in the other groups of users.

12.4.3 Prediction of health literacy and topic familiarity

It is not our main goal to predict user characteristics through the data collected in this study. Nevertheless, to help rise hypothesis that can be explored later, we decided to include a research question related to user characteristics prediction. Our ultimate goal is to investigate if it is possible to predict users health literacy or familiarity with a health topic through their past behaviors, namely through their past use of medico-scientific terminology on health searches and on specific health topics. If possible, this could allow search engines to implicitly acquire these context features from search logs.

In Section 12.4.1, based on the queries reported by the users, we found that good health literacy users employ medico-scientific terminology significantly more than elementary users. Moreover, we found that users who are not familiar with the topic significantly use medico-scientific terminology less often than somehow familiar and familiar users. To complement these findings, in this section, we analyze users' habits through their answer to a question

in the pre-search questionnaire in which users rated how frequently they use medico-scientific terminology in their health searches. This allows us to draw conclusions based on two different sources, each with different value. While the information source used in Section 12.4.1 is almost an explicit behavior, user's perception on the regularity with which they use medico-scientific terminology is representative of past behaviors.

Users' answers to the question in the pre-search questionnaire were rated in a scale of 1 (Never uses medico-scientific terminology in health searches) to 5 (Always uses medico-scientific terminology in health searches). About health literacy, as seen in Figure 12.2, we did not detect a clear trend, preventing us from raising hypothesis. Differences between levels of health literacy are not statistically significant. Concerning the topic familiarity level, Figure 12.3 allows us to raise the following hypothesis: "Users familiar with a health topic are more likely to have used medico-scientific terminology in their past searches about that topic than users with less familiarity with the topic". Yet, differences between levels of topic familiarity are not statistically significant.

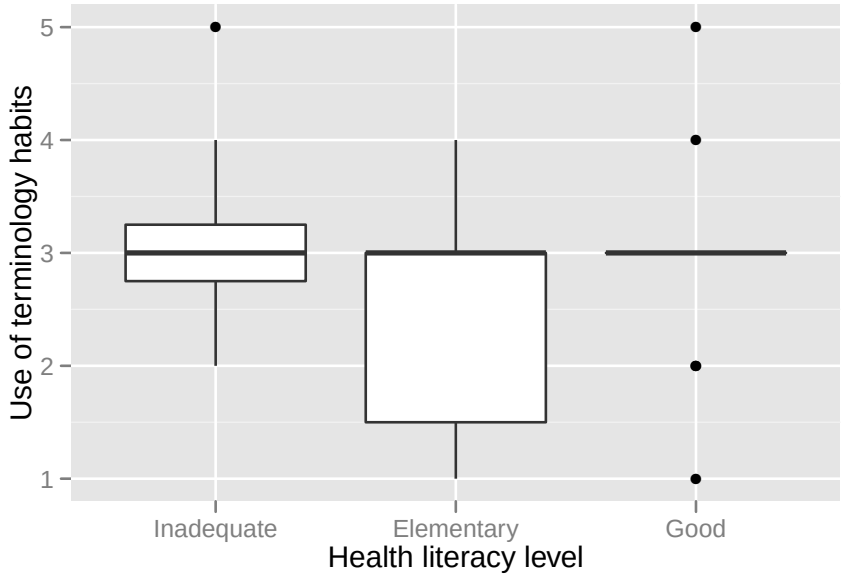

Figure 12.2: Use of medico-scientific terminology habits and users' health literacy.

12.5 DISCUSSION

We verified that health consumers rarely use medico-scientific terminology and that, as expected, users who know the scientific term use it more often. However, even these users include it in only 1 out of 4 health queries. Moreover, we found that users with good health literacy use it more often than elementary health-literate users. In terms of topic familiarity, users who are not familiar with the topic use medico-scientific terminology less often than other users. This is in accordance with previous studies (Wildemuth, 2004; Birru et al., 2004) that conclude that users with less knowledge on the topic and less literacy have less ability to include specific terms in queries.

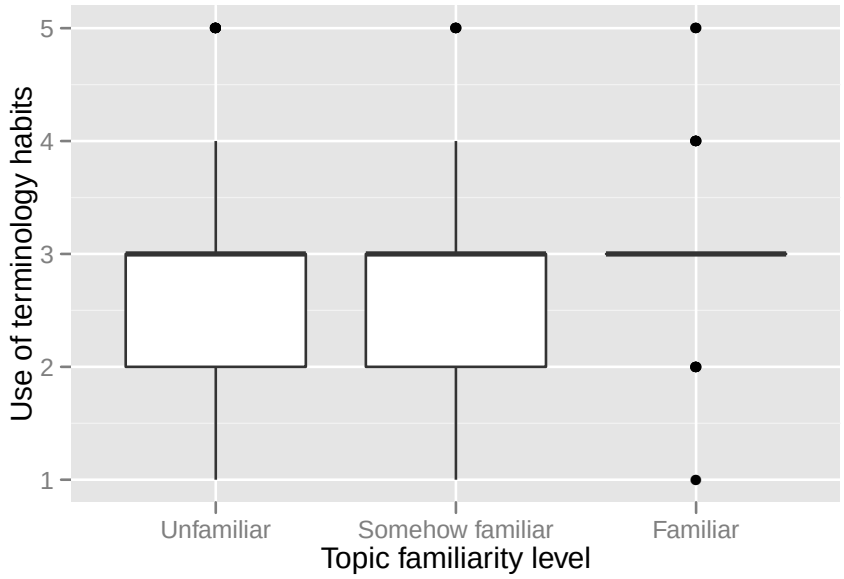

Figure 12.3: Use of medico-scientific terminology habits and topic familiarity.

If queries in question format indicate difficulties (Aula et al., 2010), the formulation of health queries is harder for inadequate health-literate users than for good health-literate users. Furthermore, the former class of users employs advanced operators less often than other users, which indicates they may have less search experience and less ability to fully exploit the potential of search engines. The weak positive association found between web search success and health literacy agrees with the above findings. Moreover, this is in accordance with Summers and Summers (2005), who found that low health literacy users avoid searches because they have difficulties formulating queries and processing results' pages. Since users with inadequate health literacy are ill prepared for conducting health web searches, search engines should focus their attention on this group of users, providing special help mechanisms in health query formulation.

Concerning query reformulation, we found that access to documents containing medico-scientific terminology encourages the use of this type of terminology in subsequent queries. In fact, after medico-scientific sessions the proportion of subsequent queries using scientific terminology is significantly higher than after lay sessions. This happens in users who didn't know the scientific term and in users who knew it. The former learn the scientific term through the documents assessed in the search session and the latter use it from their passive vocabulary.

We found that about half of the queries reformulated to include medico-scientific terminology were a result of terminology learning. The other half had to do with forcing the use of passive vocabulary where the scientific term was included. Through the analysis that distinguishes the first and the second post-search queries, we found that users who have learned the scientific term in the search session tend to use it more in the second query than in the first. This shows that these users have reluctance to use it and only do it as a further alternative.

After the medico-scientific sessions, users who are not familiar with the topic are less prone to use medico-scientific terminology than more familiar users. For this reason and because they use medico-scientific terminology less often in the initial query, these users' health query formulation should also be given special attention by search engines.

If search engines understand the differences between low and high health literacy users and between users familiar and not familiar with a topic, they can develop strategies to better support each type of user find the information they need. Strategies include new interfaces or new features that leverage users' understanding of the information retrieved (e.g.: providing definitions of medical terms). The Invisque system (Kodagoda et al., 2012) is an example of a visual interface developed to help low literacy users overcome search difficulties. In addition, systems can also adjust the ranking of the documents or develop query suggestion mechanisms in which the terminology of the suggested query is adjusted to users' knowledge. Queries can be a simple translation of the initial user query or can introduce new related terms. Both low and high health literacy/topic familiarity users can benefit from such a system. The former type of users probably benefit from translations from medico-scientific to lay terminology which can, for example, happen when users don't understand the terminology used by clinicians or the one included in medical reports and want to inform themselves. Moreover, they can benefit from lay queries using synonyms or related terms. On the other hand, users with more knowledge can also benefit from translations to medico-scientific terminology. Considering the query reformulation findings, the benefits of a query suggestion system might be twofold. It not only provides access to documents that wouldn't be reached without the given suggestions but also stimulates the use of different terms in subsequent queries. Nonetheless, to guarantee that users understand the retrieved documents, it is important to adapt the query suggestions to users' health literacy and topic familiarity.

Based on the query formulated by users in the pre-search questionnaire, we found that users with good health literacy use medico-scientific terminology more often than elementary health-literate users. Moreover, users who are not familiar with the topic use medico-scientific terminology less often than other users. Based on the answer given by users to the question about how frequently they use medico-scientific terminology in their health searches, we found that users with more topic familiarity tend to use medico-scientific terminology more often. We found no significant differences in the use of medico-scientific terminology by users with different levels of health literacy. These conclusions are drawn from two different information sources. The former is composed by only one query and therefore is not representative of past behaviors. The latter has the disadvantage of not being based on explicit behaviors. Through the combination of both sources, we are able to formulate the hypothesis: "Users familiar with a health topic are more likely to have used medico-scientific terminology in their past searches about that topic than users with less familiarity with the topic", that should be explored in future work.

12.6 CONCLUSION

In this work we analyze how the type of terminology used in past queries affect query formulation and reformulation in users with different levels of health literacy and familiarity with the topic. If not the first, this is one of the first works dealing with health literacy in the information retrieval domain. Although some of the results are predictable, we consider important to have their empirical demonstration. In addition, we also analyzed if the prediction of health literacy and topic familiarity may be done using the users past queries.

We found that, although consumers rarely use medico-scientific terminology in their queries, the ones with higher health literacy or topic familiarity do it more often. Users with low health literacy or topic familiarity were found to have more difficulties in query formulation, not only selecting and typing the appropriate medical terms but also on general aspects like the inclusion of advanced operators. The contact with documents using medico-scientific terminology encourages the use of this type of terminology in future queries. Although this is statistically significant in every user, users who did not know the medico-scientific term from the beginning seem more reluctant to use this type of terminology.

Analyzing behaviors of users with different characteristics can help search engines to define how they can provide a better experience for each type of user. This can be done in query formulation, in ranking or at the interface level. As expressed above, we believe a personalized query suggestion system that translates queries between the medico-scientific and the lay terminology can be beneficial to consumer health information retrieval.

Through the prediction analysis we also raised the hypothesis: "Users familiar with a health topic are more likely to have used medico-scientific terminology in their past searches about that topic than users with less familiarity with the topic", that should be explored in future work.

Following the knowledge acquired in studies previously described, in the next Part of this dissertation we describe and evaluate a prototype of a suggestion system offering query suggestions combining the English and Portuguese languages with the lay and medico-scientific terminology.

PART IV

QUERY SUGGESTION SYSTEM: IMPLEMENTATION AND EVALUATION

SUGGESTION SYSTEM

13.1 INTRODUCTION

Searching for health information online is the third most popular activity, following email and using a search engine, being done by 80% of the U.S. Internet users (Fox, 2011). This domain poses specific challenges to health consumers that frequently have additional difficulties in finding the right terms to include in the their queries (Zeng et al., 2006; Kriewel and Fuhr, 2010; Zhang, 2011) because they lack proper medical terms (Zhang, 2010; Toms and Latter, 2007). The misspell of medical terms is another usual problem (Kogan et al., 2001; McCray and Tse, 2003). For these reasons, support in query formulation may contribute to an improved retrieval experience. In addition, this is a domain where the quality of the retrieved contents is crucial and, since we previously concluded that certain languages could result in higher-quality contents (see Chapter 9), support in query translation may be useful.

Query formulation support may be given through query or term suggestions based on the initial query. According to Wildemuth (2004), current search systems offer little support to help searchers formulate or reformulate effective search strategies. A recent review of search interfaces in consumer health websites revealed that only 50% of the analyzed sites provided query suggestions, mainly in a list format (Zhang, 2011). According to the same study, only one of the 18 analyzed sites provided auto-completion and none provided a function to help users find proper medical terms to describe their needs.

In our research, we are interested in improving the health search experience of general users and, in particular, of those who do not have English as their primary language. In previous studies, we already concluded that search engines should propose queries using lay and medico-scientific terminology (see Chapter 10) and, for some non-English users, English queries (see Chapter 9). The suggestion mechanism should be personalized to users' English proficiency, health literacy and topic familiarity. The importance of query formulation support in health searches, the lack of this type of support and the findings of previous studies motivated the development of a system that, based on the user's initial query, suggests 4 different queries combining two languages (English and the users' native language) and two terminologies (lay and medico-scientific).

The study, described in this chapter and the two following ones, has two major goals. First, it intends to assess user's receptivity to these four types of suggestions considering three user characteristics: English proficiency, health literacy and topic familiarity. It is important to note that the effectiveness of query suggestion systems largely depends on the users' selection of suggested

queries, which is usually not considered in the evaluation of this type of systems. Second, considering the same user characteristics, it aims to assess if the use given to suggestions and their terms is beneficial to precision, medical accuracy of the obtained knowledge and motivational relevance.

Previous studies concluded that search assistance should be personalized to achieve its maximal outcome (Jansen and McNeese, 2005). Yet, little attention has been focused on how people perform query reformulations across different user groups. In earlier works we have already focused on the influence of query language and terminology in users with different English proficiency, health literacy and topic familiarity. However, this study is different for several reasons. Here, we don't compare suggestions' languages or terminologies but compare the impact of using or not using each type of suggestion. In addition, this study is less controlled in the sense that, here, users choose whether or not to use the suggestions and this is relevant to the real usage scenario of this tool. This usage awareness will be pertinent to help clarify which personalization will result in the best outcome, define which suggestions should be presented to whom and even what type of personalization is preferable. We may prefer a system that recommends suggestions or a system that executes actions automatically through, for example, changes in the set of retrieved documents. Finally, this work also innovates because, that we know of, there are no works exploring cross-language query suggestions in the health domain.

The remaining sections of this chapter describe related works reported on the literature and the implemented suggestion tool.

13.2 RELATED WORK

In the following sections, we describe works that point out the importance of query assistance in IR, propose assistance mechanisms and evaluate them. In the second subsection we specifically focus on works related to query suggestions and query expansion on the health domain.

13.2.1 Query assistance in IR

The importance of a strategic help given by the interface for information retrieval is widely recognized (Brajnik et al., 2002; Jansen, 2005). About strategic help, Bates (1990) defines several concepts related to the level of system involvement, levels of search activities, search tactics and stratagems. In the health area, Bhavnani et al. (2003) describe a system where specific search procedures are proposed during web search and suggest this procedural knowledge can improve the efficacy, efficiency, and satisfaction of searchers.

A study developed by Jansen (2005) analyzed how often and when do users seek and use automated assistance in the search process. For this purpose, the author developed an automated assistance application that provides help in structuring queries, spelling, query refinement, managing results and relevance feedback. The evaluation was done through a user study with 30 subjects interacting with the system and revealed that approximately 50% of the subjects sought assistance and, out of those, over 80% used it. Users are more willing to accept assistance after viewing results and locating relevant documents. The same author, as a co-author of another work Jansen and McNeese (2005), did a more detailed study about the interaction of users with an

automated searching assistance. In this study, authors conducted a counterbalanced, within subject, user study with 40 subjects and they identified and categorized patterns of interaction. They found that users use the assistance 71% of the times they view it and that they are most receptive when the assistance is viewed after scrolling a results page. In the same study, authors also evaluated the effectiveness of the assistance, considering two types of relevant documents: the ones selected by the user and the ones defined as relevant in a TREC collection. In both cases, only half of the users performed better on the system with assistance. Authors conclude that automated assistance should be personalized to achieve its maximal outcome.

Kriewel and Fuhr (2010) evaluate a suggestion tool available on a digital library that provides several types of suggestions, including terminological and strategic ones. Using 24 subjects in a between-subjects design, authors found a correlation between the number of documents saved and the use of suggestions, but not in all tasks. Moreover, they found that users receiving suggestions, even not using them, used the system's advanced tools more often.

While the previous studies analyzed user behavior when IR systems have automated assistance and evaluate the global utility of these mechanisms, other studies evaluate the utility of specific suggestions and specific query suggestion algorithms. Two main methods are usually adopted for this purpose. The first involves users' judgments, either directly asking for post-retrieval judgments on queries' utility (Ma et al., 2012) or through the analysis of users' behavior in search logs. For example, using search logs, some authors consider good queries the ones reformulated by users (Albakour et al., 2011; Torres et al., 2012) and others consider good queries the ones that lead to clicks on documents (Bing et al., 2011). The second method measures queries' utility through the performance of queries' results using, for example, existing test collections.

Since, in this study, we specifically focus on support through query suggestions, we will, in the following section, center on works that are directly related to query suggestion in the health domain. Moreover, since our analysis also addresses the usage of terms from the presented suggestions, we will also focus on works about query expansion. Query suggestion differs from query expansion because it suggests entire queries instead of suggesting only terms. In both cases, the selection of terms can be based on search results or knowledge structures that can depend on the collection (e.g.: use of terms clusters based on the collection of documents) or not (e.g.: use of a domain-specific thesauri) (Efthimiadis, 1996). As described in Section 3.3.1, Manning et al. (2008) classify the class of methods that use the search results as *local*, and the one using knowledge structures, as *global*. In addition, in both cases, the system involvement may range from Bates's level 3a to level 4b (Bates, 1990), that is, from a system that monitors the search process and recommends search activities (level 3), at user request (3a) or not (3b), to a system that executes actions automatically (level 4) and informs the user (4a) or not (4b). Efthimiadis (1996) did a detailed review about query expansion.

13.2.2 Query formulation support in HIR

In the consumer health information retrieval area, there is an awareness that several difficulties emerge due to the terminology gap between medical experts and medical lay people (Zielstorff, 2003). To overcome the difficulties in

query formulation some authors have proposed query expansion approaches. The Health Information Query Assistant proposed by Zeng et al. (2006) suggests terms based on their semantic distance to the original query. To compute this distance, authors use co-occurrences in medical literature, log data, and medical vocabularies' semantic relations. A user study with 213 subjects randomized in 2 groups, one using suggestions and the other without them, showed that recommendations resulted in higher rates of successful queries, that is, queries with at least one relevant result on the top-10, but not in higher rates of satisfaction. Two proposed search engines for health information retrieval — iMed (Luo and Tang, 2008) and MedSearch (Luo, 2009) — include the suggestion of medical phrases to help the user refine the query. In these systems, the phrases were extracted and ranked based on MeSH, the collection of crawled webpages and the query. Zarro and Lin (2011) present a search system that uses social tagging and MeSH to provide peer and professional terms to users. To evaluate the suggestions' impact, a user study with 10 lay subjects and 10 expert subjects was conducted. They found no differences between the groups. Both groups preferred MeSH terms because their quality was considered superior to the quality of social tags. Also in the health domain, Fattahi et al. (2008) propose a query expansion method that uses non-topical terms (terms that occur before or after topical terms to represent a specific aspect of the subject like 'about' in 'about breast cancer') and semi-topical terms (terms that do not occur alone like 'risk of' in 'risk of breast cancer') in conjunction with topical terms (terms that represent the subject content of documents like 'breast cancer'). Authors found that web search can be enhanced by the combination of these three types of terms.

Although not in the specific area of health information retrieval, the only work we found involving the proposal of query suggestions in a language different from the original query's language is done by Gao et al. (2010). Authors propose a method to translate queries using query logs and then estimate the cross-lingual query similarity using information such as word translation relations and word co-occurrence statistics. They found that these suggestions, when used in combination with pseudo-relevance feedback, improve the effectiveness of cross-language information retrieval.

13.3 SUGGESTION TOOL

We designed and developed a prototype of a suggestion tool that can be integrated in Information Retrieval (IR) systems. Given a health query, our tool suggests alternative queries in two languages, Portuguese and English, using two types of terminologies, medico-scientific and lay ones. As an example, for the Portuguese lay query 'tumor abdominal', the system suggests the following queries: 'abdominal tumor' (English lay query), 'abdominal neoplasm' (English and medico-scientific query) and 'neoplasia abdominal' (Portuguese and medico-scientific query). In Figure 13.1, we present the architecture of the suggestion tool that will be detailed in the following paragraphs.

The system uses the Consumer Health Vocabulary (CHV) described in Section 2.3.2. Consequently, according to Manning et al. (2008), this is a global method using a collection-independent knowledge structure (Efthimiadis, 1996). The CHV vocabulary was translated to Portuguese using the Google Transla-

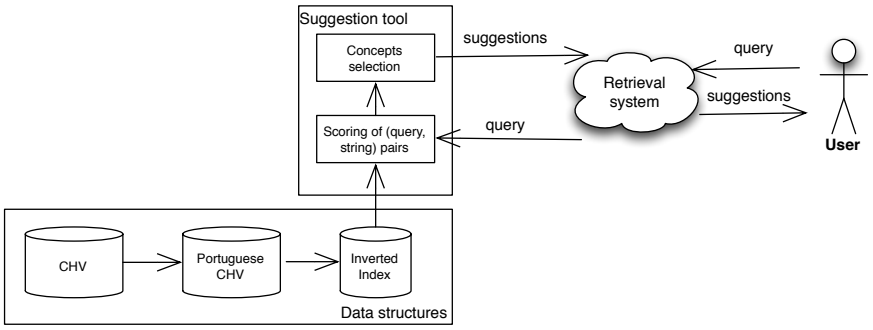

Figure 13.1: Architecture of the suggestion tool.

tor API. To evaluate the translation process we manually assessed 1620 strings, that is, 1% of the total number of strings of the CHV vocabulary. We found that 84.2% (95% CI: [82.3%, 85.9%]) of the translations were correct, 9.7% (95% CI: [8.3%, 11.3%]) of the strings were translated to Brazilian Portuguese (pt-BR), 3.6% (95% CI: [2.8%, 4.6%]) were incorrect translations and in 2.5% (95% CI: [1.8%, 3.5%]) of the strings, the authors had doubts regarding the translation.

With the Portuguese translation of the CHV, we created an inverted index where each stemmed term is associated with an inverse string frequency (isf_t) and a postings list, that is, the list of strings where it appears. The computation of the inverse string frequency is similar to the one traditionally done in information retrieval to compute the inverse document frequency. Since strings are usually small, the probability of finding a string with multiple occurrences of the same term is very small. For this reason, we decided to weight each term based only on its isf_t, ignoring its frequency in each string ($tf_{t,s}$). To determine the vocabulary of terms, strings were tokenized and stop-words were removed. In the remaining terms, letters were reduced to lower case, the accents were removed and terms were also stemmed.

The score assigned to the pairs (query, string) is defined by the following equation: $\text{score}(q, s) = \sum_{t \in q} isf_t$. Since the length of strings and queries has a very small variance, we found that the extra computational power needed to normalize the above score formula would not justify its gains.

In this initial stage of the prototype development, to limit the number of suggestions, we decided to select only the string with the maximum score for the input query. For this string, we select its associated concept and then return its CHV and UMLS preferred names in English and Portuguese. If a suggestion is identical to the query or to other suggestions, it will not be presented, that is, the output of the system only contains unique suggestions different from the query.

Users interact with the retrieval system issuing a query that may be the initial query or a subsequent query, influenced or not by the suggestions presented in the previous iteration. The integration of the suggestions in the search result page can, for example, be done as shown in Figure 13.2, that is, before the first result.

Figure 13.2: Example of a results' page including suggestions.

13.4 CONCLUSION

In this chapter we reviewed the literature pertinent to the implementation and evaluation of query suggestion systems and describe the implementation of a suggestion system prototype. This system uses the CHV vocabulary and offers the user a maximum of 4 alternative suggestions using a combination of Portuguese or English language with lay or medico-scientific terminology.

The next chapter will describe the experiment we have conducted to analyze if this system leads to a more successful health search experience. In addition, we present the results gathered after the data analysis.

EVALUATION

14.1 INTRODUCTION

In this chapter we describe the experiment – User Experiment 3 – conducted to analyze users' interaction with the suggestion prototype and to evaluate if, what and how are the suggestions useful to users with different levels of English proficiency, health literacy and topic familiarity. The experiment required the definition of information situations and the design of a task assignment strategy and an execution procedure. Interaction data has been collected automatically. Two medical doctors collaborated in the assessment of the accuracy of the answers provided by the users to the information situations. After the methodology, we present and analyze the results of this experiment.

14.2 METHODOLOGY

To learn how users with different characteristics interact with a system with the suggestion prototype and to evaluate if the query suggestion prototype described above contributes to a more successful experience, we conducted an interactive light IR experiment run on the laboratory with 40 participants.

14.2.1 Research questions

Our experiment was motivated by the following research questions.

- **RQ1**: How are the system's suggestions used? To which suggestions are users most receptive? Does this change with users' English proficiency, health literacy and topic familiarity?
- **RQ2**: Does a system that includes this suggestion tool leads to a more successful search experience in terms of precision, medical accuracy and motivational relevance?
- **RQ3**: Clicking on a suggestion leads to a more successful search experience in terms of precision (RQ3.1), medical accuracy (RQ3.2) and motivational relevance (RQ3.3)? Does this differ with the language and terminology of the suggestions? Does this differ with users' English proficiency, health literacy and topic familiarity?
- **RQ4**: Using terms from a suggestion leads to a more successful search experience in terms of precision (RQ4.1), medical accuracy (RQ4.2) and motivational relevance (RQ4.3)? Does this differ with the language and

terminology of the suggestions? Does this differ with users' English proficiency, health literacy and topic familiarity?

- **RQ5**: Could an awareness of users' English proficiency, health literacy and topic familiarity improve the performance of this suggestion tool? If so, how?

As can be inferred by the research questions, we will evaluate the query suggestion prototype analyzing users' behavior and measuring the performance of the suggestions in terms of precision, medical accuracy and motivational relevance.

14.2.2 Retrieval Systems

Two retrieval systems were used in this study, one that incorporates the developed suggestion tool and presents query suggestions (SYS1) and the other that doesn't (SYS2). Both of them used the Bing Search API to obtain the web results for users' queries. To increase the usability of the interface in terms of learning, we decided to keep the interfaces very simple and similar to the ones used in the most popular search engines (Figure 14.1). The system with suggestions had hues of blue and the other hues of pink.

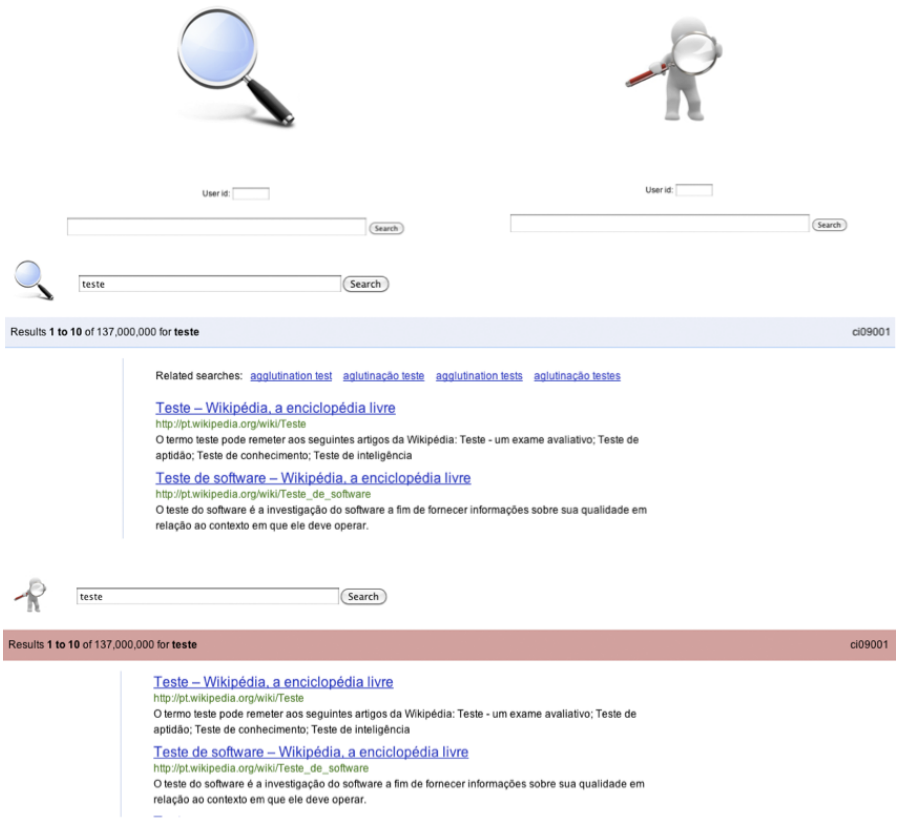

Figure 14.1: Systems' home pages (top left and top right) and result pages (middle and bottom).

In both systems we monitored users' behavior through access logs. We have developed a log mechanism that registers every action made on the sys-

tem, namely, each issued query, search result pages (SERP) presented, URL included in SERP, suggestions included in SERP and user clicks. Each action on the system is associated with a user since we required them to insert an identification code in the system's homepage before each task.

14.2.3 Information situations

To answer our research questions we conducted a user study with 40 participants. Each user was assigned a set of 8 tasks, equally distributed by both systems and each associated with one of 8 simulated work task situations defined as "a short 'cover story' that describes a situation that leads to an individual requiring to use an IR system" (Borlund, 2003b). To define the simulated situations, 20 persons were inquired about the health topic they had most recently searched for on the Web. From these topics, we randomly selected 8 and created a scenario where we included one or more specific questions that had to be answered after the task. We quantified the number of items to include in the answer to facilitate the subsequent medical evaluation of the answers. The information situations were defined as follows:

1. Your mother has just been diagnosed with breast cancer but, shocked with the news, she was not able to ask everything she wanted to the doctor. She only remembers hearing about ductal breast cancer and she is interested in knowing more about her treatment options. Help her and find 4 types of treatments for ductal breast cancer. Give a comprehensive answer, indicating what circumstances may condition each treatment (for example: type of cancer and tumor size).

2. Since the summer, your brother has been feeling a severe pain in the leg, from the buttocks to the knee. A friend has told him that he had the same symptom in the past and was diagnosed with sciatica. Help your brother knowing how this pain can be treated pointing 3 ways to reduce his symptoms.

3. Two weeks ago, someone from your family has been diagnosed with shingles. To understand what characterizes this disease you decided to what are its causes and symptoms. Find out what causes the disease and identify two common symptoms.

4. You suffer from the Irritable Bowel Syndrome. Point out 4 possible ways to alleviate the symptoms. Give a complete answer, indicating what constrains each form of reducing the symptoms.

5. Your younger brother is 2 years old and was diagnosed with atopic dermatitis. Indicate 4 ways to reduce or treat the symptoms associated with this condition.

6. You have just been painfully stung by an insect and, not knowing what insect it was, you want to know how to proceed. Point out 3 possible signs or symptoms to which you should be aware and how you should act on their presence. What is the most serious problem associated with insect bites?

7. You have been diagnosed with hypothyroidism. Indicate 5 symptoms usually associated with this disease.

8. You have been felling shortness of breath. Investigate what may be behind this symptom, indicating 5 possible causes for shortness of breath.

14.2.4 Task Assignment

In the assignment of the tasks we applied a Latin square-like procedure so that users assess the relevance of the answer to each information need exactly once, and use both retrieval systems the same number of times. We have also permuted the order of tasks to avoid possible bias of relevance assessments owing to human behavior. We also guaranteed that each iteration of relevance assessments had tasks in each retrieval system the same number of times and that each retrieval system was associated with each information need the same number of times. The described task assignment results in a within subject design with counterbalancing.

14.2.5 Procedure

Users began by answering two quizzes, one to assess their health literacy, presented in Appendix K, and the other to assess their English proficiency, presented in Appendix J. More details regarding the acquisition of these context features are given in the following section. They then answered an initial questionnaire where, among other questions, they were asked to respond to the questions included in each simulated task without consulting any information sources. After this questionnaire, available in Appendix H, users enrolled in a sequence of 8 tasks. Each task is composed of 3 iterations in which users (re)define a query and assess the 10 first retrieved documents. In the end, a post-search questionnaire is answered. In this questionnaire, presented in Appendix I, users are asked (1) to evaluate the task's topic in terms of familiarity, (2) to evaluate their feeling of success with each iteration, and (3) to, once again, answer the questions included in the simulated situation. The feeling of success with the iterations was evaluated in a 5-level scale from "extremely unsuccessful" (1) to "extremely successful" (5).

For each URL, the user had to indicate the relevance of the document to the information need considering their own context in a 3-value scale: "not relevant", "partially relevant" and "totally relevant", denoted by 0, 1 and 2, respectively.

14.2.6 User context features acquisition

To evaluate users' English proficiency we have used an instrument developed by the European Council that grades English proficiency in the Common European Framework of Reference for Languages (CEFR), a widely accepted European standard for this effect. The Faculty of Arts of the University of Porto validated the use of this instrument in the Portuguese community. This instrument is presented in Appendix J. To have a reasonable number of users in each group, we decided to analyze the data considering only the higher levels of the CEFR, that is, the *basic*, *independent* and *proficient* user.

To evaluate users' health literacy we have used the Medical Term Recognition Test (METER), a brief and self administered instrument proposed by Rawson et al. (2010), along with the suggested cutoff points of 0-20, 21-34, and 35-40 to demarcate *low*, *marginal*, and *functional* health literacy levels. The translated version of this instrument, along with the original concepts, is presented in Appendix K.

Users' familiarity with the topic was self-assessed in a five-level scale: *extremely unfamiliar* (1), *unfamiliar* (2), *neutral* (3), *familiar* (4) and *extremely familiar* (5). In the data analysis, the 5-point user familiarity ratings were further grouped into three categories: *extremely familiar*, *familiar* and *not familiar* (including *extremely unfamiliar*, *unfamiliar*, and *neutral*).

14.2.7 Medical accuracy assessment

The answers given to the questions included in the simulated situations, before and after the tasks, were evaluated in terms of medical accuracy. A committee of two medical doctors defined a list of correct answers for each question included in the simulated situations. Where applicable, this committee also defined items that should be ignored. All the elements that did not belong to these lists should be considered incorrect. To assess the reliability of this classification procedure, a medical doctor and one of the researchers assessed a random set of 30% of the total number of answers (40*8*2=640) with an approximately equal number of pre-search and post-search answers. At this point, the inter-rater agreement was computed and, in the simulated situations where the weighted Cohen's Kappa between assessments was below 0.8, the criteria were further detailed. The weighted Cohen's Kappa is an adaptation of Cohen's Kappa to ordinal scales that treats disagreements differently. With the redefined criteria, the researcher assessed, once again, the same set of answers and reached, for the correctness ratings, a weighted Cohen's Kappa, with squared weights, of 0.90 (95% CI: [0.83, 0.93]), indicating an almost perfect agreement. For the incorrectness ratings, this measure is 0.75 (95% CI: [0.60, 0.83]) indicating a substantial agreement. Since these inter-rater reliability results assure the quality of the assessment procedure, the researcher then assessed the remaining 70% of the answers. The researcher's assessments were the ones considered in the data analysis stage.

14.2.8 Summary of context features

In Table 14.1 we summarize the context features used in this study. We group features into categories and, for each of them, we present its definition, measure scale and data collection methods. The scale transformation of the features related to answers' medical accuracy is discussed in the following section.

14.2.9 Data Analysis

To evaluate the usage given to suggestions, for each type of suggestion, we computed the proportion of suggestion's terms that were used in the subsequent query (termsUsed) and the proportion of the suggestion's terms that were used in the following query and were not used in the previous query (newTermsUsed). Let Q_{it} be the set of unique stemmed terms belonging to

Table 14.1: Summary of context features used in this study.

Category	Context feature	Definition	Scale	Collection method
User	English Proficiency	Users' English skills.	Ordinal: basic, independent and proficient users.	Assessed through an instrument developed by the European Council that grades proficiency in the CEFR levels.
	Health Literacy	The "capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions" (Kutner et al., 2006).	Ordinal: low, marginal and functional health literacy.	Grade obtained in the METER health literacy instrument that was later grouped according with the suggested cutoff points.
User & Document	Relevance	It relates the task and the retrieved documents, "being inferred by criteria like usefulness in decision making, appropriateness of information in resolution of a problem and reduction of uncertainty" (Saracevic, 1996).	Ordinal: from 0 (not relevant) to 2 (totally relevant).	Users' judgment pertaining the usefulness of the document to the resolution of the problem. Part of their assessment task.
User & Task	Topic Familiarity	User's general knowledge about the topic of a search task.	Ordinal: not familiar, familiar and extremely familiar.	Assessed by the user in a 5-level scale and grouped later in three classes.
	Answer's correctness	The degree to which the answer given before and after the search task contains the adequate quantity of correct information.	Varies between 0 (least correct) and 1 (most correct).	Evaluated by a medical doctor and a researcher according to the criteria defined by a committee of two medical doctors. Assessed in an ordinal scale that varies with the information need and later transformed to a 0-1 scale.
	Answer's incorrectness	The degree to which the answer given before and after the search task contains no incorrect information.	Varies between 0 (smallest quantity of incorrect contents) and 1 (largest quantity of incorrect contents).	Evaluated by a medical doctor and a researcher according to the criteria defined by a committee of two medical doctors. Assessed in an ordinal scale that varies with the information need and later transformed to a 0-1 scale.
	Motivational relevance	It relates the user's goals and motivations with the information objects. It is expressed by the user's feeling of success and his satisfaction (Saracevic, 1996).	Ordinal: from 1 (extremely unsuccessful) to 5 (extremely successful).	Obtained through users' answer to the question "Evaluate your feeling of success with the iterations associated with this task.", after the search task.

the query of the iteration *it* and S_{it} the set of unique stemmed terms belonging to the suggestion presented in the iteration *it*, these proportions are computed as follows.

$$\text{termsUsed}_{it} = \frac{|Q_{it} \cap S_{it}|}{|S_{it}|}$$

$$\text{newTermsUsed}_{it} = \frac{|(Q_{it} \cap S_{it}) \backslash Q_{it-1}|}{|S_{it}|}$$

Considering these proportions and the clicks made on suggestions, we analyzed the use of each type of suggestion considering users' English proficiency, health literacy and topic familiarity. It is important to note that, while the proportions allow the assessment of the utility of the suggestions to query expansion, the click analysis allows the evaluation of their utility to query suggestion.

To evaluate the impact of suggestions provided by this system, we measured and compared three major outcomes: precision, medical accuracy and motivational relevance. In precision and motivational relevance, the analysis was done comparing iterations where suggestions were used with iterations where suggestions were not used. In medical accuracy, the analysis was done on a session level, that is, sessions where suggestions were used were compared with sessions were suggestions were not used. The set of three iterations in each task is considered a session. In each outcome, we analyzed the use of general suggestions and the specific use of lay, medico-scientific, English and Portuguese suggestions. Moreover, when considering specific types of suggestions, we considered users' English proficiency, health literacy and topic familiarity. Besides comparing the situation of using a suggestion with the situation of not using it, we also compared different groups of users while using or not using each type of suggestion.

In our analysis, we considered four scenarios of suggestions' use: using or not using suggestions' terms (Terms?), using or not using all the terms of a suggestion (All terms?), using or not using suggestions' terms that were not used in the previous query (NewTerms?) and clicking or not in suggestions (Click?). Note that using all the terms of a suggestion can be different from clicking a suggestion. Clicking on a suggestion always implies using all the terms from a suggestion while the converse is not true. Users may use all the terms from a suggestion without clicking it, or they can change the order of the terms or even mix them with other terms. We decided to focus on the first three scenarios because we think they are the most representative query expansion actions within a system with query suggestions. While the first two scenarios allow the evaluation of the suggestions' terms regardless their inclusion in the previous query, the *NewTerms?* scenario enables the evaluation of the utility of the suggestions' novel terms for each user. Note that the *Terms?* scenario includes situations where terms may exist in a query because they already did and not because the system suggested them and that, although possible, this is less likely in the *All terms?* scenario. These scenarios allow the assessment of the suggestions' utility as a source of terms to query expansion and the click scenario allows the evaluation of their utility as suggestions of queries. To identify and distinguish the first three scenarios we automatically

computed the intersection of the queries' terms and suggestions' terms. The click data was obtained from access logs.

We use Graded Average Precision (GAP) and Graded Precision (gP) with an equally balanced g_1 and g_2, i.e., $g_1 = g_2 = 0.5$ meaning that the levels 1 and 2 of our relevance scale have the same probability of being the grade from which the user considers the documents relevant. These precision measures are described in Section 5.2.1.

As explained in the study setup, users were required to answer the questions defined in the simulated situations, before and after the task. This allows the evaluation of the knowledge that was acquired during the search task through the use of a Δ correctness and a Δ incorrectness. These metrics represent the value after the task minus the value before the task. If positive, it means the search task contributed to increase the correct or incorrect contents. Before the computation of Δ correctness and Δ incorrectness we transformed the correct and incorrect scales to a 0-1 scale. To accomplish this, in the correctness assessments we divided the score by the maximum possible score in that simulated situation. For example, if the simulated situation asked for 3 treatments, the maximum possible score would be 3. The incorrectness assessment is simply the number of incorrect items included in the answer, so the absence of a predefined maximum made us opt to divide the incorrect score by the maximum incorrect score obtained in that simulated situation.

Motivational relevance relates the user's goals and motivations with the information objects. It is expressed by the user's feeling of success and his satisfaction (Saracevic, 1996). To assess motivational relevance, users were asked to evaluate their feeling of success with the iterations in a 5-point scale: Extremely unsuccessful Unsuccessful, Neutral, Successful and Extremely successful. To analyze motivational relevance, we grouped the 5-point success ratings into two categories: successful (including extremely successful and successful) and not successful (including extremely unsuccessful, unsuccessful and neutral).

In the analysis of suggestions' use behavior and motivational relevance we used the test of equal proportions with the chi-squared value to compare proportions between samples. In the usage behavior, and more specifically in the precision and medical accuracy analysis, we compared means between groups (with and without the use of suggestions) using the Student's t-test. When the assumption of homogeneity of variances was not verified, we applied the Welch t-test. To compare groups of users, we applied the one-way ANOVA and the Tukey's test to assess the location of the differences, whenever significant differences were found. When reporting our results, we use a * to mark significant results at $\alpha = 0.05$ and a ** to mark significant results at $\alpha = 0.01$.

14.3 RESULTS AND ANALYSIS

Forty information science undergraduate students participated in this study (24 females; 16 males) with a mean age of 23.48 years (standard deviation (sd)=7.66). The English proficiency assessment revealed an heterogeneous sample of users that, in average, had 19.93 (sd=8.86) in a scale of 0 to 40. In terms of CEFR classes, 16 users have *basic* English proficiency, 17 are *independent* and 7 are *proficient* users.

In the health literacy test, evaluated in a 0 to 40 scale, users had in average

25.55 (sd=7.40). Users' distribution by health literacy classes is the following: *low* (7 users), *marginal* (28 users) and *functional* (5 users). Users' familiarity with a topic depends on the task's subject. The pairs "user, topic" are distributed as follows: *not familiar* (86 pairs), *familiar* (114 pairs) and *extremely familiar* (120 pairs). A deeper analysis showed that two topics single out for being the least familiar (shingles and then insect bites) and one for being the most familiar (hypothyroidism).

In an open question, users were asked about their difficulties when performing health web searches. Answers revealed four major issues: the existence of medico-scientific/difficult terminology (in 55% of the answers), information credibility (35%), query formulation (16%) and finding contradictory information (3%).

14.3.1 Use of suggestions

In this analysis we only consider the iterations where suggestions could have been presented, that is, the second and third iteration of the tasks using SYS1.

On the overall experience, SYS1 did not present suggestions in only 4.7% of the iterations. In 13.1% of the iterations, suggestions were not used and in the other cases, users either used terms from one (34.1% of the iterations) or several suggestions (48.1% of the iterations). An analysis by iteration (Table 14.2) shows that, in the third iteration, SYS1 presents suggestions in a smaller number of cases. This happens because the suggestion tool does not present suggestions after English queries and these queries appear more often in the second iteration of the experience than in the first iteration. Moreover, participants use terms from a larger number of suggestions in the initial iterations where they tend — non-significant difference — to find suggestions useful more often.

Table 14.2: Suggestion use by iteration: proportions and one-sided significant differences. It = iteration. Sug = suggestion.

Usage	It2	It3	It2 vs It3
sug not presented	1.9%	7.5%	$\chi^2(1) = 4.5$, p=0.02*
terms from 1 sug	30.6%	37.5%	$\chi^2(1) = 1.39$, p=0.12
terms from several sug	57.5%	38.8%	$\chi^2(1) = 10.5$, p=6e-4**
sug not useful	10.0%	16.3%	$\chi^2(1) = 3.7$, p=0.07
all sug terms	65.0%	48.0%	$\chi^2(1) = 8.3$, p=0.002**
sug click	41.4%	24.3%	$\chi^2(1) = 9.3$, p=0.002**

After iterations with suggestions, participants used, in average, 1.34 terms from the suggestions. Considering only the new terms, that is, the terms that did not belong to the previous query, this value falls to 0.66. In the analysis by iteration (Table 14.3), we see that the number of suggestion terms used in the subsequent query is higher in earlier iterations. This is in line with what was described above.

Users employ all terms from suggestions in 56.7% of the iterations where suggestions were presented. The above tendency is still true, that is, complete suggestions are used more in the initial iterations as shown in Table 14.2.

Table 14.3: Means of terms and new terms by iteration. One sided significant differences.

	It2	It3	It2 vs It3
Terms	1.89	0.76	t(302.2)=8.7, p=2.2e-16**
New terms	0.82	0.49	t(301)=3.1, p=0.002**

If, instead of entire suggestions, we consider clicks, the proportion of iterations with clicked suggestions falls to 33.1% and is also significantly higher in the first iteration as shown in Table 14.2. Still regarding clicks, we found that 54.7% of the sessions had at least one click and only 8.8% of the sessions had two clicks. We found that a large proportion of users (87.5%) have clicked at least once in the proposed suggestions. However, only 27.5% of the users have clicked on suggestions in the two iterations where they were presented.

As shown in Table 14.4, users extract the larger number of terms from Portuguese/medico-scientific suggestions. In terms of significant differences, we found that the mean number of terms extracted from this type of suggestions is larger than the mean number of terms extracted from English suggestions (Tukey's adjusted p=0 for EN/Lay; Tukey's adjusted p=1.7e-6 for EN/MS). Moreover, as can be seen in Table 14.5, Portuguese/lay suggestions are also preferred to both types of English suggestions. Regarding the use of new terms, the English/medico-scientific suggestions are the ones with the greatest contribution to the expansion of terms in users' queries. This difference is explained by users' lack of habit to begin their searches with an English query. English/medico-scientific suggestions are also the ones with a higher proportion of clicks. On the other hand, Portuguese/Lay suggestions are the ones with lower proportion of clicks and lower proportion of new terms, showing that users consider these suggestions the least useful ones. In fact, this type of suggestions has a significantly lower mean of new terms and a significantly lower proportion of clicks with respect to all the other types of suggestions (Table 14.5).

Table 14.4: termsUsed and newTermsUsed: mean and standard deviation (SD) by type of suggestion. Proportion of clicks by type of suggestion. Boldface indicates the maximum per column.

	termsUsed		newTermsUsed		Clicks
	Mean	SD	Mean	SD	Proportion
EN/Lay	0.24	0.38	0.16	0.33	14%
EN/MS	0.30	0.41	**0.19**	0.37	**18%**
PT/Lay	0.40	0.40	0.08	0.20	5%
PT/MS	**0.46**	0.43	0.16	0.33	14%

Analysis by English proficiency

Table 14.6 presents an analysis of users' preferred language, in general and by English proficiency.

Table 14.5: Tukey's adjusted p-value for one-sided significant comparisons of the termsUsed and newTermsUsed. Holm adjusted p-value for one-sided significant comparison of proportions of clicks.

PT/Lay vs	**PT/MS**	**EN/Lay**	**EN/MS**
termsUsed (>)	-	5e-6	0.01
newTermsUsed (<)	0.002	0.002	1e-5
Click (<)	0.009	0.009	8e-6

Table 14.6: Means, proportions and one-sided differences of termsUsed, newTermsUsed and clicks by language and English proficiency. Boldface identifies the row's maximum.

			English vs Portuguese	
	English	Portuguese	test value	p value
termsUsed				
All	0.27	**0.43**	t(1214.7)=-6.9	3.3e-12**
ep1	0.25	**0.45**	t(483.5)=-5.38	6e-06**
ep2	0.24	**0.44**	t(511.4)=-6.0	2.4e-09**
ep3	**0.38**	0.35	t(209.1)=0.6	0.27
newTermsUsed				
All	**0.18**	0.12	t(1153.7)=3.2	5e-04**
ep1	**0.18**	0.12	t(455.8)=2.1	0.01*
ep2	**0.15**	0.13	t(513,2)=0.7	0.25
ep3	**0.25**	0.10	t(171.7)=3.3	5e-04**
clicks				
All	**15.9%**	9.7%	$\chi^2(1) = 10.1$	7.6e-04**
ep1	**16.8%**	8.2%	$\chi^2(1) = 0.15$	0.003**
ep2	**11.9%**	9.6%	$\chi^2(1) = 8.7$	0.24
ep3	**23.6%**	13.2%	$\chi^2(1) = 1.5$	0.04*

If we consider all the used terms, we can see that *basic* and *independent* English proficiency users prefer to use terms from Portuguese suggestions. In the *proficient* group, users tend to prefer English suggestions but this is not a significant difference. If we only consider the newly introduced terms or clicks, we found that *basic* and *proficient* users extract significantly more terms from English suggestions than from Portuguese ones.

Besides testing the differences between languages, we also tested the differences between levels of English proficiency. In this respect and regarding English suggestions, we found that *proficient* users are associated with a higher mean of *termsUsed* than *independent* (Tukey's adjusted p = 0.001**) and *basic* users (Tukey's adjusted p = 0.006**). Excluding the previously used terms (*newTermsUsed*), *proficient* users employ more terms from English suggestions than *independent* users (Tukey's adjusted p = 0.01*). The same happens in clicked suggestions (Tukey's adjusted p = 0.012*).

In the analysis by health literacy we only considered the information situations associated with a topic that is expressed differently when using lay and medico-scientific terminology. The others are not useful for this analysis.

In Table 14.7 we can observe that, in general, users prefer medico-scientific suggestions to lay ones. However, if we consider users' health literacy, we see that this is only true in the *marginal* and *functional* groups.

Table 14.7: Means and one-sided significant differences of termsUsed and newTermsUsed by terminology, health literacy and topic familiarity. Boldface identifies the row's maximum.

			Lay vs Medico-scientific	
	Lay	Medico-scientific	test value	p value
termsUsed				
All	0.14	**0.46**	t(275.1)=-7.0	9e-10**
hl2	0.14	**0.47**	t(204.6)=-6.3	7e-08**
hl3	0.06	**0.70**	t(27.1)=-5.8	1.7e-06**
tf1	0.12	**0.51**	t(136.4) = -6.3	1.8e-09**
tf2	0.16	**0.48**	t(55.3)=-3.0	0.001**
tf3	0.16	**0.36**	t(80.8)=-2.3	0.01*
newTermsUsed				
All	0.09	**0.32**	t(247.2)=-5.5	9.5e-06**
hl2	0.07	**0.32**	t(176)=-5.0	6e-05**
hl3	0.06	**0.48**	t(24.5)=-3.3	0.001**
tf1	0.00	**0.31**	t(27) = -5.3	7e-04**
tf2	0.15	**0.35**	t(57)=-2.1	0.02*
tf3	0.14	**0.30**	t(79.9)=-1.8	0.03*
clicks				
All	5.3%	**24.3%**	$\chi^2(1) = 20.4$	3e-06**
hl2	4.5%	**21.4%**	$\chi^2(1) = 12.8$	2e-04**
hl3	5.6%	**55.5%**	$\chi^2(1) = 8.4$	0.002**
tf1	2.6%	**22.4%**	$\chi^2(1) = 11.8$	3e-04**
tf2	9.4%	**37.5%**	$\chi^2(1) = 5.6$	0.009**

Besides the differences presented in Table 14.7, we also found that the use of terms from medico-scientific suggestions significantly increases with the health literacy of the users. In fact, the *low* health literacy group uses fewer terms than the *marginal* (Tukey's adjusted p = 0.01*) and *functional* literacy group (Tukey's adjusted p = 5e-04**). In addition, the *marginal* health literacy group uses fewer terms than the *functional* one (Tukey's adjusted p = 0.05*). We also found that the *functional* group has a larger proportion of clicks than the *low* (Tukey's adjusted p =0.013**) and *marginal* group (Tukey's adjusted p = 0.008**).

Analysis by topic familiarity

Still in Table 14.7 we can see that topic familiarity is not discriminative in terms of lay versus medico-scientific terminology use. All topic familiarity levels significantly prefer the medico-scientific terminology. The only exception occurs in clicks by users *extremely familiar* with the topic. In this group we did not found significant differences between terminologies.

Regarding terms from lay suggestions not used in the previous query, we also found that users *extremely familiar* with a topic use them more than *non-familiar* users (Tukey's adjusted p = 0.05*).

14.3.2 Impact of suggestions on precision

Computing the mean GAP for the three iterations we found that this value decreases with the iterations (0.86 in the first iteration, 0.82 in the second iteration and 0.8 in the third one). After discovering significant differences in the mean GAP between iterations, we found that the first iteration has a higher mean GAP than the second iteration (Tukey's adjusted p = 0.009**) and the third iteration (Tukey's adjusted p = 3.3e-06**). These differences may be explained by users' criteria in judging relevance. We believe that documents with reoccurring contents, because they are no longer useful, are assigned lower relevance scores.

Since we found significant differences of the mean GAP between iterations, we decided to conduct our analysis using the GAP difference between iterations (Δ GAP) instead of just the mean GAP. For each iteration we have therefore computed a $\Delta GAP_{it} = GAP_{it}\text{-}GAP_{it-1}$. We found no significant differences between ΔGAP_2 and ΔGAP_3.

SYS1 tends to have a higher Δ GAP mean (-0.031 against -0.033 in SYS2) but this difference is not significant. In all the four scenarios of suggestions' use (Terms?, All terms?, NewTerms? and Click?), we did not find significant differences between using or not using suggestions. Iterations where suggestions are not used tend to have a higher precision in the four analyzed scenarios.

After this general analysis, we repeated it by suggestion's language and terminology. Almost all the comparisons were non-significant. The only exceptions happen in the use of new terms from suggestions, where the mean Δ GAP is higher without Portuguese suggestions than with them (t(147.5)=2.4, p=0.01**). In this same scenario, the mean Δ GAP is also higher without lay suggestions than with them (t(139.5)=2.78, p=0.003**).

Analysis by English proficiency

To analyze the relation between suggestions' language and users' English proficiency, we compared the mean Δ GAP in the four scenarios mentioned above, in each group of English proficiency (Table 14.8). With respect to Portuguese suggestions, although we haven't found significant differences, the general tendency is to have higher precision when users do not use the suggestions. In terms of English suggestions, Δ GAP tends to be higher when a suggestion is clicked or when all the terms of a suggestion are used. Yet, this tendency is only significant when *proficient* users click English suggestions. In the two other scenarios, there is not a clear tendency. Advanced proficiency users tend to have higher precision when they use new terms from an English suggestion

and, surprisingly, the same happens with *basic* proficiency users when they use terms from English suggestions.

Table 14.8: Δ GAP means by language. Boldface represents the maximum in each group and scenario. Square brackets are used there are significant differences between scenarios.

	Terms? [w/o | w/]	All terms? [w/o | w/]	NewTerms? [w/o | w/]	Click? [w/o | w/]
PT				
EP1	-0.05 | **-0.03**	**-0.04** | -0.05	**-0.04** | -0.08	**-0.04** | -0.05
EP2	**-0.01** | -0.04	**-0.02** | -0.06	**-0.02** | -0.06	-0.02 | -0.02
EP3	**-0.03** | -0.08	**-0.04** | -0.05	**-0.03** | -0.10	-0.04 | -0.04
EN				
EP1	-0.04 | **-0.03**	-0.04 | **-0.03**	**-0.04** | -0.06	-0.04 | **-0.03**
EP2	**-0.02** | -0.05	-0.02 | **-0.01**	**-0.02** | -0.08	-0.02 | **-0.01**
EP3	**-0.04** | -0.05	-0.05 | **0.01**	-0.04 | **-0.02**	[-0.05 | **0.03**]*

In each scenario, we compared the mean Δ GAP of the several proficiency levels when using English and Portuguese suggestions but we did not find significant differences.

Analysis by health literacy

Comparing the precision of lay and medico-scientific queries by level of health literacy, we found few significant differences. As can be seen in Table 14.9, two of the three exceptions occur when *marginal* (t(89.9)=2.3, p=0.01*) and *functional* (t(33.7)=2.0, p=0.03*) health literate users use terms from lay suggestions they have not used before. In these cases, the precision is lower than when they do not use these suggestions. The use of lay suggestions tends to be beneficial to precision when *low* health literate users use all the suggestion's terms and when these users and the *marginal* health literate users click in the suggestions.

In medico-scientific suggestions, the use of all the suggestion' terms tends to be favorable to precision in all levels of health literacy. Moreover, the use of new terms from suggestions and suggestions' clicks are beneficial to precision in the *low* and *functional* health literacy groups. Of these, the only significant difference occurs when *low* literate users click medico-scientific suggestions (t(35.4)=-1.9, p=0.03*).

In each scenario, we compared the mean Δ GAP of the several health literacy levels when using lay and medico-scientific suggestions but we did not find significant differences.

Analysis by topic familiarity

Considering topic familiarity, whose results are also presented in Table 14.9, we found that *familiar* users (TF2) have significantly higher precision in iterations where they do not use new terms from lay (t(72.9)=2.2, p=0.01*) or medico-scientific suggestions (t(76.4)=2.0, p=0.02*). We also found that *extremely familiar* users tend to have higher precision with medico-scientific suggestions

Table 14.9: Δ GAP means by terminology. Boldface represents the maximum in each group and scenario. Square brackets are used there are significant differences between scenarios.

	Terms? [w/o | w/]	All terms? [w/o | w/]	NewTerms? [w/o | w/]	Click? [w/o | w/]
Lay				
HL1	**-0.04** | -0.06	-0.05 | **-0.03**	**-0.04** | -0.07	-0.05 | **0.00**
HL2	**-0.02** | -0.04	**-0.02** | -0.04	[**-0.02** | -0.08]**	-0.03 | **-0.01**
HL3	**-0.04** | -0.10	**-0.05** | 0.10	[**-0.05** | -0.11]**	**-0.05** | -0.09
TF1	**0.01** | -0.03	**0.00** | -0.03	**0.00** | -0.06	0.00 | **0.02**
TF2	**-0.04** | -0.06	**-0.04** | -0.05	[**-0.03** | -0.11]*	-0.04 | -0.04
TF3	**-0.04** | -0.05	-0.05 | **-0.04**	**-0.04** | -0.06	-0.05 | **-0.01**
MS				
HL1	**-0.04** | -0.04	-0.04 | **-0.05**	-0.04 | **-0.04**	[-0.05 | **0.01**]*
HL2	**-0.02** | -0.03	-0.03 | **-0.03**	**-0.02** | -0.06	**-0.03** | -0.04
HL3	**-0.04** | -0.08	-0.06 | **-0.03**	-0.05 | **-0.05**	-0.05 | **-0.04**
TF1	**0.00** | -0.01	0.00 | **0.02**	0.00 | **0.01**	0.00 | **0.03**
TF2	**-0.04** | -0.06	**-0.04** | -0.06	[**-0.03** | -0.10]*	**-0.04** | -0.06
TF3	-0.05 | **-0.04**	-0.05 | **-0.04**	-0.05 | **-0.04**	-0.05 | **-0.04**

or when they click or use all the terms from lay suggestions. *Non-familiar* users also seem to benefit from clicks in lay suggestions and from medico-scientific suggestions.

Comparing the mean Δ GAP of the several topic familiarity levels in each scenario and type of terminology, we found that, when using new terms from medico-scientific suggestions, *non-familiar* users have a significantly higher precision than *familiar* users (Tukey's adjusted p=0.01*). This is simultaneously due to the increase in precision in *non-familiar* users and the significant decrease found in the *familiar* users, when using these suggestions.

14.3.3 Impact of suggestions on medical accuracy

In general, the quantity of correct contents after the search task significantly increases (t(319)=-21.62, p<2.2e-16**). The Δ correctness has a mean of 0.45 and a standard deviation of 0.37. The quantity of incorrect contents does not significantly differ (t(319)=0.24, p=0.81) before and after the search task. In fact the mean of the Δ incorrectness is 0 and its standard deviation is 0.32. Contrary to what happens with correct contents, in incorrect contents, the lower the value, the better. This happens because a negative value shows a reduction in the quantity of incorrect contents after the search task. From this point forward, when we simply mention correctness, we mean Δ correctness and when we simply mention incorrectness, we mean Δ incorrectness.

While in the previous section the analysis was based on the iteration, here the analysis will be done on a session level. In fact, the answer is given only after the third iteration. For this reason, in this section, when we mention the use of a suggestion, we mean the use of a suggestion in the session, that is, in

the set of the three iterations that form the task.

A comparison between SYS1 and SYS2 shows us that the system with suggestions outperforms the other system. In terms of Δ correctness, the difference is not significant (0.47 in SYS1 against 0.43 in SYS2) but in Δ incorrectness, we found that SYS1 significantly contributes to reduce answers' incorrect contents more than SYS2 (-0.03 in SYS1; 0.03 in SYS2; t(312)=-1.7, p=0.04*).

As can be seen in Table 14.10, the use of suggestions tends to improve the medical accuracy of the knowledge obtained in the search task, simultaneously contributing to increase the quantity correct contents and to diminish the quantity of incorrect contents. In terms of significant differences we see that the use of all the suggestions' terms significantly increases the correct contents and significantly decreases the incorrect contents. Clicking in suggestions also significantly increases the quantity of correct contents acquired in the session. Moreover, the use of the system with suggestions and the use of suggestions' terms significantly decrease the quantity of incorrect contents.

Table 14.10: Δcorrectness and Δincorrectness comparisons and one-sided differences between systems and the use of suggestions. Boldface identifies the value that shows the highest quality improvement.

	Means		Differences	
	without	with	test value	p
Correctness				
System with sug?	0.43	**0.47**	t(316.5)=1.1	0.12
Terms?	0.43	**0.47**	t(308.3)=-0.9	0.18
All terms?	0.42	**0.5**	t(278.0)=-1.9	0.02*
NewTerms?	0.44	**0.49**	t(156.8)=-1.1	0.14
Click?	0.42	**0.52**	t(161.6)=-2.0	0.02*
Incorrectness				
System with sug?	0.03	**-0.03**	t(312)=-1.7	0.04*
Terms?	0.03	**-0.04**	t(318)=2.1	0.02*
All terms?	0.02	**-0.04**	t(202.3)=1.7	0.05*
NewTerms?	0	**-0.02**	t(174)=0.7	0.47
Click?	0	**-0.02**	t(168.3)=0.6	0.28

Beside the above general analysis, we also did an analysis by type of suggestion. In Table 14.11 we present the results by suggestion's language and, in Table 14.12, the results by suggestion's terminology. As shown in these tables, in general, all the suggestions tend to improve the medical accuracy of the answers. In terms of significant differences, excluding the first scenario where terms from suggestions are used (Terms?), English suggestions always contribute to a higher quantity of correct contents. On the other hand, Portuguese suggestions contribute to reduce the quantity of incorrect contents. Regarding the suggestions' terminology, the use of entire lay or medico-scientific suggestions, through a click or not, increases the quantity of correct contents. The same happens with the use of new terms from medico-scientific suggestions. On the other hand, the use of terms from lay and medico-scientific suggestions

decreases the quantity of incorrect contents.

Analysis by English proficiency

The use of suggestions, in both languages and in the four scenarios, tends to increase the quantity of correct contents. In terms of significant differences, we found that the use of English entire suggestions significantly increases the quantity of correct contents in *basic* proficiency users (w/o: 0.41, w/: 0.55, t(47.3)=-2.0, p=0.05*). The same happens when these users click on English suggestions (w/o: 0.4, w/: 0.56, t(43.0)=-2.9, p=0.01*). We also found that, in general, Portuguese suggestions tend to decrease the quantity of incorrect contents. On the other hand, in all the scenarios, English suggestions tend to have the same effect but only in the *independent* and *proficient* groups. In Table 14.13 we present the significant results found in terms of incorrectness. As can be seen, disregarding the scenario, the use of terms from Portuguese suggestions significantly decreases the incorrect contents in *basic* and *proficient* users. The same effect is also found when *proficient* users click in Portuguese suggestions or when they use English suggestions.

Pertaining the significant differences between levels of English proficiency, whose Tukey's adjusted p-values are presented in Table 14.14, we found that *proficient* users give answers with more correct contents than *basic* and *independent* users, with and without the use of Portuguese suggestions. The two only exceptions occur with clicks in and use of new terms from Portuguese suggestions, where the differences between *proficient* and *basic* users don't exist. The same differences are found when English suggestions are not used. Since we did not find the above differences with the use of English suggestions and, with these suggestions, the quantity of correct contents increase in all groups, we hypothesize that English suggestions contribute to fade the differences between these groups of users.

Surprisingly, we also found that *proficient* users, when not using English suggestions, have a greater increase in incorrect contents than *basic* users. Surprised with this, we found out in part this is due to their different behavior in the pre-search answers, where *basic* proficiency users give answers with significantly more incorrect contents than *proficient* users (0.54 against 0.12, t (99.7) = 3.17, p= 0.001**). In the post-search answers we found no significant differences between both groups in terms of incorrectness. Nonetheless, *basic* users reduce the quantity of incorrect contents when they do not use English suggestions while *proficient* users increase it.

When using English suggestions, we found no significant differences between groups of English proficiency. With English suggestions we found that *basic* users tend to increase the incorrectness of their answers while *independent* and *proficient* users decrease it.

Analysis by health literacy

An analysis by health literacy and terminology, whose significant results are presented in Table 14.15, reveals that the *functional* health group significantly gives answers with more correct contents when using medico-scientific suggestions in all the scenarios in which the users use terms from the suggestions. Moreover, the same happens with the *low* health literacy group when using all

Table 14.11: Δcorrectness and Δincorrectness comparisons and one-sided differences by suggestions' language. Boldface identifies the value that shows the highest quality improvement.

	Portuguese means					English means				
	without	with	degrees of freedom	t statistic	p value	without	with	degrees of freedom	t statistic	p value
Correctness										
Terms?	0.43	**0.47**	281.8	-1.0	0.15	0.42	**0.45**	115.6	-0.5	0.31
All terms?	0.43	**0.5**	177.7	-1.6	0.05	0.42	**0.55**	116.1	-2.7	0.003**
NewTerms?	0.43	**0.5**	127.8	-1.3	0.09	0.42	**0.53**	146.4	-2.4	0.01*
Click?	0.44	**0.51**	60.8	-1.2	0.12	0.42	**0.56**	94.6	-2.6	0.005**
Incorrectness										
Terms?	0.03	**-0.06**	312.6	2.6	0.004**	0	**-0.02**	165.3	0.5	0.31
All terms?	0.02	**-0.06**	203.3	2.2	0.01*	**-0.01**	0	107.3	-0.3	0.40
NewTerms?	-0.01	**-0.04**	160.1	1.37	0.08	0	**-0.01**	137.3	0.2	0.43
Click?	0	**-0.04**	72.3	0.88	0.19	**-0.01**	0	85.8	-0.18	0.43

Table 14.12: Δcorrectness and Δincorrectness comparisons and one-sided differences by suggestions' terminology. Boldface identifies the value that shows the highest quality improvement.

	Lay means					Medico-scientific means				
	without	with	degrees of freedom	t statistic	p value	without	with	degrees of freedom	t statistic	p value
Correctness										
Terms?	0.45	0.45	292.1	0.02	0.49	0.42	**0.48**	310.3	-1.3	0.09
All terms?	0.43	**0.5**	176.7	-1.7	0.05*	0.42	**0.51**	244.8	-2.3	0.01*
NewTerms?	0.44	**0.49**	163.2	-1.2	0.12	0.42	**0.52**	197.2	-2.2	0.01*
Click?	0.43	**0.54**	87.9	-2.3	0.01*	0.43	**0.51**	123.3	-1.8	0.04*
Incorrectness										
Terms?	0.03	**-0.05**	293.8	2.3	0.01*	0.03	**-0.05**	316.9	2.1	0.01*
All terms?	0	**-0.02**	172.0	0.7	0.2	0.01	**-0.04**	243.4	1.5	0.07
NewTerms?	0.01	**-0.04**	144.7	1.1	0.13	0.01	**-0.03**	199.7	0.8	0.2
Click?	**-0.01**	0	78.7	-0.2	0.4	0	**-0.02**	129.2	0.6	0.3

Table 14.13: Δincorrectness comparisons and one-sided significant differences by language and English proficiency. Boldface identifies the value showing the highest quality improvement.

	Means		Differences	
	w/o	w/	test value	p
Terms?				
ep1/PT	0.03	**-0.1**	t(125.0)=2.3	0.01*
ep3/PT	0.1	**-0.01**	t(36.9)=1.8	0.04*
ep3/EN	0.11	**-0.02**	t(50.6)=2.0	0.02*
All terms?				
ep1/PT	0.02	**-0.12**	t(97.1)=2.3	0.01*
ep3/PT	0.08	**-0.02**	t(52.3)=1.9	0.03*
ep3/EN	0.09	**-0.03**	t(49.2)=1.9	0.03*
NewTerms?				
ep1/PT	0.01	**-0.12**	t(63.5)=2.0	0.03*
ep3/PT	0.08	**-0.02**	t(50.2)=1.9	0.03*
ep3/EN	0.1	**-0.02**	t(53.8)=2.0	0.02*
Click?				
ep3/PT	0.07	**-0.02**	t(51.5)=1.9	0.03*
ep3/EN	0.09	**-0.03**	t(39.4)=1.8	0.04*

Table 14.14: Tukey's adjusted p-value for one-sided significant comparisons of the *proficient* group with the other groups in terms of Δcorrectness and Δincorrectness.

	Terms?	All terms?	NewTerms?	Click?
ep3 >	ep1|ep2	ep|ep2	ep1|ep2	ep1|ep2
Correct.				
w/o PT	0.04|0.01	0.02|0.001	0.01|0.003	0.008|0.003
w/ PT	0.02|0.01	0.04|0.04	-|0.03	-|0.02
w/o EN	0.008|0.001	0.009|0.02	0.02|0.004	0.006|0.002
Incorrec.				
w/o EN	0.04|-	0.03| -	0.04| -	0.03|-

the terms from lay suggestions and also when *functional* literate users click on lay suggestions. In terms of incorrect contents we found that the *marginal* health literacy group benefits from the use of lay or medico-scientific suggestions. Contrary to what we expected, *functional* health literate users give answers with more incorrect contents when using terms, new or not, from medico-scientific suggestions. The same happens when these users click on lay suggestions.

Regarding the correctness differences between groups of health literacy, shown in Table 14.16, we found that, when not using suggestions, *low* health literate users give answers with less correct contents than *marginal* health lit-

Table 14.15: Δcorrectness [Cor] and Δincorrectness [Inco] comparisons and one-sided significant differences by terminology (Lay/MS) and health literacy (hl1, hl2, hl3). Boldface identifies the value that shows the highest quality improvement.

	Means		Differences	
	w/o	w/	test value	p
Terms?				
hl2/Lay [Inco]	0.03	**-0.1**	t(215.9)=3.2	0.0007**
hl2/MS [Inco]	0.04	**-0.1**	t(220.2)=3.3	0.0006**
hl3/MS [Cor]	0.38	**0.68**	t(37.5)=-2.5	0.01**
hl3/MS [Inco]	**0.01**	0.19	t(34.4)=-1.7	0.05*
All terms?				
hl1/Lay [Cor]	0.26	**0.44**	t(36.8)=-1.8	0.04*
hl2/Lay [Inco]	-0.01	**-0.08**	t(127.7)=1.7	0.04*
hl2/MS [Inco]	0.01	**-0.09**	t(176.5)=2.3	0.01**
hl3/MS [Cor]	0.43	**0.65**	t(34.8)=-1.7	0.05*
NewTerms?				
hl2/Lay [Inco]	0	**-0.12**	t(91.5)=2.6	0.005**
hl2/MS [Inco]	0	**-0.1**	t(138.2)=2.3	0.01*
hl3/MS [Cor]	0.42	**0.66**	t(37.4)=-1.9	0.03*
hl3/MS [Inco]	**0.01**	0.22	t(27)=-1.9	0.03*
Click?				
hl2/Lay [Cor]	0.45	**0.58**	t(43.3)=-1.9	0.03*
hl2/MS [Inco]	0	**-0.11**	t(79.8)=2.3	0.01*
hl3/Lay [Inco]	**0.02**	0.26	t(13.5)=-1.8	0.05*

eracy users. The same happens when *low* literate users do not use all the terms from lay suggestions or do not click them and are compared with *functional* users. Since we found that suggestions tend to increase the quantity of correct contents in almost all the scenarios and levels of health literacy, we conclude that suggestions help vanish the differences between the above groups of users, increasing the abilities of the *low* literate users to answer with a higher quantity of correct contents. When using medico-scientific suggestions, *functional* health literate users stand out from the *low* literate group with a larger quantity of correct contents. *Marginal* users also stand out from *low* literate users when they click in medico-scientific suggestions. When using terms from medico-scientific suggestions, *functional* health literate users also give more correct answers than *marginal* users (Tukey's adjusted p=0.032).

Surprisingly, we found that, with suggestions, the *functional* health literate group gives answers with more incorrect contents than the *marginal* literate group (Table 14.17). When using all the terms from lay suggestions, this is also true when *functional* users are compared with *low* literacy users.

The unexpected findings on incorrectness for the *functional* literate group using suggestions led us to further investigations. We believe the higher quantity of incorrect contents of these users when using suggestions and their worst

Table 14.16: Tukey's adjusted p-value for one-sided significant comparisons of the *low* health literacy group with the other groups in terms of Δcorrectness.

	Terms?	All terms?	NewTerms?	Click?
hl1 <	hl2|hl3	hl2|hl3	hl2|hl3	hl2|hl3
w/o Lay	0.012|-	0.005|0.022	0.027|-	0.08|0.03
w/o MS	0.008|-	0.025| -	0.04|-	-|-
w/ MS	-|0.015	-|0.035	-|0.022	0.04|0.02

Table 14.17: Tukey's adjusted p-value for one-sided significant comparisons of the *functional* health literacy group with the other groups in terms of Δincorrectness.

	Terms?	All terms?	NewTerms?	Click?
hl3 >	hl1|hl2	hl1|hl2	hl1|hl2	hl1|hl2
w/ Lay	-|0.001	0.005|0.002	-|0.001	-|0.003
w/ MS	-|3e-4	-|0.001	-|1e-4	-|0.001

results, when compared with users with lower levels of health literacy, is due to their longer answers. Although this may also be an argument to explain the higher quantity of correct answers, the effect of long answers in the correctness analysis is not as strong as its effect on the incorrectness analysis. In fact, the number of correct items in an answer has a predefined maximum while the number of incorrectness items has not. With this hypothesis in mind, we repeated the previous analysis for the incorrectness variable but now for the variation in the length of the answer. Similarly to what we found in terms of incorrectness, when *functional* users use medico-scientific terms from suggestions, their answers grow more than without these suggestions (w/: 843.1; w/o: 387.4; t (18.4)=-1.9, p= 0.04*). Comparing groups of users, we found several significant differences. Table 14.18 presents the p-values of the significant results between the *functional* group and the other groups. Besides the differences reported on Table 14.18, we also found that, with medico-scientific suggestions, the *low* literacy group gives answers with a smaller variation in length than the *marginal* group when using all terms (p=0.03*), new terms (p=0.03*) and when they click a suggestion (p=0.02*). These results corroborate our initial suspicion that the higher quantity of incorrect contents may be due to longer answers.

Analysis by topic familiarity

The significant results of the analysis by topic familiarity are presented in Table 14.19. These results show that *non-familiar* users give answers with a larger quantity of correct content when they use suggestions, either lay or medico-scientific ones. Surprisingly, the use of terms from lay suggestions by users *extremely familiar* with a topic decreases the correctness of their answers. However, these users benefit from lay suggestions with respect to incorrect contents, in the four scenarios. When using terms from medico-scientific suggestions, *extremely familiar* users also give answers with less incorrect contents.

Table 14.18: Tukey's adjusted p-value for one-sided significant comparisons of the *functional* health literacy group with the other groups in terms of answer length variation.

	Terms?	All terms?	NewTerms?	Click?
hl3 >	hl1|hl2	hl1|hl2	hl1|hl2	hl1|hl2
w/ Lay	1.2e-4|0.009	0.018|-	0.003|0.01	-|-
w/ MS	2.3e-5|1.4e-4	0.002|-	0.005|-	0.02|-
w/o Lay	-|-	-|-	-|-	0.02|-
w/o MS	-|-	-|-	0.04|0.04	0.03|0.02

Table 14.19: Δcorrectness [Cor] and Δincorrectness [Inco] comparisons and one-sided significant differences by terminology (Lay/MS) and topic familiarity (tf1, tf2, tf3). Boldface identifies the value that shows the highest quality improvement.

	Means		Differences	
	w/o	w/	test value	p
Terms?				
tf1/Lay [Cor]	0.38	**0.57**	t(55.2)=-2.2	0.015*
tf1/MS [Cor]	0.34	**0.56**	t(80.6)=-2.7	0.004**
tf3/Lay [Cor]	**0.46**	0.35	t(93.9)=1.7	0.045*
tf3/Lay [Inco]	0.04	**-0.14**	t(94.4)=3.2	0.001**
tf3/MS [Inco]	0.01	**-0.1**	t(100.2)=1.9	0.031*
All terms?				
tf1/Lay [Cor]	0.38	**0.7**	t(30.9)=-3.9	2.5e-04**
tf1/MS [Cor]	0.38	**0.55**	t(69.7)=-2.1	0.019*
tf3/Lay [Inco]	0.02	**-0.16**	t(57.2)=2.9	0.003**
NewTerms?				
tf1/Lay [Cor]	0.4	**0.62**	t(36)=-2.8	0.004*
tf1/MS [Cor]	0.38	**0.57**	t(67.1)=-2.3	0.012**
tf3/Lay [Inco]	0.01	**-0.15**	t(46.2)=2.3	0.012**
Click?				
tf1/Lay [Cor]	0.4	**0.69**	t(17.6)=-3.1	0.003*
tf1/MS [Cor]	0.39	**0.57**	t(41.9)=-1.9	0.03**
tf3/Lay [Inco]	0	**-0.19**	t(24.7)=2.5	0.01**

In the comparison of the several familiarity groups, whose significant results are presented in Table 14.20, we found that, when using terms from lay suggestions, the *extremely familiar* group gives answers with less correct content than *non-familiar* users and answers with less incorrect content than *familiar* users. Moreover, the use of all the terms from lay suggestions make *extremely familiar* users be the worse group in what concerns the quantity of correct contents and the best group in terms of incorrect contents. This happens because, in every scenario, the use of lay suggestions simultaneously tends to increase the Δcorrectness in *non-familiar* and *familiar* users and to have the

opposite effect on *extremely familiar* group. As stated above, the differences in the *non-familiar* group are significant. We also found that *extremely familiar* users behave better in terms of incorrect contents than users with less familiarity when clicking lay suggestions.

Table 14.20: Tukey's adjusted p-value for one-sided significant comparisons of the *extremely familiar* group with the other groups in terms of Δcorrectness and Δincorrectness.

	Terms?	All terms?	NewTerms?	Click?
tf3 <	tf1|tf2	tf1|tf2	tf1|tf2	tf1|tf2
Correct.				
w/ Lay	0.011|-	0.001|0.023	0.022|-	-|-
Incorrec.				
w/ Lay	-|0.042	0.04|0.004	-|0.04	0.04|0.006

14.3.4 Suggestions impact on motivational relevance

Motivational relevance was evaluated through users' feeling of success with each iteration. In terms of significant differences, we found that iterations where terms from suggestions were used had a smaller proportion of successes (w/ sug.: 65.6%; w/o sug.: 79.5%; $\chi^2(1)=13.1$, p=1e-4**). Similarly, the same happened in iterations where NewTerms from suggestions were used (w/ sug.: 64.2%; w/o sug.: 77.3%; $\chi^2(1)=6.7$, p=0.005**) and in iterations were suggestions were clicked (w/ sug.: 67.3%; w/o sug.: 76.8%; $\chi^2(1)=3.6$, p=0.03*). Although it is non-significant, the only situation where the proportion of successes is higher with the use of suggestions happens when users use entire suggestions that only contain unused terms (w/ sug.: 79%, w/o sug.: 74.8%).

The comparison between both retrieval systems shows a very similar proportion of successes in both systems (71% in SYS1 and 71.3% in SYS2), a non-significant difference.

The tendency to have more success without suggestions applies to all types of suggestions, that is, Portuguese, English, lay and medico-scientific suggestions. In these comparisons, two significant differences occur: in Portuguese suggestions clicks (w/ sug.: 61.0%, w/o sug.: 76.0%, $\chi^2(1)=6.0$, p=0.007**) and in the use of terms from lay suggestions (w/ sug.: 68.9%, w/o sug.: 77%, $\chi^2(1)=6.8$, p=0.004**). The only exception to this tendency happens when users employ entire English suggestions; in this case the proportion is 76.2% against 75.2% when users do not use them.

Analysis by English proficiency

An analysis by English proficiency revealed that the use of terms from Portuguese suggestions by the *independent* proficiency users result in an smaller proportion of succeeded iterations (w/o sug.: 82.6%, w/ sug.: 72.7%, $\chi^2(1)=6.0$, p= 0.007**) than without them. The same happens when these users employ new terms from Portuguese suggestions (w/o sug.: 80.9%, w/ sug.: 69.2%, $\chi^2(1)=3.3$, p= 0.04*) and when *basic* users click in Portuguese suggestions (w/o sug.: 70.0%, w/ sug.: 45.0%, $\chi^2(1)=4.6$, p= 0.016*).

Through the significant results obtained in the comparison of groups of English proficiency, available in Table 14.21, we conclude that, without suggestions, *basic* proficiency users feel less successful in their searches when compared with the other groups. When using all terms from Portuguese suggestions or new terms from these suggestions, the differences in relation with the *independent* group still apply.

Table 14.21: Tukey's adjusted p-value for one-sided significant comparisons of the *basic* English proficiency group with the other groups in terms of motivational relevance.

	Terms?	All terms?	NewTerms?	Click?
ep1<	ep2|ep3	ep2|ep3	ep2|ep3	ep2|ep3
w/o EN	0.0001|0.005	0.007|0.005	0.0002|0.015	0.005|0.004
w/o PT	1.5e-5|0.003	0.002|-	0.0002|-	0.0003|0.02
w/ PT	-|-	0.04|-	0.046|-	-|-

Analysis by health literacy

An analysis by health literacy only revealed that, when using terms from lay suggestions, *marginal* health literate users feel less successful than without them ($\chi^2(1)$=9.4, p = 0.002**). In terms of differences between groups of health literacy, when terms from lay suggestions are not used, *low* literate users are less successful than *marginal* literate users (Tukey's adjusted p=0.027*).

Analysis by topic familiarity

In the analysis by topic familiarity, we found that *non-familiar* ($\chi^2(1)$=4.8, p=0.014*) and *familiar* ($\chi^2(1) = 2.8$, p = 0.047) users feel more successful when they don't use terms from lay suggestions. Moreover, *non-familiar* users feel more successful with medico-scientific suggestions than with lay ones ($\chi^2(1)$=5.1, p=0.012*). Table 14.22 shows the significant differences between groups of topic familiarity.

Table 14.22: Tukey's adjusted p-value for one-sided significant comparisons of the *extremely familiar* group with the other groups in terms of motivational relevance.

	Terms?	All terms?	NewTerms?	Click?
	tf1|tf2	tf1|tf2	tf1|tf2	tf1|tf2
tf3>				
w/o Lay	4e-5|8e-3	1.5e-6|8e-4	1e-6|3e-3	4e-7|5e-4
w/ Lay	6e-4|2e-2	3e-2|-	4e-2|4e-2	-|-
w/o MS	4e-8|3e-3	1e-7|5e-3	1e-7|9e-3	3e-7|2e-4
tf2>				
w/o Lay	-|-	-|-	3e-2|-	4e-2|-
w/o MS	4e-3|-	2e-2|-	1.5e-2|-	-|-

We found that, generally, without using terms from medico-scientific suggestions, users feel more successful as their familiarity with the topic increases. The same happens when new terms from lay suggestions are not used and when lay suggestions are not clicked. With the use of terms from lay suggestions, new or not, the *extremely familiar* group surpasses the others in proportion of successes. With the use of all the terms from lay suggestions, the *extremely familiar* group only surpasses the *non-familiar* group.

14.4 CONCLUSION

In this chapter we describe the methodology followed to answer our research questions. Moreover, we present and analyze the results we found. In Chapter 15 we discuss the results presented in this chapter and describe our main conclusions.

DISCUSSION OF RESULTS

15.1 INTRODUCTION

In the previous chapter we described the user experiment we have conducted to evaluate the suggestion prototype. The results gathered during the experiment were also presented and analyzed in that chapter. In this chapter, we discuss these results in the context of this dissertation. We focus on results that relate to the research questions presented in the previous chapter. Each one of the 5 main research questions is discussed in a separate section. In the end we summarize our discussion and examine the implications of our findings.

15.2 SUGGESTIONS' USE BEHAVIOR

One of our research questions is concerned with suggestions' usage behavior. Regarding this aspect, we found that suggestions are used more often in the beginning of the search sessions. In fact, in the initial iterations, users not only click and use entire suggestions more often but they also use more terms from suggestions. From the sessions where suggestions were presented, almost 55% had at least one click and almost 87% had a query in which at least one of the terms was extracted from one of the suggestions. From the 40 participants involved in the study, only 5 did not click in any suggestion during their tasks. This shows a good acceptance rate of the suggestions, similar to what Jansen and McNeese (2005) have found.

The suggestions using Portuguese and lay terminology are the ones with a smaller proportion of clicks and the ones from where new terms are extracted in a lower quantity. This shows that these suggestions are the least useful, which is not strange since queries formulated by Portuguese lay people without assistance will most probably use Portuguese and lay terminology. To support this statement we verified that all the queries formulated in the first iteration, and therefore without assistance, were Portuguese and used lay terminology. Ignoring the terms used in previous queries, both types of Portuguese suggestions are preferred to both types of English suggestions. Yet, if we deepen our analysis considering users' English proficiency, we see that this only happens in the lower levels of English proficiency. Advanced proficiency users tend to prefer terms from English suggestions. Moreover, along with the *basic* proficiency users, *proficient* users click more often and extract more new terms from English suggestions than from Portuguese ones. These last findings about *basic* proficiency users surprised us because we thought a language

in which they are not proficient wouldn't attract them. However, this probably happened because these suggestions had a great degree of novelty and this might have aroused their curiosity. As expected, *proficient* users click more often and use more new terms from English suggestions than *independent* users.

In general, users prefer medico-scientific suggestions to lay ones. However, if we consider users' health literacy, we see that this is only true in the *marginal* and *functional* groups. We also found that the use of terms from medico-scientific suggestions significantly increases with the health literacy of the users. Moreover, the *functional* group has a larger proportion of clicks in medico-scientific suggestions when compared with the *low* and *marginal* group.

Topic familiarity was not found to be discriminative in terms of lay versus medico-scientific terminology use.

15.3 COMPARISON OF THE RETRIEVAL SYSTEMS

The system with suggestions (SYS1) has a better performance in terms of precision, correctness and incorrectness of the answers. Yet, the only significant difference was found in terms of incorrect contents where SYS1 reduced the quantity of incorrect contents. Users were slightly more satisfied with the system that did not present suggestions but this is a very small difference. Through these findings we can conclude that a retrieval system that includes the proposed suggestion tool contributes to a better retrieval experience in an outcome that is particularly important in health searches, namely helping to reduce the quantity of incorrect contents in the acquired knowledge.

15.4 IMPACT OF SUGGESTIONS' CLICKS

On Table 15.1 we aggregate the significant findings reported in the previous chapter about the use of suggestions, in general, by language and by terminology. As can be seen in this table, clicking in suggestions, regardless of its type, leads to answers with more correct content and less motivational relevance than not clicking suggestions. A deeper analysis shows that only Portuguese suggestions don't show a significant benefit to answers' correctness. Moreover this is also the only type of suggestion that degrades motivational relevance. English suggestions are advantageous to *proficient* users in both precision and incorrectness of the answers. Surprisingly, *basic* English proficiency user give more correct contents with English suggestions than without them. Even not having much English proficiency, these users are still capable of extracting accurate information from English documents. Although English suggestions tend to increase the incorrectness outcome of *basic* proficiency users, we did not find this to be significant. Since these users are not significantly affected in any retrieval outcome by this type of suggestions, we hypothesize that, although English queries had a worse behavior than Portuguese queries in the experiment described in Chapter 9, it might be better to use them at least once than not to use them at all.

Excluding the unexpected increase of incorrect contents in the *functional* health literate group with lay suggestions, the use of lay and medico-scientific suggestions has a good effect on precision and medical accuracy. Surprisingly,

Table 15.1: Summary of the significant findings pertaining click suggestion. ↑ denotes increases and ↓ decreases in each outcome.

	Precision	**Correctness**	**Incorrectness**	**Mot. relevance**
General		↑		↓
English	proficient EP (↑)	general (↑) basic EP (↑)	proficient EP (↓)	
Portuguese			proficient EP (↓)	general (↓) basic EP (↓)
Lay	non-familiar (↑) extremely familiar (↑)	general (↑) functional HL (↑)	functional HL (↑) extremely familiar (↓)	
Medico-scientific	low HL(↑) non-familiar (↑) extremely familiar (↑)	general (↑) non-familiar (↑)	marginal HL (↓)	

lay suggestions increase precision not only in *non-familiar* users but also in the *extremely familiar* group. This seems to indicate that, although in Chapter 10 we concluded that medico-scientific queries tend to have higher precision than lay queries when users are more familiar with the topic, these users also benefit from lay suggestions. This happens not only because they have higher precision when clicking lay suggestions but also because they give answers with less incorrect contents in this situation. We cannot clearly state that lay queries are beneficial to *functional* health literate users because they show an increase in both correct and incorrect contents. Although the length of their answers probably explains this, the presence of more incorrect contents cannot be ignored and the use of lay queries in *functional* health literate users has to be further explored. The precision increase found when *low* health literate users click on medico-scientific suggestions is consistent with what has been found in a previous study, that is, "less subject expertise seems to lead to more lenient and relatively higher relevance ratings" (Saracevic, 2007b). This means that these users may be assessing documents regarding their relation with the topic instead of their utility to themselves. Precision findings in *non-familiar* users could be explained the same way but, since we also found that medico-scientific suggestions increase their answers' correct contents and tends to decrease their incorrect contents, we have reasons to believe this is not the case. Moreover, this agrees with what we found in the study described in Chapter 10, where we concluded that health literacy is more important to comprehend medico-scientific documents than topic familiarity.

In Table 15.2 we present the significant differences found when comparing groups of users, clicking or not in certain type of suggestions, in the four outcomes of the retrieval process. In general, groups with higher English proficiency, health literacy or topic familiarity have a better retrieval experience than users below their level in terms of acquired correct contents and motivational relevance.

Table 15.2: Summary of the significant differences found in groups' comparisons pertaining click suggestion.

		Prec.	Correctness	Incorrectness	Mot. relevance
EN	**w/o**		EP_3>{EP_1, EP_2}	EP_3>EP_1	{EP_3,EP_2}>EP_1
	w/				
PT	**w/o**		EP_3>{EP_1, EP_2}		{EP_3,EP_2}>EP_1
	w/		EP_3>EP_2		
Lay	**w/o**		{HL_3, HL_2}>HL_1		TF_3>{TF_2, TF_1}
					TF_2>TF_1
	w/			HL_3>HL_2	
				TF_3<{TF_2, TF_1}	
MS	**w/o**				TF_3>{TF_2, TF_1}
	w/		{HL_3, HL_2}>HL_1	HL_3>HL_2	

As can be seen, when not clicking suggestions, *proficient* users give answers with more correct contents than *basic* and *independent* users. This shows

that, in a retrieval system without suggestions, these users are better prepared to search for health information and/or to answer medical questions. However, when using English suggestions, the *proficient* users' superiority is no longer significant. In addition, when using Portuguese suggestions, the *proficient* group superiority only stands when they are compared with the *independent* proficiency group. Since, with the use of suggestions, the quantity of correct contents increases in every group of users, we hypothesize that English suggestions contribute to fade the differences between groups of users. Regarding the incorrectness outcome, since the use of English suggestions tends to simultaneously diminish the incorrect contents in the *proficient* group and increase them in the *basic* proficiency group, this type of suggestions contributes to eliminate the difference that exists between these groups when they do not click English suggestions.

When not clicking lay suggestions, *low* health literate users give answers with less correct contents than *marginal* and *functional* users, a difference that does not exist when these users click in this type of suggestions. Since we found that clicking in lay suggestions tends to increase the quantity of correct contents in all levels of health literacy, we conclude that these suggestions help vanish the differences between the above groups of users, increasing the abilities of the *low* literacy users to answer with a higher quantity of correct contents. On the other hand, although clicking in medico-scientific suggestions tends to increase the quantity of correct contents in all levels of health literacy, the rise is bigger in the higher levels of health literacy making a significant difference emerge between these groups. This corroborates a finding reported in Chapter 10 showing that users with higher levels of health literacy are more apt to assimilate medico-scientific documents. Probably caused by the length of their answers, the *functional* health literate group behaved worse than the *marginal* literate group in terms of incorrect contents when using suggestions. In terms of motivational relevance, the *basic* English proficiency group is the least satisfied with their search tasks when suggestions are not clicked. On the other hand, when suggestions are not used, the greater the familiarity with the topic, the greater the satisfaction with the task. This confirms a tendency we already found in the study of Chapter 9.

15.5 IMPACT OF SUGGESTIONS' TERMS

One of our research questions is related to a different use given to suggestions, that is, to use them as a source of terms without clicking them. In Tables 15.3, 15.4 and 15.5 we summarize the significant findings related to the three scenarios of query expansion we have considered: use of terms from suggestions, use of all suggestions' terms and use of new terms from suggestions. A global analysis of these three tables shows us that the three scenarios of query expansion are prejudicial to precision and motivational relevance but beneficial to the medical accuracy of the acquired knowledge, simultaneously increasing the quantity of correct contents and decreasing the quantity of incorrect contents.

The most favorable scenario of term expansion involves the use of all the terms of a suggestion where only significant advantages were found. This is the scenario closest to the click scenario. When using all their terms, Portuguese suggestions contribute to decrease the quantity of incorrect contents

Table 15.3: Summary of the significant findings pertaining the use of terms from suggestions (Terms). ↑ denote increases and ↓ decreases in each outcome.

	Precision	Correctness	Incorrectness	Mot. relevance
General			↓	↓
English			proficient EP (↓)	
Portuguese			general (↓) basic EP (↓) proficient EP (↓)	independent EP (↓)
Lay		non-familiar (↑) extremely familiar (↓)	general (↓) marginal HL (↓) extremely familiar (↓)	general (↓) marginal HL (↓) non-familiar (↓) familiar (↓)
Medico-scientific		functional HL (↑) non-familiar (↑)	general (↓) marginal HL (↓) functional HL (↑) extremely familiar (↓)	

Table 15.4: Summary of the significant findings pertaining the use of all suggestions' terms (All Terms). ↑ denote increases and ↓ decreases in each outcome.

	Prec.	Correctness	Incorrectness	Mot. relev.
General		↑	↓	
EN		general (↑) basic EP (↑)	proficient EP (↓)	
PT			general (↓) basic EP (↓) proficient EP (↓)	
Lay		general (↑) low HL (↑) non-familiar (↑)	marginal HL (↓) extremely familiar (↓)	
MS		general (↑) functional HL (↑) non-familiar (↑)	marginal HL (↓)	

and English, lay and medico-scientific suggestions to increase the quantity of correct contents. Similarly to what happens when *basic* English proficiency users click in English suggestions, when they use all their terms, these users' answers are improved in terms of correct contents. The two groups with lower health literacy benefit from using all terms of lay suggestions, the *low* literate users increase their correct contents and the *marginal* literate users decrease their incorrect contents. On the other hand, the two groups with highest literacy benefit from using all terms from medico-scientific suggestions, the *functional* group increases the correct contents of its answers and the *marginal* group decreases its incorrect contents.

The use of suggestions, in both languages and every scenario of term expansion, tends to increase the quantity of correct contents. In terms of significant differences they are only visible when users employ all or new terms from English suggestions. On the other hand, in general, Portuguese suggestions tend to decrease the quantity of incorrect contents, being only significant when users employ terms or all terms from these suggestions. English suggestions tend to have the same effect but only in the *independent* and *proficient* users. The only significant difference is found when *proficient* users use terms from English suggestions. About language comparisons, during the medical accuracy analysis we found that 8 answers had incorrect content due to a commonly misunderstood false friend involving the words *constipation* in English and *constipação*, in Portuguese, where the latter means a cold. At first we supposed this error was due to users' translation but, later, we found one retrieved document with this exact error. This reinforces the importance of high-quality content in health searches, where the inclusion of medical certification in the set of criteria used by search engines, according to the conclusions of Chapter 6, or the use of English queries, according to the conclusions of Chapter 9, may help.

Table 15.5: Summary of the significant findings pertaining the use of new terms from suggestions (New Terms). ↑ denote increases and ↓ decreases in each outcome.

	Precision	**Correctness**	**Incorrectness**	**Mot. relevance**
General				↓
English		general (↑)	proficient EP (↓)	
Portuguese	general (↓)		basic EP (↓) proficient EP (↓)	independent EP (↓)
Lay	general (↓) marginal HL (↓) functional HL (↓) familiar (↓)	non-familiar (↑)	marginal HL (↓) extremely familiar (↓)	
Medico-scientific	familiar (↓)	general (↑) functional HL (↑) non-familiar (↑)	marginal HL (↓) functional HL (↑)	

In every scenario, the *functional* health literate group significantly gives answers with more correct contents when using medico-scientific suggestions. In the use of terms and new terms from these suggestions, this group of users also gives answers with more incorrect contents. The *marginal* health literate group gives answers with less incorrect contents using lay and medico-scientific suggestions in all the scenarios of term expansion.

In terms of familiarity, in all three scenarios, *non-familiar* users increase the correct contents in their answers when using lay or medico-scientific suggestions. Moreover, *extremely familiar* users decrease their incorrect contents with the use of lay suggestions in the three scenarios and with the use of terms from medico-scientific suggestions. Comparably to click suggestions, we conclude that lay suggestions are useful to users of every familiarity level performing query expansion. Likewise, *non-familiar* users are also able to digest medico-scientific contents and benefit from this type of suggestions.

Regarding precision, we only found significant differences in the use of new terms from suggestions where, in general, precision decreases with the use of Portuguese and lay suggestions.

Regarding motivational relevance, generally, users tend to be less satisfied when they use terms or new terms from suggestions. In general, lay and Portuguese suggestions result in less satisfied users. Since these results contradict the medical accuracy results, we conjecture that users may feel unsatisfied after using suggestions because they felt the need to use them, that is, they only used them because they were unsatisfied and that feeling has not disappeared after the search task. The fact that *non-familiar* users feel more successful when they use terms from medico-scientific suggestions than from lay ones reinforces our previous conclusion that *non-familiar* users benefit from this type of suggestions.

Tables 15.6, 15.7 and 15.8 present a summary of the significant findings encountered in the three scenarios of term expansion when comparing groups of users. It is clear that users feel more satisfied when they have more English proficiency and topic familiarity. Excluding some cases related to the use of lay suggestions, the same happens in terms of the correctness of the acquired knowledge with the English proficiency, topic familiarity and health literacy.

Without suggestions, *proficient* users give answers with more correct contents than users with less proficiency. While, with Portuguese suggestions, the above differences still exist, with English suggestions they disappear. Since we also found that English suggestions tend to improve the increase in correctness in all levels of English proficiency, we conclude that these suggestions contribute to approximate users with less proficiency to more proficient users.

Without English suggestions, *proficient* users were found to give answers with more incorrect content than *basic* users. This can be, in part, explained by the fact that pre-search answers of *basic* users had significantly more incorrect contents than *proficient* users, making it easier for them to reduce that quantity after a search session. With English suggestions, this difference disappears because, in this situation, *basic* users give answers with the same or more incorrect contents than the answers given before the task while, without English suggestions, their post-search answers have less incorrect contents than before it. Moreover, the opposite happens in *independent* or *proficient* users.

Without suggestions, the *marginal* health literate group gives answers with more correct content than *low* literate users. In general, lay suggestions in-

crease the correctness of users' answers but its effect is stronger in *low* health literate users, thus eliminating the above difference. We therefore conclude that lay suggestions help to increase the abilities of the *low* literacy users to answer with a higher quantity of correct contents even when using partial suggestions (when using all terms from suggestions, this effect was already found to be significant). On the other hand, medico-scientific suggestions have an increased benefit in the correctness of the *functional* literate users and contribute to differentiate these from the users with lowest health literacy. Surprisingly, with suggestions, the *functional* literate group has answers with more incorrect contents than the *marginal* group. In part, this can be explained by their longer answers.

When partially using lay suggestions, *extremely familiar* users perform worse than *non-familiar* users in the correctness outcome but perform better than *familiar* users in the incorrectness aspect. When using all terms from lay suggestions the *extremely familiar* group becomes the worst group in correctness and the best group in incorrectness. This can be explained by the significant benefit of lay suggestions in the correctness of *non-familiar* users and in the incorrectness of *extremely familiar* users. The precision difference between *familiar* and *non-familiar* users when new terms from medico-scientific suggestions are used is explained by the prejudicial effect of this scenario and type of suggestions to *familiar* users and its advantage to *non-familiar* users that, in these circumstances, have higher precision than the other familiarity groups.

Table 15.6: Summary of the significant differences found in groups' comparisons pertaining the use of terms from suggestions (Terms).

		Prec.	Correctness	Incorrectness	Mot. relevance
EN	**w/o**		EP_3>{EP_2, EP_1}	EP_3>EP_1	{EP_3,EP_2}>EP_1
	w/				
PT	**w/o**		EP_3>{EP_2, EP_1}		{EP_3,EP_2}>EP_1
	w/		EP_3>{EP_2, EP_1}		
Lay	**w/o**		HL_2>HL_1		HL_2>HL_1
					TF_3>{TF_2, TF_1}
	w/		TF_3<TF_1	HL_3>HL_2	TF_3>{TF_2, TF_1}
				TF_3<TF_2	
MS	**w/o**		HL_2>HL_1		TF_3>{TF_2, TF_1}
					TF_2>TF_1
	w/		HL_3>{HL_2, HL_1}	HL_3>HL_2	

15.6 PERSONALIZATION STRATEGIES

In the click and use of all terms scenarios, we found that *basic* English proficiency users not only significantly prefer English to Portuguese suggestions but they also significantly benefit, in correctness, from the former type of sug-

Table 15.7: Summary of the significant differences found in groups' comparisons pertaining the use of all the terms of a suggestion (All Terms).

		Prec.	Correctness	Incorrectness	Mot. relevance
EN	**w/o**		EP_3>{EP_2, EP_1}	EP_3>EP_1	{EP_3,EP_2}>EP_1
	w/				
PT	**w/o**		EP_3>{EP_2, EP_1}		EP_2>EP_1
	w/		EP_3>{EP_2, EP_1}		EP_2>EP_1
Lay	**w/o**		{HL_3, HL_2}>HL_1		TF_3>{TF_2, TF_1}
	w/		TF_3<{TF_2, TF_1}	HL_3>{HL_2, HL_1}	TF_3>TF_1
				TF_3<{TF_2, TF_1}	
MS	**w/o**		HL_2>HL_1		TF_3>{TF_2, TF_1}
					TF_2>TF_1
	w/		HL_3>HL_1	HL_3>HL_2	

Table 15.8: Summary of the significant differences found in groups' comparisons pertaining the use of new terms of a suggestion (New Terms).

		Precision	Correctness	Incorrectness	Mot. relevance
EN	**w/o**		EP_3>{EP_2, EP_1}	EP_3>EP_1	{EP_3,EP_2}>EP_1
	w/				
PT	**w/o**		EP_3>{EP_2, EP_1}		EP_2>EP_1
	w/		EP_3>EP_2		EP_2>EP_1
Lay	**w/o**		HL_2>HL_1		TF_3>{TF_2, TF_1}
					TF_2>TF_1
	w/		TF_3<TF_1	HL_3>HL_2	TF_3>{TF_2,TF_1}
				TF_3<TF_2	
MS	**w/o**		HL_2>HL_1		TF_3>{TF_2, TF_1}
					TF_2>TF_1
	w/	TF_2<TF_1	HL_3>HL_1	HL_3>HL_2	

gestions. Since we did not find any significant disadvantage in the use of English suggestions, we conclude that, although Portuguese suggestions may be preferable to English suggestions to these users, a conclusion of the study reported in Chapter 9, it is better to have English suggestions than to have no suggestions at all. Yet, we suspect the choices of *basic* English proficiency users are not the ideal. For example, when clicking Portuguese suggestions, we did not find a significant difference in the correctness outcome because few users chose this type of suggestions. Moreover, similarly to what we found in Chapter 9, in *basic* English proficiency users, we found that Portuguese suggestions tend to be more effective in reducing answers' incorrectness than English suggestions. This has to be further explored.

Since *independent* proficiency users tend to have better correctness and incorrectness performance without clicking Portuguese suggestions, we stick with the conclusions described in Chapter 9 and think these users should only be provided with English suggestions. Since this is the only group that does not significantly use English suggestions more often than Portuguese ones, we think the proposal of solely English suggestions may be giving them access to higher-quality health content, as stated in Chapter 9.

Advanced proficiency users benefit from both types of suggestions in terms of incorrectness. Since the decrease in incorrect contents is higher with English suggestions than with Portuguese suggestions, we remain with the conclusion reached in earlier studies, that is, these users should only be presented with English suggestions.

Lay suggestions tend to be advantageous to *low* health literate users and differences are significant in terms of correct contents when they use all the suggestions' terms. We therefore conclude these users, along with *marginal* health literate ones, should be provided this type of suggestions. Note that the choices of the users are aligned with our conclusions, being the *low* literate group the only group that doesn't use medico-scientific suggestions more often than lay ones. Since *marginal* health literate users have significantly less incorrect contents with lay and medical accuracy suggestions, we think they should be provided both types of suggestions. *Functional* health literate users had contradictory behaviors in terms of correctness and incorrectness when clicking lay suggestions and using terms from medico-scientific suggestions. Considering the conclusions of Chapter 10 and the absence of significant bad results in the clicking scenario, we think these users should be provided medico-scientific suggestions. Doubts exist on whether or not to suggest alternative lay queries to these users.

Our results indicate that lay and medico-scientific suggestions are useful to users of all topic familiarities. In fact, this is a characteristic that is not discriminant to the personalization of the suggestion system. In Table 15.9 we summarize the personalization conclusions we have reached so far.

Table 15.9: Personalization pertaining English proficiency and health literacy.

	EP1	**EP2**	**EP3**		**HL1**	**HL2**	**HL3**
EN	?	√	√	**Lay**	√	√	?
PT	√	-	-	**MS**	-	√	√

To evaluate the value of the proposed personalization strategy and to study whether or not to present English suggestions to *basic* proficiency users and whether or not to present lay suggestions to *functional* health literate users, we decided to compare both systems in five personalization scenarios:

- Scenario 1 – No personalization at all – the system presents all the suggestions to every user;

- Scenario 2 – Table 15.9 scenario **with** English suggestions to EP1 and **with** lay suggestions to HL3;

- Scenario 3 – Table 15.9 scenario **with** English suggestions to EP1 and **without** lay suggestions to HL3;
- Scenario 4 – Table 15.9 scenario **without** English suggestions to EP1 and **with** lay suggestions to HL3;
- Scenario 5 – Table 15.9 scenario **without** English suggestions to EP1 and **without** lay suggestions to HL3;

.

To simulate the scenarios 2-5 we restricted the set of SYS1 iterations to the ones in which there was a click in one of the suggestions that would be presented to each particular user, according to the scenario being considered. For example, in the second and third scenarios, for a *basic* proficiency user (EP1) with *low* health literacy (HL1), we only consider two types of suggestions: Lay/English and Lay/Portuguese. For the same user, in the fourth and fifth scenarios, only the Lay/Portuguese suggestion is considered.

Systems were compared in four outcomes: precision, correctness, incorrectness and motivational relevance. In all the scenarios we found the same tendencies, that is, SYS1 tends to have higher precision, higher correctness, less incorrectness and less motivational relevance. However, as shown in Table 15.10, we only found significant differences in terms of correctness and incorrectness. In the second and third scenarios of personalization, the incorrectness significant difference that existed in the first scenario disappears and emerges a significant difference in terms of correctness with a lower p-value, that is a lower probability of obtaining a test statistic as extreme as the observed one, assuming the null hypothesis. From Table 15.10 it is possible to conclude that the best personalization scenarios are the fourth and fifth ones where we simultaneously found significant differences in correctness and incorrectness. From these two scenarios, we consider the fifth scenario the best one for having a lower p-value in terms of precision and incorrectness. It is important to note that, in the health domain, we consider more important to lower the quantity of incorrect contents than to rise the quantity of correct contents due to the possible damages that incorrect contents can cause.

From these findings we can conclude that the best personalization strategy is the one presented in Table 15.11 and that the awareness of users' English proficiency and health literacy is advantageous to health information retrieval, particularly to the medical accuracy of the knowledge obtained in the search session. We believe that, in the implementation of a personalized suggestion tool in a production system, where the user will be biased toward the suggestions more appropriate for him, more significant findings will probably be encountered.

15.7 CONCLUSION

This study had multiple goals. We intended to evaluate a prototype that, given a health query, suggests alternative queries in two languages, Portuguese and English, using two types of terminologies, lay and medico-scientific. Moreover, we wanted to assess its acceptance within the users of a search system and to analyze if and how an awareness of users' English proficiency, health

Table 15.10: Means and significant differences between systems in each personalization scenario.

			Sys 1 vs Sys2	
	Sys1	Sys2	test value	p value
Scenario 1				
Incorrectness	**-0.04**	0.03	t(312) = -1.7	0.045*
Scenario 2				
Correctness	**0.56**	0.43	t(141.4)=2.6	0.005**
Scenario 3				
Correctness	**0.56**	0.43	t(128.2)=2.6	0.007**
Scenario 4				
Correctness	**0.56**	0.43	t(84.9)=2.3	0.012*
Incorrectness	**-0.06**	0.03	t(88.6) = -1.7	0.049*
Scenario 5				
Correctness	**0.56**	0.43	t(76.3)=2.1	0.021*
Incorrectness	**-0.07**	0.03	t(80) = -1.8	0.036*

Table 15.11: Personalization strategy to be implemented.

	EP1	**EP2**	**EP3**		**HL1**	**HL2**	**HL3**
EN	-	√	√	**Lay**	√	√	-
PT	√	-	-	**MS**	-	√	√

literacy and topic familiarity might be useful to personalize such system. The evaluation of the suggestion tool was done considering the utility of the suggestions as whole suggestions and as sources of terms to be included in new queries. The personalization analysis complements the conclusions of the previous studies.

Suggestions were found to be used more often and, when contributing to term expansion, also in a larger quantity, in the initial stages of a search session. In general, suggestions had a good acceptance by the users and the novelty aspect seems to be important in the choice of which suggestion to use. Excluding the scenario in which terms from suggestions are used, useful to assess the quality of suggestions' terms but not so useful to assess their utility to the users, English suggestions tend to be preferred to Portuguese ones in all levels of English proficiency, a significant preference in the *basic* and *proficient* users. On the other hand, medico-scientific suggestions tend to be preferred to lay ones in the higher levels of health literacy and the extraction of new terms from these suggestions increases with health literacy.

A retrieval system that includes the implemented suggestion system without any type of personalization tends to be better than a system without suggestions in terms of precision, correctness and incorrectness and tends to be slightly worse in terms of motivational relevance. Of these, only the incorrectness difference is significant, an outcome extremely relevant in the health

domain. The best behavior of the suggestion tool is achieved when users use it strictly as a suggestion tool, that is, when they click in suggestions, and when they use all terms from suggestions. However the medical accuracy of the obtained knowledge can also be improved from suggestions as sources of terms.

Our findings allowed us to slightly update the personalization conclusions we have reached in previous studies. A comparison of several personalization scenarios allowed us to conclude that a retrieval system that proposes suggestions considering users' English proficiency and health literacy outperforms a system without personalization. This personalized system bias users towards the suggestions more beneficial to them. The best personalization strategy involves presenting English suggestions two the higher levels of English proficiency and Portuguese suggestions to *basic* proficiency users. Moreover, lay suggestions should be provided to the lowest levels of health literacy and medico-scientific ones to the highest levels.

We proved that the personalized suggestion of medical concepts related to the initial query using different languages and terminologies could be beneficial to the medical accuracy of the knowledge obtained in the search session.

The next chapter is included in a new part of this dissertation and regards the automatic acquisition of context features. The next part's single chapter reports preliminary work on the automatic identification and categorization of health queries. We propose several methods and compare them with existing ones. This kind of methods allows retrieval systems to detect when they are dealing with health search sessions and, in some methods, to even gather more specific information, such as the semantic types of queries.

PART V

AUTOMATIC CONTEXT ACQUISITION

HEALTH QUERIES IDENTIFICATION

16.1 INTRODUCTION

Before implementing personalized HIR strategies, retrieval systems have to be aware that the user is performing a health search through the query he submits. Therefore, the ability to correctly identify health-related queries is important. By health query we mean a query that intends to retrieve health-related information and is motivated by a health information need.

Some research works, like the one from Spink et al. (2004), manually classify queries, a process that is slow and requires the availability of one or more human classifiers. In some cases, the huge volume of queries may even make this classification impracticable. For these reasons, automatic methods to perform the query classification task can be useful.

Some years ago, Eysenbach and Kohler (2003) proposed a method to automatically classify search strings as health-related based on the proportion of pages on the Web containing the search string plus the word "health" and the number of pages containing only the search string. Besides this method, no other automatic mechanism with this goal was found reported in the literature. The nearest, but broader, topic is generic automatic query classification. An extensive state of the art about this topic is done by Beitzel et al. (2005). Yet, as our goal is restricted to the health domain, we believe some simpler and more targeted strategies may be developed.

Our goal with this research is to propose new automatic methods to detect health queries and to compare them with three variants of the one described by Eysenbach and Kohler (2003). Based on the knowledge that most health queries contain terms that can be mapped to health/medical vocabularies (McCray et al., 1999; Zeng et al., 2006), we have decided to use this type of vocabularies to detect the presence of health terms in queries through several different strategies.

16.2 METHODS

We evaluate several automatic methods to detect health-related queries that can be grouped in three distinct categories. One category, entitled *co-occurrence methods* contains 3 variants of the method proposed by Eysenbach and Kohler (2003) that is based on the idea that health-related terms should co-occur with the word "health" more often than non-health terms. The two other categories include methods that use the CHV vocabulary instead of the UMLS for its greater adequacy to health consumers' terminology. The first category

of methods has a binary output that can be "health" if the query has terms that are included in the CHV subset in use or "non-health", otherwise. From this point forward, this class of methods is denoted by *CHV methods with binary output*. The second category of methods computes a continuous output that quantifies the query's degree of association with the health domain. From this point forward, this class of methods is designated by *CHV methods with continuous output*.

In methods with a continuous output, that is, in *co-occurrence methods* and in *CHV methods with continuous output*, different thresholds may be applied to predict if the query is or is not a health query. The logic and implementation behind the CHV methods with continuous output allow, besides classifying queries as being health-related or not, the computation of an association degree of queries with UMLS specialized health categories.

Since one of the main disadvantages of a method based on vocabularies is its dependence on the language in which it was created, in the CHV methods with continuous output we have tested if they can be applied with a Portuguese translated version of the CHV without much penalty on the results.

16.2.1 Co-occurrence methods

As previously mentioned, these methods are based on the idea that health-related terms should co-occur together with the word "health" more often than non-health terms. For each query (Q) in the pool, two queries were submitted to a search engine: one (Q1) with the terms of the query Q and another (Q2) with the terms of Q plus the word "health". The health co-occurrence rate (cooc) of Q is calculated by the proportion of the total number of results of Q2 and the total number of results of Q1 as expressed in equation 16.1, where terms_Q is the set of terms that compose the query Q. If $\#\,\text{results}(\text{terms}_Q) = 0$, $\text{cooc}(Q) = 0$.

$$\text{cooc}(Q) = \frac{\#\,\text{results}(\text{terms}_Q \cap \text{health})}{\#\,\text{results}(\text{terms}_Q)} \tag{16.1}$$

This proportion indicates the relatedness of the query Q to the health domain because it represents the frequency of occurrence of Q's search terms and the word "health" in web pages.

In the work of Eysenbach and Kohler (2003), where this method was proposed, Google was the used search engine. Here, we have used Google and Yahoo! to determine the number of results. For example, in Google, the query 'diabetes symptoms' has a health co-occurrence rate of $\frac{478000}{929000} = 0.51$ and the query 'Pavarotti' has a health co-occurrence rate of $\frac{359000}{6440000} = 0.06$. We have also proposed a variant of these methods that combines both search engines' number of results. We have, therefore, implemented 3 methods with different health co-occurrence rates as expressed in Equations 16.2, 16.3 and 16.4.

$$G_{cooc_Q} = \frac{\#google(terms_Q \cap health)}{\#google(terms_Q)} \tag{16.2}$$

$$Y_{cooc_Q} = \frac{\#yahoo!(terms_Q \cap health)}{\#yahoo!(terms_Q)} \tag{16.3}$$

$$Y + G_{cooc_Q} = \frac{\#google(terms_Q \cap health) + \#yahoo!(terms_Q \cap health)}{\#google(terms_Q) + \#yahoo!(terms_Q)} \tag{16.4}$$

The differences detected in the number of results of both search engines, also stated in Chitu (2007), made us combine the number of results returned by the two search engines in the third method.

As shown in Figure 16.1, we have developed scripts, one for each search engine, to automatically get the number of results returned for each query in Google and Yahoo! through each search engine's API. Each of these scripts was then used by another script (`classifyQueries.pl`) that, for each query, asks the `numberofResults.pl` for the number of results of the query and the query plus the word "health". These values are then used to compute the health co-occurrence rate.

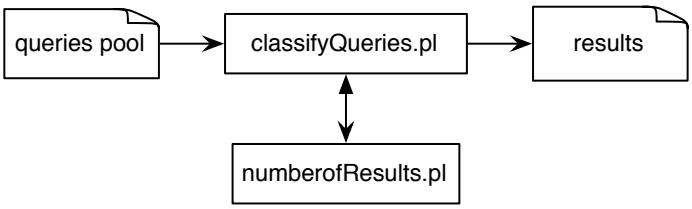

Figure 16.1: Co-occurrence methods global architecture - dataset files and Perl scripts

After the computation of the health co-occurrence rate, this value was compared with several thresholds (0; 0.05; 0.1; 0.15; 0.2; ...; 0.95; 1). In each comparison, if the health co-occurrence rate was larger than or equal to the threshold, the query was considered to be a health-related query at that threshold.

16.2.2 CHV methods with binary output

This category's methods differ on the subset of the terms used to classify the queries. The presence of one term in a query is sufficient to classify it as a health query.

The CHV vocabulary contains concepts of several categories and some of them contain strings (e.g.: car, driving) that, when isolated from other health terms or concepts, are not useful to identify a health query. To avoid false positives we decided to obtain different subsets of the CHV vocabulary besides using the complete CHV. The criteria behind the definition of the CHV subsets were defined empirically in an iterative process fed by the data analysis of the variants defined at that moment. Different results could have led to different criteria (e.g. use more terms if the previous results were showing performance improvements).

The 11 variants with different lists of terms are: CHV1 (all terms), CHV2 (terms associated with the 200 most frequent concepts), CHV3 (terms associated with the 400 most frequent concepts), CHV4 (terms associated with the 600 most frequent concepts), CHV5 (terms associated with the 800 most frequent concepts), CHV6 (terms associated with the 1,000 most frequent concepts), CHV7 (terms existing in the UMLS preferred names), CHV8 (terms

existing in the CHV preferred names), CHV9 (terms existing in the UMLS and CHV preferred names), CHV10 (6,000 most frequent terms[1]) and CHV11 (10,000 more frequent terms[1]).

As shown in Figure 16.2, we used two Perl scripts: one (`generateTermsList.pl`) that generates a subset of health terms and another one, similar in all CHV methods, that classifies queries. The `generateTermsList.pl` also removes stop-words, using a list of stop-words provided by the University of Glasgow[2], and replaces special characters that may be misunderstood by regular expressions that are used later to parse the files. The `classifyQueries.pl` simply checks if any of the query terms is present in the terms list. If present, the query is classified as health-related.

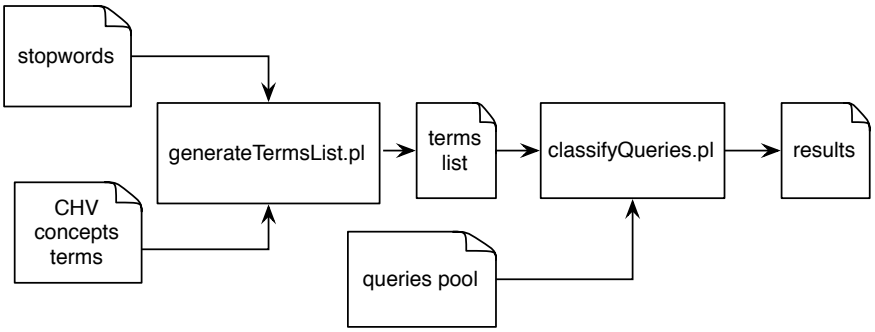

Figure 16.2: CHV methods global architecture - dataset files and Perl scripts

16.2.3 CHV methods with continuous output

For the same reasons expressed above, we decided to obtain different subsets of the CHV vocabulary instead of using only the complete CHV. We defined four subsets: one with concept strings from UMLS categories containing concepts more likely to occur in consumer health queries (HEALTH), one with the consumer preferred string for each concept in the CHV (CHVP), one with the UMLS preferred string for each concept in the CHV (UMLSP) and the other with the MedlinePlus Health Topics source vocabulary concept strings (MEDP).

The HEALTH subset was created including all the strings associated with concepts pertaining the UMLS semantic types that had a greater probability of containing terms used by health consumers on their health searches. The semantic types containing mostly concepts related to the biology and chemistry fields were excluded because their inclusion in health queries is unlikely. All the concepts directly under the following semantic types[3] were included in the HEALTH subset:

```
A1.2 Anatomical Structure
A1.2.1 Embryonic Structure
A1.2.3 Fully Formed Anatomical Structure
A1.2.3.1 Body Part, Organ, or Organ Component
```

[1] Obtained from the CHV website.

[2] http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words. Archived by WebCite at http://www.webcitation.org/6Dixv4Uc8

[3] The numeration is the one presented in the Semantic Network Browser.

```
A1.2.3.2 Tissue

A1.4 Substance
A1.4.1.1.1 Pharmacologic Substance
A1.4.1.1.1.1 Antibiotic

A2.1.4.1 Body System

A2.1.5.2 Body Location or Region

A2.2 Finding
A2.2.2 Sign or Symptom

B1 Activity
B1.1 Behavior
B1.3.1 Health Care Activity
B1.3.1.2 Diagnostic Procedure
B1.3.1.3 Therapeutic or Preventive Procedure

B2.2.1 Biologic Function
B2.2.1.1.2 Organ or Tissue Function
B2.2.1.2 Pathologic Function
B2.2.1.2.1 Disease or Syndrome
B2.2.1.2.1.1 Mental or Behavioral Dysfunction
B2.2.1.2.1.2 Neoplastic Process
```

For each subset, we created an inverted index containing the unique terms mapped to a list of unique identifiers for each concept string in the subset and their association degree with the concept string. The association degree of a term t to a concept string c, w_t^c, is computed as the ratio $\mathrm{tf}_t^c / |c|$, where the numerator is the term frequency of t in the concept string c and the denominator is the number of terms in concept string c. If we consider the CHV strings *tooth* and *dental infection*, the terms *dental* and *infection* would be associated with the second string with a probability of 0.5 and with the first string with a probability of 1.

In the classification process, queries are tokenized and, for each term, we retrieve the corresponding posting list from the inverted index. We then combine the posting lists of every term into a single list to which we call query list. As shown in Figure 16.3, two combination methods were tested. The first joins the lists and, when an identifier appears more than once, the w_t^c are added. The resulting list contains the weights of each CHV string in the query, w_c^q. This way we can easily identify if a query contains parts or entire health CHV strings. The second method (M2), joins the lists as M1, but also counts the occurrences of each CHV string in the query ($\mathrm{cf}_{c,q}$). As a final step, we adjust the weights calculated in the first method as $w_c^q \times \frac{\mathrm{cf}_{c,q}}{|q|}$, where $\mathrm{cf}_{c,q}$ is the frequency of c in query q and $|q|$ the number of unique terms in q.

After obtaining the query list, we calculate the final score that will be used to classify the query as health related or not. To do this, we propose some variants for the two previous methods, as presented in Table 16.1. In that table, Query is the query list obtained after the combination of the terms' lists in each method, $\mathrm{tf}_{h,q}$ is the number of terms in query q included in the inverted index, and $|q|$ is the number of unique terms in q. M1Max and M1MaxBoost use the maximum weight of the $Query$ list under the assumption that, if a query is completely matched by a health concept, it is a health query. In M1Avg and M1AvgBoost we computed the average of the 5 largest probabilities in the query list.

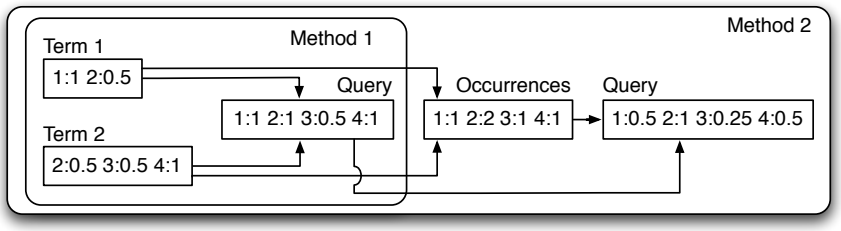

Figure 16.3: Joining posting lists in Methods 1 and 2.

Table 16.1: Variants applied to the different methods.

Variant	Formula	Boost
M1Max	$\max(Query) \times (tf_{h,q} \div \lvert q \rvert)$	No
M1MaxBoost		Yes
M1Avg	$avg(top_1^5(Query)) \times (tf_{h,q} \div \lvert q \rvert)$	No
M1AvgBoost		Yes
M2Max	$\max(Query)$	No
M2MaxBoost		Yes
M2Avg	$avg(Query)$	No

The product used in the M1 variants lowers the score of the queries that have non-health terms even if the query matches an entire concept, because a concept may change when a term is added. Consider, for example, the query "tooth piercing". Since "tooth" is a CHV concept and the term "piercing" is not, without the final product the above query would be scored with 1 instead of 0.5 as it is with the multiplication. This is not needed in the M2 variants because the M2 already uses the occurrences of each CHV concept string in the whole query.

To promote the queries that contain terms that appear more frequently in the CHV vocabulary, we decided to test the application of a boost value b to the term weights in a CHV string ($b \times w_t^c$). This boost is similar to the document frequency df used in Information Retrieval and is equal to the number of strings in the CHV in which the term appears.

Queries that have the final score above a specific threshold will be classified as being health-related. We also used the UMLS semantic network to assign health categories to each query. For this purpose, we created an index similar to the one described above where terms are replaced by CHV strings and the posting lists contain categories and not strings. After obtaining the query list as explained above we create another list with the category associated to each CHV string in the query list and the weight, w_c^q, previously associated with the string. If a category appears more than once, we select its maximum weight.

To evaluate the efficacy of our method in Portuguese we used the *Google Translator API*. We manually evaluated 1% of the total number of translated strings and concluded that 84.2% (95% CI: [82.3%, 85.9%]) of the translations were good, which is very satisfactory.

The evaluation of each method was done through the comparison of the classification made by a team of human assessors and the output classification of each method. In the CHV methods with binary output, the classification is immediately computed after the execution of the described scripts. In the methods with the continuous output, the classification only occurs after the calculation of the cooc/final rate and its comparison with each threshold. The best thresholds are determined after the analysis of all collected data.

We have used a collection of 20,000 web queries, randomly sampled from AOL Search in the Fall of 2004. Beitzel et al. (2005) used this collection in a research project where queries were classified into 20 topical categories by a team of approximately ten human assessors. One of the topical categories is health, where 1,197 queries are included. In the evaluation of the co-occurrence methods and the CHV methods with binary output we used the 20,000 queries.

In the evaluation of the CHV methods with continuous output we have used two datasets, one for each language. In Portuguese (PT) we have used a collection of 1,522 queries manually classified by medical students. The initial set of queries was composed of 1553 queries extracted from the SAPO Saúde search engine and several assessors classified each query. When classification ties occurred, we excluded the query. In the final dataset, 55.6% queries were health queries.

In the English language, to obtain a dataset of a similar size, we have used 1,647 queries from the AOL Search dataset where 1,197 are health queries (72.7% of the entire sample).

For each method, measures like sensitivity, specificity and accuracy were calculated. These can be expressed in terms of probabilities of the following events: H_{hc} (query is classified as health-related in a human classification), NH_{hc} (query is classified as non-health-related in a human classification), H_{ac} (query is classified as health-related in an automatic classification) and NH_{ac} (query is classified as non-health-related in an automatic classification).

Sensitivity (SEN) is the number of true positives divided by the sum of true positives with false negatives. It can be expressed as the conditional probability of having an automatic classification of health-related, given that the query was classified as health-related by a human: $P(H_{ac}|H_{hc})$.

Specificity (SPC) is the number of true negatives divided by the sum of true negatives with false positives. It can be expressed as the conditional probability of having an automatic classification of non-health-related when the query was classified as non-health-related by a human: $P(NH_{ac}|NH_{hc})$.

Accuracy (ACC) is the tax of correct classifications (either as health-related or as non-health-related) and is expressed as stated in Equation 16.5.

$$\frac{P(H_{ac} \cap H_{hc}) + P(NH_{ac} \cap NH_{hc})}{P(H_{hc}) + P(NH_{hc})} \tag{16.5}$$

Besides the computation of these measures, two Receiver Operating Characteristics (ROC) graphs for comparing the several discrete classifiers methods and the several continuous classifiers methods were also drawn. A ROC graph is a two-dimensional graph in which sensitivity is plotted on the Y-axis and the false positive rate (1-specificity) is plotted on the X-axis. It is a technique that

depicts relative tradeoffs between benefits (true positives) and costs (false positives), being useful for visualizing, organizing and selecting classifiers based on their performance (Fawcett, 2006).

16.3.1 Co-occurrence methods

As mentioned in Section 16.2, co-occurrence methods are continuous classifiers because they produce a continuous output (health co-occurrence rate) that may be considered an estimate of queries health-relatedness probability. Each method has its own health co-occurrence rate with the distribution presented in the histograms of Figures 16.4, 16.5 and 16.6. In these histograms, only health co-occurrence rates between 0 and 1 are represented. In the three methods we detected queries with health co-occurrence rates greater than 1: Google has 3,174, Yahoo! has 693 and Yahoo!Google has 1,417 queries. Google has a co-occurrence average of 0.45, Yahoo! of 0.32 and Yahoo!Google of 0.39. The standard deviation is also greater in Google (0.305), followed by Yahoo!Google (0.243) and Yahoo! (0.228).

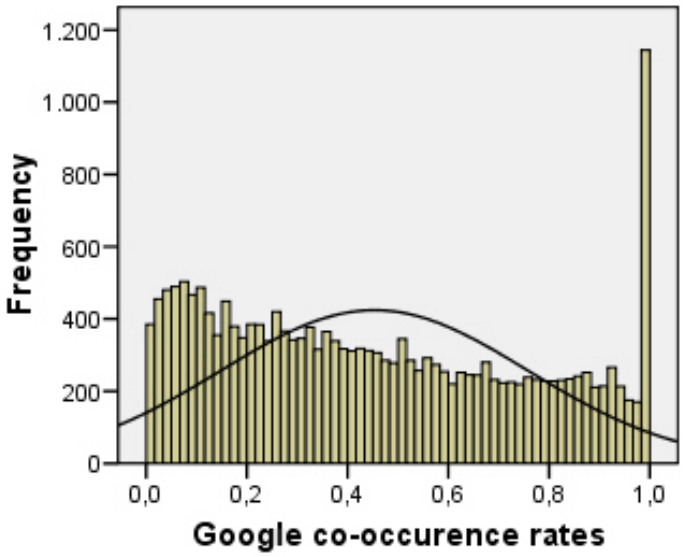

Figure 16.4: Google health co-occurrence rate histogram

To predict each query health-relatedness, this continuous output was then compared with different thresholds (ranging from 0 to 1). Sensitivity, specificity, accuracy, sum of sensitivity and specificity (SEN + SPC) and the distance of each method to the optimal point (0,1) in ROC space (ROCD), for the several thresholds in each method, are presented in Table 16.2. Each column's maximum value is highlighted in bold (except in the last column where the minimum value is the indicator of a best performance). The inclusion of the *SEN + SPC* value doesn't intend to be an indicator of the best method because sensitivity may be preferred over specificity in some cases and vice-versa. It is just a helpful measure to see which method has the greatest overall sum of sensitivity and specificity.

The ROC curves for each co-occurrence method are presented in Figure 16.7. Each point in the curve corresponds to a threshold value, starting on 1 at the left side of the graph.

Table 16.2: Sensitivity, specificity, accuracy and other measures for co-occurrence methods

	Sensitivity			Specificity			Accuracy			Sensitivity + Specificity			(0,1) ROC Distance		
Threshold	Yahoo!	Google	Y+G	Yahoo!	Google	Y+G	Yahoo!	Google	Y+G	Yahoo!	Google	Y+G	Yahoo!	Google	Y+G
1	0.07	0.21	0.12	**0.97**	**0.82**	**0.93**	**0.92**	**0.78**	**0.88**	1.04	1.02	1.05	0.93	0.82	0.88
0.95	0.08	0.28	0.15	0.97	0.80	0.92	0.92	0.77	0.88	1.05	1.08	1.07	0.92	0.74	0.85
0.9	0.13	0.37	0.21	0.96	0.77	0.91	0.92	0.75	0.87	1.09	1.14	1.13	0.87	0.67	0.79
0.85	0.19	0.43	0.29	0.96	0.74	0.90	0.91	0.72	0.87	1.15	1.17	1.19	0.81	0.63	0.72
0.8	0.27	0.49	0.36	0.95	0.71	0.88	0.91	0.70	0.85	1.22	1.20	1.24	0.73	0.59	0.65
0.75	0.36	0.54	0.43	0.93	0.68	0.86	0.90	0.67	0.84	1.29	1.22	1.29	0.65	0.56	0.59
0.7	0.44	0.58	0.51	0.92	0.65	0.84	0.89	0.65	0.82	1.36	1.24	1.35	0.56	0.54	0.52
0.65	0.53	0.63	0.58	0.90	0.62	0.81	0.88	0.62	0.80	1.42	1.25	1.39	0.48	0.53	0.46
0.6	0.60	0.68	0.65	0.87	0.59	0.77	0.85	0.59	0.77	1.47	1.27	1.42	0.42	**0.52**	0.42
0.55	0.67	0.72	0.70	0.83	0.55	0.73	0.82	0.56	0.73	1.50	**1.27**	1.43	0.37	0.53	0.40
0.5	0.73	0.75	0.76	0.79	0.51	0.68	0.79	0.53	0.69	**1.52**	1.27	**1.44**	**0.34**	0.55	**0.40**
0.45	0.77	0.79	0.80	0.74	0.48	0.62	0.74	0.49	0.63	1.50	1.26	1.42	0.35	0.57	0.43
0.4	0.81	0.81	0.84	0.67	0.43	0.56	0.68	0.45	0.57	1.48	1.24	1.40	0.38	0.60	0.47
0.35	0.85	0.85	0.88	0.60	0.39	0.48	0.62	0.41	0.51	1.45	1.24	1.36	0.42	0.63	0.53
0.3	0.88	0.87	0.92	0.52	0.34	0.41	0.54	0.37	0.44	1.40	1.21	1.32	0.49	0.67	0.60
0.25	0.90	0.89	0.93	0.44	0.29	0.33	0.46	0.32	0.36	1.34	1.18	1.26	0.57	0.72	0.67
0.2	0.92	0.91	0.94	0.36	0.24	0.25	0.39	0.27	0.29	1.28	1.15	1.19	0.65	0.77	0.75
0.15	0.93	0.94	0.96	0.27	0.18	0.18	0.31	0.22	0.22	1.20	1.12	1.14	0.73	0.82	0.82
0.1	0.95	0.97	0.98	0.19	0.13	0.11	0.23	0.17	0.15	1.14	1.09	1.08	0.81	0.87	0.89
0.05	0.96	0.99	0.99	0.11	0.06	0.05	0.16	0.11	0.10	1.07	1.05	1.04	0.89	0.94	0.95
0	**1.00**	**1.00**	**1.00**	0.00	0.00	0.00	0.05	0.05	0.05	1.00	1.00	1.00	1.00	1.00	1.00

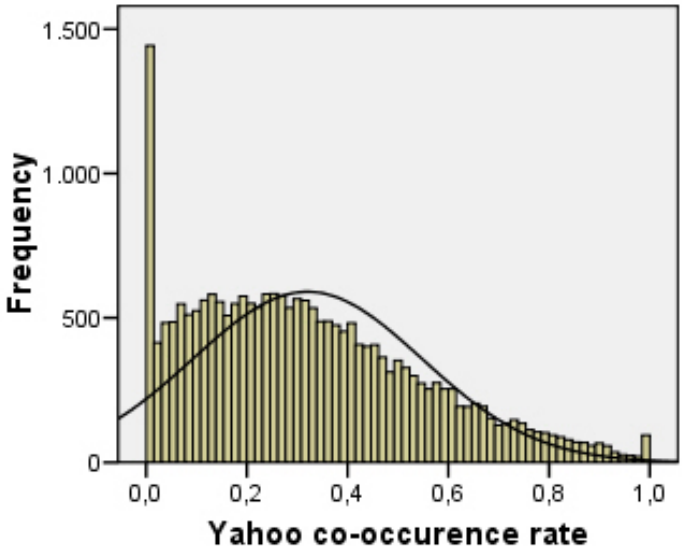

Figure 16.5: Yahoo! health co-occurrence rate histogram

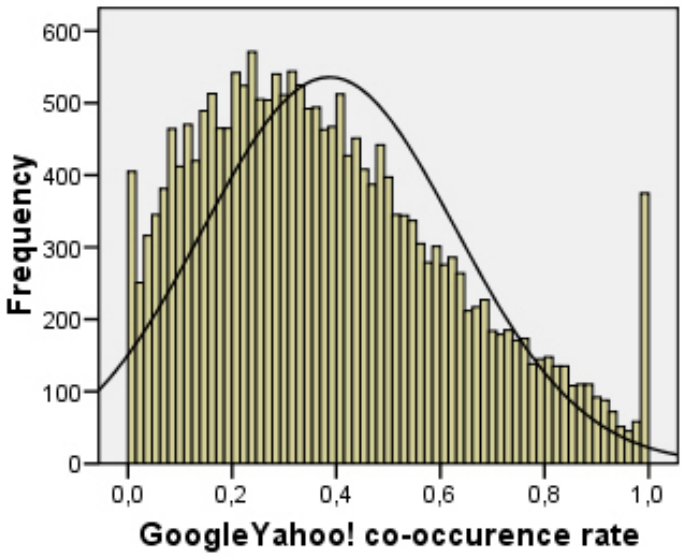

Figure 16.6: Yahoo!Google health co-occurrence rate histogram

16.3.2 CHV methods with binary output

Table 16.3 presents, for each CHV method, the number of terms used in the classification method (Terms), sensitivity, specificity, accuracy, sum of sensitivity and specificity and the distance of each method to the optimal point (0,1) in ROC space. Each column's maximum value is highlighted in bold (except in the last column where the minimum value is the indicator of a best performance). Just as in the co-occurrence methods, the sum of sensitivity and specificity does not intend to be a single evaluation measure of the optimal threshold.

To aid the comparison of the several methods, a ROC graph was drawn (Figure 16.8) for each method.

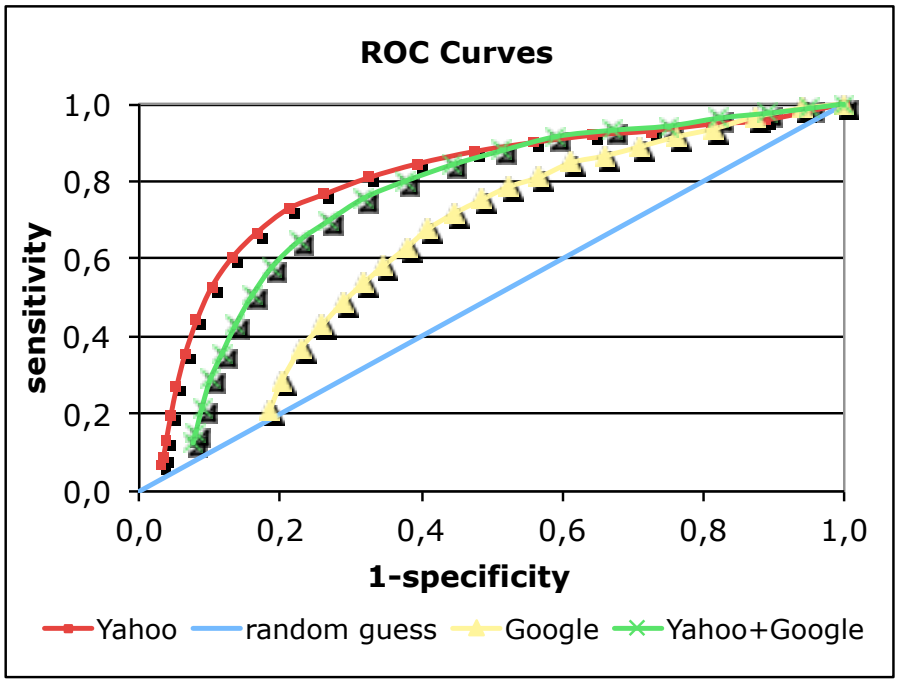

Figure 16.7: Co-occurrence methods ROC graph

Table 16.3: Number of terms, Sensitivity, Specificity, Accuracy and other Measures for CHV methods with binary output.

Method	Terms	SEN	SPC	ACC	SEN + SPC	ROC dist
1	**158783**	**0.73**	0.35	0.37	1.08	0.71
2	1616	0.42	**0.85**	**0.83**	1.27	0.60
3	2897	0.51	0.80	0.79	**1.31**	0.53
4	4404	0.55	0.75	0.74	1.30	0.51
5	5622	0.57	0.73	0.72	1.30	**0.51**
6	20354	0.67	0.52	0.53	1.18	0.59
7	27657	0.43	0.73	0.72	1.17	0.63
8	58655	0.63	0.49	0.50	1.12	0.63
9	66398	0.65	0.48	0.49	1.13	0.63
10	5898	0.69	0.59	0.60	1.28	0.51
11	9872	0.71	0.52	0.53	1.23	0.56

16.3.3 CHV methods with continuous output

Initial tests shows that the HEALTH subset produces the best results with respect to accuracy and distance to the ROC optimal point. However, the MEDP subset revealed a better specificity (86%-87%) due to a lower number of concept strings and its strong focus on consumers. In terms of sensitivity, M1Max using the UMLSP subset and M1Max using the CHV entire vocabulary had the best results with 68%. The UMLSP, despite having fewer strings than the CHV subset, has the same sensitivity probably because it contains almost all of the concept strings that led to query classification. In general, almost all methods have sensitivity and accuracy values above 60%.

Table 16.4 shows the results of each method used in the classification of the sample collections in both languages with the HEALTH Subset. As shown,

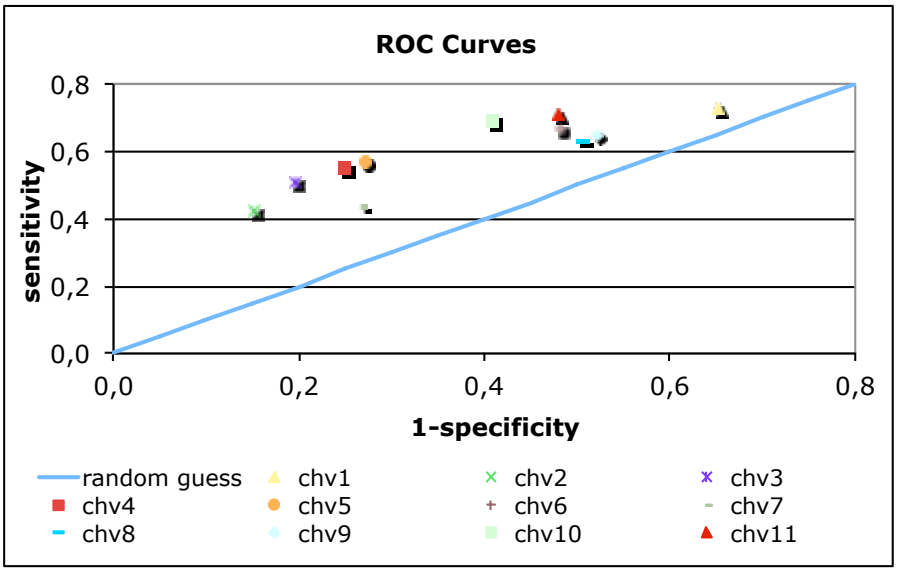

Figure 16.8: CHV methods with binary output ROC graph

the best method is M2Max with a threshold of 0.17 using the English vocabulary. In Portuguese the best method is M1Max with a threshold of 0.5. We can therefore conclude that translation has impact on the results. The difference in sensitivity and accuracy is negligible. However, differences in the distance to the ROC optimal point and specificity are more expressive. We believe our results can be improved by removing unspecialized terms that, used alone, are not health-related.

Table 16.4: Best results with the HEALTH subset. T=threshold, L=language.

M	T	L	SEN	SPE	ACC	SEN + SPE	ROCD
M1Max	0.2		**0.76**	0.67	**0.73**	1.43	0.41
M1Avg	0.2		0.66	**0.80**	0.70	1.46	0.39
M1MaxBoost	0.2		0.71	0.71	0.71	1.42	0.41
M1AvgBoost	0.75	EN	0.72	0.67	0.71	1.39	0.43
M2Max	0.17		0.68	0.79	0.71	**1.47**	**0.38**
M2Avg	0.1125		0.67	0.68	0.68	1.35	0.46
M2MaxBoost	0.35		0.71	0.71	0.71	1.42	0.41
M1Max	0.5		0.65	0.69	**0.67**	**1.34**	**0.46**
M1Avg	0.2		0.65	0.68	0.66	1.33	0.47
M1MaxBoost	0.75		0.66	0.67	**0.67**	1.33	0.47
M1AvgBoost	0.2	PT	0.67	0.65	0.66	1.32	0.48
M2Max	0.5		0.63	**0.70**	0.61	1.33	0.48
M2Avg	0.1		**0.68**	0.60	0.65	1.28	0.51
M2MaxBoost	0.75		0.66	0.67	0.66	1.33	0.47

16.4 DISCUSSION

Before analyzing the co-occurrence methods, we would like to mention the existence of health co-occurrence rates greater than 1. Theoretically, these val-

ues shouldn't exist because the default operator between terms in both search engines (Google and Yahoo!) is the logic "AND", meaning that all terms in a query without operators should appear in the retrieved documents. In theory, adding terms should only result in a maintenance or decrease of the number of results. The number of queries in this situation is higher in Google than in Yahoo! (3,174 against 693). The query "go carts" is one example (with 3,230,000 results in Google) and the query "go carts health" (with 8,470,000 results in Google). This may be explained by the fact that the number of results returned by search engines is usually just an estimate. Google Help Center (Google, 2012) explains that not providing the exact count allows them to return search results faster. Yet, the high number of these cases is still surprising.

Figures 16.4, 16.5 and 16.6 show that the Yahoo!Google health co-occurrence rate is the closest to the Normal distribution, followed by the Yahoo! health co-occurrence rate. It's also possible to verify the existence of a strange peak at the right side of the Google histogram and at the left side of the Yahoo! histogram. The higher frequency of values near 1 in Google histogram shows that, a large number of queries retrieves documents where the term "health" co-occurs with the other terms of the query. The peak in Yahoo! shows that a large number of queries return 0 results.

Analyzing the measures of Table 16.2 it is possible to verify that, as expected, sensitivity is 1 at a threshold of 0 (health co-occurrence rates are always bigger than 0 making all queries to be classified as health-related). Naturally, at this same threshold, specificity is 0 (since there aren't queries classified as non-health related). Mainly due to high specificity values at threshold of 1, accuracy is also maximized at this threshold. The sum of sensitivity and specificity measure has the best value at a threshold of 0.5 of the Yahoo! method (with 73% of sensitivity and 79% of specificity) just as the Yahoo!Google method. The Google method has its best sensitivity+specificity value at a 0.55 threshold. The analysis of the distance to the optimal point in the ROC Space keeps the threshold of 0.5 as the best of the Yahoo! method. Using Google, the best threshold value changes to 0.6 in the analysis of this last measure.

In the ROC graph of Figure 16.7 it is clear the dominance of Yahoo! over Google (always above it). In this graph it is also possible to detect the closer points of each method to the point (0,1).

The idea of joining the estimates of Yahoo! and Google into the third method hasn't produced the expected results (improvements when compared to the two other methods). As can be seen in Figure 16.7 and Table 16.2, the Yahoo!Google method has an intermediate performance, being probably better than Google due to Yahoo! performance.

Google results in this sample of 20,000 queries are different from the results of Eysenbach and Kohler (2003). In their work, the threshold of 35% was considered an optimal trade-off between sensitivity (85.2%) and specificity (80.4%). The sample used in their study was composed of 2985 queries. Comparatively, our study had worse sensitivity values (68% or 72%), specificity values (59% or 55%) and different optimal threshold values (0.6 or 0.55). The larger sample used in our study make us believe our results are a better portray of reality.

In Figure 16.8 it is possible to see that all CHV methods are better than a random guess (represented by a diagonal line) as they are located above it. In ROC graphs, the point (0,1) represents a perfect classification, so better perfor-

mances are closer to this point. Yet, no method has reached the results initially expected. In fact, the best methods, as can be seen in Figure 16.8, are CHV2, CHV3, CHV4 and CHV5 (methods that use the list of terms of the 200, 400, 600 and 800 most frequent concepts) and their sensitivity doesn't exceed 57%. The specificity and accuracy is greater in CHV2 but sensitivity has a low value (42%) in this method. CHV3 is the method with the larger sum of sensitivity (51%) and specificity (80%). CHV5 is the closest to the (0,1) point, the optimal point in ROC Space.

We can also see that the relation between the number of health terms and sensitivity is not a direct proportion. For example, CHV10 has less terms but higher sensitivity and specificity than CHV6. This means there are terms more related to the health context than others and that the performance of this type of methods could be improved by a careful selection of terms. Generally, all CHV methods present a low sensitivity.

The results of the CHV methods with continuous output in the English language show that these methods outperform the CHV methods with binary output and most of the co-occurrence methods. In fact, the best CHV method with continuous output has a ROC distance of 0.38 whereas the best CHV method with binary output has a ROC distance of 0.51 and the best co-occurrence method of 0.34 with the Yahoo! search engine. The CHV methods with continuous output have an extra advantage of allowing the association of queries with UMLS semantic types, which can improve the categorization of health queries.

The Portuguese results can not be compared with the results of the other type of methods but allow us to conclude that, although the translation has impact on the results, it can be a good strategy to apply CHV methods in non-English languages with further improvements in the translation process.

We would like to emphasize that the methods indicated as optimal may be discarded when compared to others if sensitivity is preferable to accuracy or vice-versa. For example, in a situation where we want to filter the number of queries to be categorized by a human assessor without the risk to eliminate a large number of health-related queries, it is preferable to have good sensitivity instead of specificity.

16.5 CONCLUSIONS

In this chapter we evaluated three types of methods. Two of them were proposed by us and use terms from the CHV vocabulary. One produces a binary output and the other a continuous one. A third one was proposed by Eysenbach and Kohler (2003) and evaluates query relatedness to health through the health co-occurrence rate of query terms with the word "health" in search engines' results.

While Yahoo! performed better than Google in the co-occurrence methods, its results were worse than Eysenbach and Kohler's results. In their work, at a threshold of 35%, sensitivity was 85.2% and specificity was 80.4%, while in our Yahoo! method, at a threshold of 0.5, sensitivity was 73% and specificity was 79%. We think our results depict reality more accurately since our sample of queries is an order of magnitude larger: 20,000 against 2,985 queries.

None of the binary methods that used subsets of terms of health vocabularies behaved as well as the Yahoo! method. Yet, some of CHV methods behaved better than the Google method (CHV3, CHV4 and CHV5 had better or similar performance than the Google method).

The continuous CHV methods outperformed most of the other methods and, not less important, allow the association of queries to the UMLS semantic tree and their classification into categories like *Disease or Syndrome* or *Anatomical Structure*. The output of our method can be useful to search engines that can, for example, use it to provide contextualized query suggestions or even information about the health subject searched for. The evaluation of this set of methods in a language different from the vocabulary language showed that the influence of the translation process in the proposed method is noticeable but does not compromise its overall effectiveness.

The next chapter is included in a new part of this dissertation that aims to conclude this dissertation and present lines of future work.

PART VI

CONCLUSION

CONCLUSIONS

17.1 INTRODUCTION

Although, traditionally, IR systems ignore context features surrounding search, we believe these features can be useful to improve the retrieval process. For that reason, in this dissertation we have investigated how context features affect consumer health information retrieval and how they can be used in query formulation to improve retrieval. In Part II we analyzed the influence of several context features in the retrieval process. Motivated by some of the findings obtained in Part II, in Part III we studied how the language and the terminology of a query affect various outcomes of the retrieval process, in users with different levels of language proficiency, health literacy and topic familiarity. Findings of these studies stimulate the development of a query suggestion system that is described in Part IV, where the system is also evaluated considering the users' characteristics mentioned above. In Part V we describe preliminary work regarding the automatic identification of health queries. In this chapter we conclude and summarize the main contributions of the dissertation.

17.2 EXPLORATORY STUDIES

In Part II we explored the influence of context features in health IR. In the first study, we compare several search engines in different health situations. We concluded that generalist search engines surpass health-specific ones in terms of precision, and also that Google is the preferred engine and the one with better precision. Since the superiority of this engine is more expressive in the top of the rank, Google's first results page is a good place to start a health search session. However, because health-specific engines are more balanced in the results provided for conditions with different levels of severity, it is also a good practice to use them to refine results. This behavior may help prevent escalations on medical concerns. Another good strategy in health IR involves transforming narrower clinical query types, like the prognosis one, into overview queries whenever the former has bad results. We also found that the medical specialties with higher precision at the top-5 results, psychiatry and gynecology, are also the most popular in number of web pages and number of web searches. Moreover, these specialties make us hypothesize that the Web is richer in contents associated with sensitive issues in which anonymity may be desirable.

In Chapter 6 we analyzed the impact of certification, concluding that users value the diversity provided by generalist search engines even if this means in-

cluding non-certified documents. However, if generalist search engines want to make certain the medical accuracy of the knowledge obtained in the session, they have to ensure that users understand the retrieved documents. This should be done retrieving documents with a readability and terminology adjusted to users' knowledge. Another factor affecting the medical accuracy is the presence of, even if only a few, documents with unreliable information. Therefore, to assure the credibility of the top results, the ones receiving more attention, it is advisable to incorporate the medical certification in the set of criteria currently in use by the search engine. This suggestion is enforced by users' behavior that do not consistently check if the retrieved information is or is not certified.

The last exploratory study considers the largest number of context features. It studies their impact on query formulation and on relevance assessment. We concluded that the use of medico-scientific terminology leads to higher rates of successful health searches and is used more often by users with a higher familiarity with the topic. Although English queries led to lower relevance scores, we think this was caused by the low English literacy of the users. The larger number of English queries in tasks with more familiar topics or clearer definitions suggests the translation of terms to their English synonym might be a good strategy to improve search in users with good health literacy or users familiar with the topic. Findings also make us suspect that certain types of clinical queries, like the prognosis/outcome one, need more user context to be successful. Similarly to what we found in the first exploratory study, the Web seems to be richer in subjects where anonymity is preferable. Finally, we found that situational and motivational relevance are not always in harmony, suggesting that evaluation models should incorporate several types of success measures.

17.3 USE OF CONTEXT IN QUERY FORMULATION SUPPORT

In Part III we focused on query formulation support. In Chapter 9, based on the assumption that a language popular on the Web has a higher probability of having high-quality documents, we studied the impact of translating a query to the English language in users with different levels of English proficiency. Some of our findings corroborate our assumption. In fact, we found that English health content has a larger proportion of health-certified documents, is more suited to disseminate health information and is associated with less HTTP errors. Results suggest that translation approaches should be used only on users with, at least, elementary English proficiency. Despite having a higher precision on English queries, low English proficiency users have a lower degree of comprehension of English documents, obtain less accurate knowledge through English queries and feel less satisfied in the tasks with this type of queries. A cross-lingual assistance personalized to users' English proficiency could improve non-English consumer health retrieval and could be helpful in an educational sense, enabling non-English speaking users to learn English medical terminology. Moreover, it may also be helpful to trigger new search strategies and to help the user construct queries that give access to documents that may not be reached otherwise. Similarly to what we have found in the

second exploratory study, we found that the readability of documents should be a criterion for ranking, especially if the user is proficient in the documents' language.

In Chapter 10 we studied the impact of translating the terminology of a query, in users with different levels of health literacy and topic familiarity. Findings suggest that users with inadequate health literacy and users who are unfamiliar with the topic should be provided with recommendations of lay queries. On the other hand, users with higher health literacy or higher topic familiarity should be given alternative queries using medico-scientific terminology. This would not only give access to new types of documents but would also foster the learning of terminology that can be used in future queries. Since domain expertise evolves over time, the query suggestion system should be dynamic, providing continuous learning to users in the lowest levels of health literacy and topic familiarity. We have also concluded that users with inadequate health literacy should be provided with documents adequate to them, either with pictorial contents or with higher levels of readability. Reinforcing previous findings, we verified that readability is important to every health consumer, using both types of queries, and should be incorporated in search engines' ranking algorithms. In fact, we found that the relevance of a document highly depends on its comprehension. Health websites who want to provide information to consumers should also be aware that, if they need to use medico-scientific terminology, they should, at least, simplify the remaining contents.

In Chapter 11, we study how readability, comprehension, precision, medical accuracy and motivational relevance influence each other, considering the terminology of the query. Once again, we conclude that readability is essential for a document to be at least partially relevant and it is even more important if the document contains medico-scientific terminology. Moreover, readability was found to be crucial to assure the lowest levels of satisfaction and not so essential to the highest levels. This shows that readability is a *basic need* of a successful health search. Since our results suggest that documents with more accurate medical contents are significantly harder to read, low literacy users face a serious obstacle in health IR. We found that the relevance of a document highly depends on its comprehension and that unsuccessful tasks have lower precision than more successful tasks. The medical accuracy was found to be associated with relevance assessments, an association stronger when a lay query is used. This shows that users can, at least in part, relate their relevance assessments with the medical accuracy of the documents. In lay queries, comprehension is more crucial to the accuracy of the resulting knowledge than in medico-scientific queries. In the latter, either the user understands documents worse but is still able to assimilate at least part of the contents or the higher accuracy of medico-scientific documents, when compared with lay documents, counteracts users' lower comprehension of these documents. Comprehension is also an important and influent factor to motivational relevance.

In the last chapter of Part III we analyze how health literacy, topic familiarity and users' previous interaction affect the terminology used in queries. We conclude that health consumers rarely use medico-scientific terminology in their queries, a suspicion that also arouse in the third exploratory study. However, users with higher health literacy or topic familiarity, do it more often. On the other hand, low health literacy and topic familiarity users have

more difficulties in query formulation, not only on selecting and typing the appropriate medical terms but also on general aspects like the inclusion of advanced operators. This shows that low health literacy and topic familiarity users need features that enable them to find what they look for. Concerning query reformulation, we found that access to documents containing medico-scientific terminology encourages the use of this type of terminology in subsequent queries showing that the benefits of a query suggestion system are twofold. It not only provides access to documents that would be missed without the given suggestions but also stimulates the use of different terms in subsequent queries. Nonetheless, to guarantee that users understand the retrieved documents, it is important to adapt the query suggestions to users' health literacy and topic familiarity. Finally, this study has suggested that: "Users familiar with a health topic are more likely to use medico-scientific terminology in their past searches about that topic than users with less familiarity with the topic".

17.4 QUERY SUGGESTION SYSTEM

In Part IV we describe the suggestion system we developed to offer the user a maximum of 4 alternative suggestions using a combination of Portuguese or English language with lay or medico-scientific terminology. This system uses the original CHV vocabulary and also a translated version of it. In Chapter 14 we describe the experiment we conducted to evaluate the suggestion system, present and analyze the gathered results. In the last chapter of this part, we discuss our results.

We found that suggestions are used more often and more extensively in the initial stages of a search session. The suggestions offered by the system have a good acceptance and those with a higher degree of novelty seem to be preferred. English suggestions are preferred to Portuguese ones in *basic* and *proficient* users. Medico-scientific suggestions tend to be preferred to lay ones in the higher levels of health literacy and the extraction of new terms from these suggestions increases with health literacy.

A retrieval system that includes the implemented suggestion system without any type of personalization tends to be better than a system without suggestions in terms of precision, correctness and incorrectness and tends to be slightly worse in terms of motivational relevance. Of these, only the incorrectness difference is significant, an outcome extremely relevant in the health domain. The best behavior of the suggestion tool is achieved when users use it strictly as a suggestion tool, that is, when they click in suggestions, and when they use all terms from suggestions. However the medical accuracy of the obtained knowledge can also be improved using suggestions as sources of terms.

Moreover, we also found that the personalization of this system to users' English proficiency and health literacy, biasing users towards the suggestions more beneficial to them, outperforms the system without personalization, in terms of medical accuracy of the obtained knowledge.

17.5 CONTEXT PREDICTION

In Chapter 16 we propose two types of methods to identify health queries and compare them with a method proposed by other authors. Both of our proposals use the CHV vocabulary, one producing a binary output and the other producing a continuous output. We propose several variants in each type of methods. In the overall comparison, the CHV method with continuous output outperforms most of the other methods' variants. Moreover, the CHV method with continuous output has the advantage of allowing the association of queries to the UMLS semantic tree and their classification into categories like *Disease or Syndrome* or *Anatomical Structure*. This information can be useful to search engines that can, for example, use it to provide information about the health subject searched for.

FUTURE WORK

18.1 INTRODUCTION

The work described in this dissertation has explored several issues related to the use of context in health information retrieval. In this chapter we describe lines of future work that have emerged and can take this work further. We organize and discuss these lines of future work in four sections. We begin by describing possible improvements for the developed suggestion system. Then, we identify lines of future work to automatically acquire context and explain our ideas for the development of a Portuguese consumer health vocabulary. Finally, in the last subsection, we present ideas for various studies and experiments.

18.2 QUERY FORMULATION SUPPORT

The evaluation of the implemented suggestion system showed promising results. However, there is space for several improvements to this initial version of the suggestion system. First, considering the strategy defined in Chapter 15, the system should be personalized to users' English proficiency and health literacy, providing only appropriate suggestions to each user. Ideally, these characteristics should be automatically inferred from past behaviors. However, in a first stage, users could explicitly provide this information.

The data structures used by the system can also be improved and extended. The translation of the CHV can be refined and an English index can be built to cope with English queries. In addition, UMLS lexical tools could be used in the construction of the indices to deal with cases like multi-word terms. It is also easy to visualize other information sources that can be used to enhance the suggestion system. In terms of data structures, since there is a connection point between CHV and UMLS concepts, other UMLS information like semantic types and semantic relationships or even designations used by other thesauri could be stored. Moreover, data about previous user interaction like clicked suggestions, terms used from suggestions or even clicked documents could be logged.

In the developed suggestion tool, the score assigned to the pairs (query, string) is the sum of the inverse string frequency of each query term as defined in: $\text{score}(q, s) = \sum_{t \in q} \text{isf}_t$. Similarly to what has been suggested in the indices construction, the UMLS lexical tools could be useful when matching query terms with indices terms. Moreover, MetaMap, a program that maps text to the UMLS Metathesaurus, could also be a valuable resource.

In a first stage of development, following the computation of scores for pairs (query, string), we decided to select only the string with the maximum score for the query. The concept associated with the selected string is the one used to formulate the four suggestions presented to the user. It is easy to envision ways to enhance the concept selection using the information sources mentioned above. The preferences of users that interacted with the system in the past in identical circumstances can be useful to boost some strings in the ranking. The formulation of the suggestions could also take into account the semantic type of the main concept, its semantic relationships and its designations in other UMLS thesauri like the ones in Portuguese. Moreover, other web resources like Wikipedia or medical websites can be used to extract useful information. For example, definitions about the main concept and terms that frequently co-occur with its strings can also be used in alternative suggestions. Measures of semantic similarity like the ones proposed by Pedersen et al. (2007) could be used to compute the similarity of the query and string or to reach concepts similar to the main concept.

18.3 AUTOMATIC ACQUISITION OF CONTEXT

Although the automatic acquisition of context was not one of our main goals, some of our conclusions help to envision how certain features can be predicted. This is one of the main lines of future work. The prediction of user characteristics avoids disturbing the user and allows personalization. The classification of documents according to the degree of expertise they demand is useful for personalized document ranking. The automatic characterization of work task contexts is useful to assess the context the system needs to fit in and is one of the innovative approaches for studying contextual factors, as pointed by Agosti et al. (2012).

One of the user characteristics explored in this work was the proficiency in other languages, English in our case. Our findings showed that users with higher English proficiency tend to use English queries more often than users with less proficiency. We would therefore like to specifically study if past behaviors like the language used in past queries can be used to predict user proficiency in that language.

Topic familiarity was another context feature addressed in this work. Although it was not found to be discriminant for the personalization of the suggestion system, it might be useful to detect this feature automatically to intervene in other stages of the retrieval process. Two of our studies, based on different experiments, showed that the use of medico-scientific terminology tends to be associated with higher familiarity with the topic. Therefore, we would like to analyze how topic familiarity can be predicted through past queries, other types of retrieval behaviors or even information pertaining the users' clinical history. By means of past queries this can be done, for example, through the proportion of queries containing medico-scientific terminology or the degree of technicality of the queries using a model like the one proposed by Yang et al. (2011). Users' clinical history can be useful to obtain the time elapsed since the diagnosis of a condition which imply higher background knowledge on that specific disease/condition. The dynamic nature of topic familiarity makes this a challenging goal.

The automatic detection of health literacy is also challenging. This is a characteristic that is less related with users' past behaviors on a search engine. Yet, our findings have suggested that web search expertise may be related with health literacy, and this should be explored and perhaps used in the prediction. Note that web search expertise can, more easily, be inferred from users' past queries. Moreover, it would also be interesting to analyze if health literacy, similarly to what happens in generic information literacy (Kutner et al., 2006), is related with demographic and socioeconomic factors. If so, similarly to what has been done by Weber and Castillo (2010), Census information along with location information, explicitly provided by the user or, less accurately, gathered from IP addresses, could be used to infer health literacy. The model proposed by Yang et al. (2011) to assess the degree of technicality of a text or word can also be used to predict this user characteristic.

The classification of documents according to the degree of expertise they demand can be used to adjust document ranking to the user conducting the search. We have found that the HONcode categorization of documents as "for patients" and "for health professionals" can contribute to the above classification. Moreover, several of our studies also suggest that documents' readability should be considered to predict the degree of expertise demanded by a document.

Several characteristics of the work task can also be automatically assessed. First, and as an extension of our preliminary work on the identification of health queries, we would like to evaluate the use of the *normalized Google distance*, proposed by Cilibrasi and Vitányi (2006). Moreover, we would like to use co-occurrence rates, match scores with vocabulary concepts and the *normalized Google distance* as features in machine learning techniques. The use of these features can be done separately or in combination.

Similarly to what is done in the CHV method with continuous output, described in Chapter 16, classifying queries into health categories like *Disease or Syndrom* or *Anatomical Structure*, we would like to predict the medical specialty, the clinical query type and the severity of the condition behind a health query. This might be done using resources like the UMLS, non-NLM thesauri or even the Wikipedia that contains several health categories. Simple co-occurrence methods, like the one proposed by Eysenbach and Kohler (2003), or more complex ones, like the one proposed by Cilibrasi and Vitányi (2006), can be used for this purpose. For example, the co-occurrence with the word "death" or other severe concepts might be useful to predict the severity of the condition behind a query. To predict the severity of a condition, it would be interesting to see if seeking behaviors specific to severe diseases/non-severe diseases can be used to predict this context feature.

18.4 PORTUGUESE CONSUMER HEALTH VOCABULARY

The lack of Portuguese vocabularies with health consumer terminology led us to translate the Consumer Health Vocabulary using the Google Translator API. A manual assessment of 1% of the translated strings showed that 84.2% were good translations, which is very satisfactory. However, we still feel that this translation could be improved. The idea is to use a crowdsourcing ap-

proach, ideally composed of health professionals that, using a pre-defined strategy, could rectify translations and validate the overall translation. After this step, difficulty measures like the ones included in the CHV for each string could be computed for the Portuguese language. The strategies behind these computations could, in a first stage, be similar to the ones used in the original CHV. In the end of this process, the translated version of CHV could be made available to the public as another resource of the Consumer Health Vocabulary Initiative.

18.5 FURTHER STUDIES

Of diverse nature, several ideas for further studies have emerged during this work. They are briefly described next.

Exploration of task context features

Following hypotheses raised during the exploratory studies, we would like to compare the retrieval performance of different medical specialties in a more focused study. Are there specialties like the dermatology one, in which it is harder to textually explain the problems, associated with worst results? Are specialties like the psychiatry and gynecology ones, in which there are several sensitive issues, associated with better results? Note that the Web may be a preferable medium to the discussion of certain health subjects and, as a consequence, have more information about these subjects.

Another study should focus on clinical query types. One hypothesis that has come out is that certain types of clinical queries, like the Prognosis/Outcome type, might need more user context to be successful. A study about this could provide insights on how to deal with different types of clinical queries, gathering more or less context in conformity with the query. For example, overview tasks have a more exploratory nature and, probably, simple queries are enough to obtain a general idea about the topic, as suggested by Aula (2003).

Personalized cross-language IR studies

In the sequel of the study presented in Chapter 9, has emerged the idea of testing the validity of our conclusions in different languages. It would be interesting to analyze if languages other than Portuguese would still benefit of translations of queries to English. Note that even Russian, the second most popular language (W3Techs, 2013), is still very far from the English language (54.9% the Web against 6.1% of the Web).

Moreover, considering that the personalized cross-language strategy we have proposed only benefits users with at least some English proficiency, we would like to explore how users in the lowest levels of proficiency can be helped. A good strategy might be to have query translation complemented with machine translation of the retrieved documents. This way, even these users could have access to higher-quality English content.

Improve the search experience of users with low health literacy and low topic familiarity

Our concerns with the users with low English proficiency also apply to users with low health literacy and topic familiarity. Unlike users with more literacy or expertise in the topic, these users don't benefit from all the alternative suggestions. It is, therefore, necessary to analyze how these users can be helped to be more successful without hindering learning over time. Three key guidelines, proposed by Summers and Summers (2005), to help low health literacy users read and navigate websites, could be useful in the context of IR. These guidelines suggest enhancing readability, simplifying navigational structure, opting for linear information paths, and presenting information in visual formats. In a first stage, these suggestions could be explored at the ranking and interface levels of an IR system. Note that several of our studies also point to the incorporation of contents' readability in the criteria for ranking.

Document ranking

In this work we have acted on query formulation support mainly because it is less demanding in terms of technical infrastructures and specific data sets. However, some of our conclusions suggest that the use of context features in the ranking stage could also lead to positive results. A conclusion that emerged in several studies encourages the inclusion of readability in the criteria used for ranking. Similarly to what Summers and Summers (2005) has concluded, we found readability is beneficial for every type of user. Moreover, we found it to be especially important if the user is proficient in the documents' language. Medical certification is another characteristic that would be important to assure in the first set of retrieved documents.

We would also like to study if and how documents' terminology could be used as a criterion for ranking, in users with different levels of health literacy and topic familiarity. In this work we have used a binary scale to assess the terminology of the queries. However, to assess the technicality degree of a document, we could use a non-binary scale. We might, for example, use the 3-level scale proposed by Crain et al. (2010) or even the continuous scale proposed by Yang et al. (2011).

Indexing and Searching

The development of the suggestion system also contributed to some ideas related to the Indexing and Searching stage. Following the mapping of queries to medical concepts, the same operation could be applied in documents, that is, both queries and documents could be mapped to concepts and search would then be based on matching concepts instead of terms. This matching could be based on the concepts' name or could use measures of semantic similarity. For example, it might be useful to retrieve documents about diseases that have symptoms included in the query. Moreover, additional information about the concepts, like their definition, gathered from one or more resources, could be used in this stage. It would also be interesting to analyze the utility of co-occurrence information to indexing and searching.

BIBLIOGRAPHY

Abel, Fabian; Henze, Nicola; and Krause, Daniel. 2008. *Ranking in folksonomy systems: can context help?* In Proceeding of the 17th ACM conference on Information and knowledge management, CIKM '08, pp. 1429–1430. ACM, New York, NY, USA. ISBN 978-1-59593-991-3. doi:10.1145/1458082.1458316. URL http://dx.doi.org/10.1145/1458082.1458316. Cited on p. 45.

Adams, Anne and Blandford, Ann. 2005. *Digital libraries' support for the user's 'information journey'*. In JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, pp. 160–169. ACM Press, New York, NY, USA. ISBN 1581138768. doi:10.1145/1065385.1065424. URL http://dx.doi.org/10.1145/1065385.1065424. Cited on p. 30.

Agosti, Maristella; Berendsen, Richard; Bogers, Toine; Braschler, Martin; Buitelaar, Paul; Choukri, Khalid; Di Nunzio, Giorgio M.; Ferro, Nicola; Forner, Pamela; Hanbury, Allan; Heppin, Karin F.; Hansen, Preben; Järvelin, Anni; Larsen, Birger; Lupu, Mihai; Masiero, Ivano; Müller, Henning; Peruzzo, Simone; Petras, Vivien; Piroi, Florina; de Rijke, Maarten; Santucci, Giuseppe; Silvello, Gianmaria; and Toms, Elaine. Sep. 2012. *PROMISE Retreat Report: Prospects and Opportunities for Information Access Evaluation*. Tech. rep., PROMISE network of excellence, grant agreement no. 258191. Cited on pp. 13, 42, and 272.

Ahn, Jinhyun; Jung, Hyosook; Kim, Heejin; Sun, Dongeun; and Park, Seongbin. 2008. *A System for Contextual Search*. In iwsca, vol. 0, pp. 96–98. doi:10.1109/iwsca.2008.20. URL http://dx.doi.org/10.1109/iwsca.2008.20. Cited on p. 45.

Al-Maskari, Azzah and Sanderson, Mark. 2010. *A review of factors influencing user satisfaction in information retrieval*. In Journal of the American Society for Information Science and Technology, vol. 61, no. 5, pp. 859–868. ISSN 1532-2890. doi:10.1002/asi.21300. URL http://dx.doi.org/10.1002/asi.21300. Cited on pp. 152 and 157.

Al-Maskari, Azzah and Sanderson, Mark. Sep. 2011. *The effect of user characteristics on search effectiveness in information retrieval*. In Information Processing and Management: an International Journal, vol. 47, no. 5, pp. 719–729. Cited on p. 152.

Albakour, M. Dyaa; Kruschwitz, Udo; Nanas, Nikolaos; Kim, Yunhyong; Song, Dawei; Fasli, Maria; and De Roeck, Anne. 2011. *AutoEval: an evaluation methodology for evaluating query suggestions using query logs*. In Proceedings of the 33rd European conference on Advances in information retrieval, ECIR'11, pp. 605–610. Springer-Verlag, Berlin, Heidelberg.

ISBN 978-3-642-20160-8. URL http://portal.acm.org/citation.cfm?id=1996889.1996966. Cited on p. 199.

Allan, James; Aslam, Jay; Belkin, Nicholas; Buckley, Chris; Callan, Jamie; Croft, Bruce; Dumais, Sue; Fuhr, Norbert; Harman, Donna; Harper, David J.; Hiemstra, Djoerd; Hofmann, Thomas; Hovy, Eduard; Kraaij, Wessel; Lafferty, John; Lavrenko, Victor; Lewis, David; Liddy, Liz; Manmatha, R.; Mccallum, Andrew; Ponte, Jay; Prager, John; Radev, Dragomir; Resnik, Philip; Robertson, Stephen; Rosenfeld, Roni; Roukos, Salim; Sanderson, Mark; Schwartz, Rich; Singhal, Amit; Smeaton, Alan; Turtle, Howard; Voorhees, Ellen; Weischedel, Ralph; Xu, Jinxi; and Zhai, Chengxiang. 2003. *Challenges in information retrieval and language modeling*. In SIGIR Forum, vol. 37, no. 1, pp. 31–47. Cited on pp. 3, 4, 6, 7, 42, and 152.

Anagnostopoulos, Ioannis and Maglogiannis, Ilias. 2007. *Monitoring browsing behaviour and search services evolution adaptation with a capture-recapture Internet-based programming technique: A case-study over medical portals*. In Information Services and Use, vol. 27, no. 3, pp. 105–122. Cited on p. 33.

Aronson, A. R. and Rindflesch, T. C. 1997. *Query expansion using the UMLS Metathesaurus*. In Proceedings of the AMIA Annual Fall Symposium, pp. 485–489. ISSN 1091-8280. URL http://view.ncbi.nlm.nih.gov/pubmed/9357673. Cited on p. 52.

Aula, A. 2003. *Query Formulation in Web Information Search*. In Proc. IADIS International Conference WWW/Internet, vol. I, pp. 403–410. Cited on pp. 98 and 274.

Aula, Anne; Khan, Rehan M.; and Guan, Zhiwei. 2010. *How does search behavior change as search becomes more difficult?* In Proceedings of the 28th international conference on Human factors in computing systems, CHI '10, pp. 35–44. ACM, New York, NY, USA. ISBN 978-1-60558-929-9. doi:10.1145/1753326.1753333. URL http://dx.doi.org/10.1145/1753326.1753333. Cited on pp. 188 and 191.

Baeza-Yates, Ricardo and Ribeiro-Neto, Berthier. May 1999. *Modern Information Retrieval*. Addison Wesley, 1st edn. ISBN 020139829X. URL http://www.amazon.com/Modern-Information-Retrieval-Ricardo-Baeza-Yates/dp/020139829X. Cited on p. 3.

Bai, J. and Nie, J. Nov. 2008. *Adapting information retrieval to query contexts*. In Information Processing & Management, vol. 44, no. 6, pp. 1901–1922. ISSN 03064573. doi:10.1016/j.ipm.2008.07.006. URL http://dx.doi.org/10.1016/j.ipm.2008.07.006. Cited on p. 45.

Baker, D. Sep. 1999. *Development of a brief test to measure functional health literacy*. In Patient Education and Counseling, vol. 38, no. 1, pp. 33–42. ISSN 07383991. doi:10.1016/s0738-3991(98)00116-5. URL http://dx.doi.org/10.1016/s0738-3991(98)00116-5. Cited on p. 125.

Barry, Carol L. Apr. 1994. *User-defined relevance criteria: an exploratory study*. In J. Am. Soc. Inf. Sci., vol. 45, no. 3, pp. 149–159. ISSN 0002-8231. doi:10.1002/(sici)1097-4571(199404)45:3\%3C149::aid-asi5\%3E3.0.co;

2-j. URL http://dx.doi.org/10.1002/(sici)1097-4571(199404)45:3%3C149::aid-asi5%3E3.0.co;2-j. Cited on p. 99.

Bates, Marcia J. Jan. 1990. *Where should the person stop and the information search interface start?* In Information Processing & Management, vol. 26, no. 5, pp. 575–591. ISSN 03064573. doi:10.1016/0306-4573(90)90103-9. URL http://dx.doi.org/10.1016/0306-4573(90)90103-9. Cited on pp. 198 and 199.

Baujard, Vincent; Boyer, Celia; and Geissbühler, Antoine. 2011. *Evolution of Health Web certification, through the HONcode experience.* In User Centred Networked Health Care - Proceedings of MIE 2011, vol. 169, pp. 53–57. IOS Press, The Netherlands. ISBN 978-1-60750-805-2. Cited on pp. 87, 88, and 121.

Becker, Shirley A. Dec. 2004. *A study of web usability for older adults seeking online health resources.* In ACM Trans. Comput.-Hum. Interact., vol. 11, no. 4, pp. 387–406. ISSN 1073-0516. doi:10.1145/1035575.1035578. URL http://dx.doi.org/10.1145/1035575.1035578. Cited on p. 147.

Beitzel, Steven M.; Jensen, Eric C.; Frieder, Ophir; Lewis, David D.; Chowdhury, Abdur; and Kolcz, Aleksander. 2005. *Improving Automatic Query Classification via Semi-Supervised Learning.* In Proceedings of the Fifth IEEE International Conference on Data Mining, ICDM '05, pp. 42–49. IEEE Computer Society, Washington, DC, USA. ISBN 0-7695-2278-5. doi:10.1109/icdm.2005.80. URL http://dx.doi.org/10.1109/icdm.2005.80. Cited on pp. 247 and 253.

Belkin, Nicholas J. Jun. 2008. *Some (what) grand challenges for information retrieval.* In SIGIR Forum, vol. 42, no. 1, pp. 47–54. ISSN 0163-5840. doi:10.1145/1394251.1394261. URL http://dx.doi.org/10.1145/1394251.1394261. Cited on p. 6.

Berland, G. K.; Elliott, M. N.; Morales, L. S.; Algazy, J. I.; Kravitz, R. L.; Broder, M. S.; Kanouse, D. E.; Muñoz, J. A.; Puyol, J. A.; Lara, M.; Watkins, K. E.; Yang, H.; and McGlynn, E. A. Maro 2001. *Health information on the Internet: accessibility, quality, and readability in English and Spanish.* In JAMA : the journal of the American Medical Association, vol. 285, no. 20, pp. 2612–2621. ISSN 0098-7484. doi:10.1001/jama.285.20.2612. URL http://dx.doi.org/10.1001/jama.285.20.2612. Cited on p. 31.

Bhavnani, Suresh K.; Christopher, Bichakjian K.; Johnson, Timothy M.; Little, Roderick J.; Peck, Frederick A.; Schwartz, Jennifer L.; and Strecher, Victor J. 2003. *Strategy hubs: next-generation domain portals with search procedures.* In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '03, pp. 393–400. ACM, New York, NY, USA. ISBN 1-58113-630-7. doi:10.1145/642611.642680. URL http://dx.doi.org/10.1145/642611.642680. Cited on p. 198.

Bierig, R. and Göker, Ayse. 2006. *Time, location and interest: an empirical and user-centred study.* In Proceedings of the 1st international conference on Information interaction in context, IIiX, pp. 79–87. ACM, New York, NY,

USA. ISBN 1-59593-482-0. doi:10.1145/1164820.1164838. URL http://dx.doi.org/10.1145/1164820.1164838. Cited on pp. 6, 42, and 97.

Bin, L. Jul. 2001. *The retrieval effectiveness of medical information on the web*. In International Journal of Medical Informatics, vol. 62, no. 2-3, pp. 155–163. ISSN 13865056. doi:10.1016/s1386-5056(01)00159-9. URL http://dx.doi.org/10.1016/s1386-5056(01)00159-9. Cited on pp. 69, 70, and 73.

Bing, Lidong; Lam, Wai; and Wong, Tak L. 2011. *Using query log and social tagging to refine queries based on latent topics*. In Proceedings of the 20th ACM international conference on Information and knowledge management, CIKM '11, pp. 583–592. ACM, New York, NY, USA. ISBN 978-1-4503-0717-8. doi:10.1145/2063576.2063663. URL http://dx.doi.org/10.1145/2063576.2063663. Cited on p. 199.

Birru, Mehret S.; Monaco, Valerie M.; Charles, Lonelyss; Drew, Hadiya; Njie, Valerie; Bierria, Timothy; Detlefsen, Ellen; and Steinman, Richard A. Sep. 2004. *Internet usage by low-literacy adults seeking health information: an observational analysis*. In Journal of medical Internet research, vol. 6, no. 3. ISSN 1438-8871. doi:10.2196/jmir.6.3.e25. URL http://dx.doi.org/10.2196/jmir.6.3.e25. Cited on pp. 161, 184, and 190.

Borlund, Pia. Aug. 2003a. *The concept of relevance in IR*. In J. Am. Soc. Inf. Sci., vol. 54, no. 10, pp. 913–925. ISSN 1532-2882. doi:10.1002/asi.10286. URL http://dx.doi.org/10.1002/asi.10286. Cited on pp. 99 and 100.

Borlund, Pia. 2003b. *The IIR evaluation model: a framework for evaluation of interactive information retrieval systems*. In Information Research, vol. 8, no. 3. URL http://informationr.net/ir/8-3/paper152.html. Cited on pp. 13, 59, 65, 66, 119, 123, 155, and 205.

Borlund, Pia and Ingwersen, Peter, eds. Oct. 2006. *Proceedings of the Workshop on Information Interaction in Context (IIiX) Symposium*. Copenhagen, Denmark. Cited on p. 42.

Bosworth, Adam. May 2007. *Putting Health into the Patient's Hands - Consumerism and Health Care*. Opening Plenary Session and Keynote Address. Cited on p. 5.

Brajnik, Giorgio; Mizzaro, Stefano; and Tasso, Carlo. May 2002. *Strategic help in user interfaces for information retrieval*. In J. Am. Soc. Inf. Sci. Technol., vol. 53, no. 5, pp. 343–358. ISSN 1532-2882. doi:10.1002/asi.10035. URL http://dx.doi.org/10.1002/asi.10035. Cited on p. 198.

Brenner, S. H. and Mckinin, E. J. Oct. 1989. *CINAHL and MEDLINE: a comparison of indexing practices*. In Bulletin of the Medical Library Association, vol. 77, no. 4, pp. 366–371. ISSN 0025-7338. URL http://view.ncbi.nlm.nih.gov/pubmed/2676049. Cited on p. 23.

Brézillon, Patrick. May 1999. *Context in problem solving: a survey*. In Knowl. Eng. Rev., vol. 14, no. 1, pp. 47–80. ISSN 0269-8889. doi:10.1017/s0269888999141018. URL http://dx.doi.org/10.1017/s0269888999141018. Cited on p. 37.

Bricon-Souf, Nathalie and Newman, Conrad R. Jan. 2007. *Context awareness in health care: A review*. In International Journal of Medical Informatics, vol. 76, no. 1, pp. 2–12. ISSN 13865056. doi:10.1016/j.ijmedinf.2006.01.003. URL http://dx.doi.org/10.1016/j.ijmedinf.2006.01.003. Cited on pp. x, 40, 41, 42, and 43.

Bush, Vannevar. Jul. 1945. *As We May Think*. In The Atlantic Monthly, vol. 1, no. 176, pp. 101–108. URL http://www.theatlantic.com/doc/194507/bush. Cited on p. 3.

Callan, Jamie; Allan, James; Clarke, Charles L. A.; Dumais, Susan; Evans, David A.; Sanderson, Mark; and Zhai, Chengxiang. Dec. 2007. *Meeting of the MINDS: an information retrieval research agenda*. In SIGIR Forum, vol. 41, no. 2, pp. 25–34. ISSN 0163-5840. doi:10.1145/1328964.1328967. URL http://dx.doi.org/10.1145/1328964.1328967. Cited on p. 4.

Capra, Robert G. and Pérez-Quiñones, Manuel. 2006. *Factors and Evaluation of Refinding Behaviors*. In SIGIR 2006 Workshop on Personal Information Management, pp. 16–19. Cited on p. 150.

Casson, Lionel. 2002. *Libraries in the Ancient World*. Yale University Press. ISBN 9780300097214. URL http://books.google.co.uk/books?id=ECBkVPQkNSsC. Cited on p. 3.

Chahine, C. Abi; Chaignaud, N.; Kotowicz; and Pécuchet, J. P. 2008. *Context and Keyword Extraction in Plain Text Using a Graph Representation*. In Proceedings of the 2008 IEEE International Conference on Signal Image Technology and Internet Based Systems (SITIS '08), pp. 692–696. IEEE Computer Society, Washington, DC, USA. ISBN 978-0-7695-3493-0. URL http://dx.doi.org/10.1109/SITIS.2008.47. Cited on p. 45.

Chignell, Mark H.; Gwizdka, Jacek; and Bodner, Richard C. May 1999. *Discriminating meta-search: a framework for evaluation*. In Inf. Process. Manage., vol. 35, no. 3, pp. 337–362. ISSN 0306-4573. doi:10.1016/s0306-4573(98)00065-x. URL http://dx.doi.org/10.1016/s0306-4573(98)00065-x. Cited on p. 68.

Chitu, Alex. Dec. 2007. *Google Finds Less Search Results*. Available from: http://googlesystem.blogspot.pt/2007/12/google-finds-less-search-results.html [accessed 16 January, 2013]. Cited on p. 249.

Chu, Heting and Rosenthal, Marilyn. Oct. 1996. *Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology*. In Proceedings of the American Society for Information Science Annual Meeting, vol. 33, pp. 127–135. American Society for Information Science, Baltimore, Maryland, USA. Cited on p. 68.

Cilibrasi, Rudi and Vitányi, Paul M. B. 2006. *The Google Similarity Distance*. In Kolmogorov Complexity and Applications. URL http://arxiv.org/abs/cs.CL/0412098. Cited on p. 273.

Cimino, J. J. Jun. 1996. *Linking patient information systems to bibliographic resources*. In Methods of information in medicine, vol. 35, no. 2, pp. 122–126.

ISSN 0026-1270. URL http://view.ncbi.nlm.nih.gov/pubmed/8755385. Cited on p. 49.

Cimino, J. J.; Aguirre, A.; Johnson, S. B.; and Peng, P. Apr. 1993. *Generic queries for meeting clinical information needs.* In Bulletin of the Medical Library Association, vol. 81, no. 2, pp. 195–206. ISSN 0025-7338. URL http://view.ncbi.nlm.nih.gov/pubmed/8472005. Cited on p. 51.

Cimino, J. J.; Elhanan, G.; and Zeng, Q. 1997. *Supporting infobuttons with terminological knowledge.* In Proceedings of the AMIA Annual Fall Symposium., pp. 528–532. ISSN 1091-8280. URL http://view.ncbi.nlm.nih.gov/pubmed/9357682. Cited on p. 51.

Cimino, James J.; Johnson, Stephen B.; Aguirre, Anthony; Roderer, Nancy; and Clayton, Paul D. 1992. *The MEDLINE button.* In Proc Annu Symp Comput Appl Med Care, pp. 81–85. Cited on p. 51.

Cimino, James J. and Li, Jianhua. 2003. *Sharing Infobuttons to Resolve Clinicians' Information Needs.* In AMIA Annu Symp Proc., p. 815. Cited on p. 47.

Cleverdon, Cyril. 1967. *The Cranfield Tests on index language devices.* In Aslib Proceedings, vol. 19, no. 6, pp. 173–194. Cited on p. 4.

Cline, R. J. W. and Haynes, K. M. Dec. 2001. *Consumer health information seeking on the Internet: the state of the art.* In Health education research, vol. 16, no. 6, pp. 671–692. ISSN 0268-1153. doi:10.1093/her/16.6.671. URL http://dx.doi.org/10.1093/her/16.6.671. Cited on pp. 30, 131, and 147.

Coletti, Margaret H. and Bleich, Howard L. Jul. 2001. *Medical Subject Headings Used to Search the Biomedical Literature.* In Journal of the American Medical Informatics Association, vol. 8, no. 4, pp. 317–323. ISSN 1067-5027. doi:10.1136/jamia.2001.0080317. URL http://dx.doi.org/10.1136/jamia.2001.0080317. Cited on p. 22.

Cool, C. Sep. 2002. *Issues of context in information retrieval (IR): an introduction to the special issue.* In Information Processing & Management, vol. 38, no. 5, pp. 605–611. ISSN 03064573. doi:10.1016/s0306-4573(01)00054-1. URL http://dx.doi.org/10.1016/s0306-4573(01)00054-1. Cited on p. 42.

Crain, Stevan P.; Yang, Shuang-Hong; Zha, HongYuan; and Jiao, Yu. 2010. *Dialect Topic Modeling for Improved Consumer Medical Research.* In Proceedings of the AMIA 2010 Annual Symposium. Cited on pp. 148, 149, and 275.

Crestani, Fabio and Ruthven, Ian. Apr. 2007. *Introduction to special issue on contextual information retrieval systems.* In Information Retrieval, vol. 10, no. 2, pp. 111–113. ISSN 1386-4564. doi:10.1007/s10791-007-9022-z. URL http://dx.doi.org/10.1007/s10791-007-9022-z. Cited on p. 42.

Croft, Bruce W. and Harper, David J. 1979. *Using Probabilistic Models of Document Retrieval Without Relevance Information.* In Journal of Documentation, vol. 35, no. 4, pp. 285–295. Cited on p. 4.

Davis, T. C.; Long, S. W.; Jackson, R. H.; Mayeaux, E. J.; George, R. B.; Murphy, P. W.; and Crouch, M. A. Jun. 1993. *Rapid estimate of adult literacy in medicine: a shortened screening instrument.* In Family medicine, vol. 25, no. 6, pp. 391–395. ISSN 0742-3225. URL http://view.ncbi.nlm.nih.gov/pubmed/8349060. Cited on p. 125.

Deng, Yu; Devarakonda, M.; Rajamani, N.; and Zadrozny, W. 2008. *Improving Information Access for a Community of Practice Using Business Process as Context*. In IEEE 24th International Conference on Data Engineering (ICDE 2008)., pp. 1537–1539. doi:10.1109/ICDE.2008.4497615. URL http://dx.doi.org/10.1109/ICDE.2008.4497615. Cited on p. 45.

Dervin, Brenda. 1997. *Given a context by any other name: methodological tools for taming the unruly beast*. In ISIC '96: Proceedings of an international conference on Information seeking in context, pp. 13–38. Taylor Graham Publishing, London, UK, UK. ISBN 0-947568-719. URL http://portal.acm.org/citation.cfm?id=267191. Cited on p. 37.

Dey, A. K. and Abowd, G. D. 2000. *Towards a Better Understanding of Context and Context-Awareness*. In CHI 2000 Workshop on the What, Who, Where, When, and How of Context-Awareness. Cited on pp. x, 38, 39, 40, 41, 42, 43, 44, and 97.

Dhanapal, R. Aug. 2008. *An intelligent information retrieval agent*. In Knowledge-Based Systems, vol. 21, no. 6, pp. 466–470. doi:10.1016/j.knosys.2008.03.010. URL http://dx.doi.org/10.1016/j.knosys.2008.03.010. Cited on p. 45.

Doan, Bich L.; Jose, Joemon; and Melucci, Massimo, eds. 2007. *Proceedings of the Workshop CIR-07: Context-Based Information Retrieval at the Sixth International and Interdisciplinary Conference on Modeling and Using Context*. Roskilde University, Denmark. Cited on p. 43.

Doan, Bich-Liên; Jose, Joemon; Melucci, Massimo; and Tamine-Lechani, Lynda, eds. Apr. 2009. *Contextual Information Access, Seeking and Retrieval Evaluation*. URL http://www.irit.fr/CIRSE09/000_ECIR%202009%20Workshop_proceedings.pdf. Cited on p. 42.

Doan, Bich-Liên; Jose, Joemon; Melucci, Massimo; and Tamine-Lechani, Lynda, eds. Mar. 2010. *Proceedings of the 2nd International Workshop on Contextual Information Access, Seeking and Retrieval Evaluation*. URL http://ceur-ws.org/Vol-569/cirse2010-proceedings.pdf. Cited on p. 42.

Doms, Andreas and Schroeder, Michael. Jul. 2005. *GoPubMed: exploring PubMed with the Gene Ontology.* In Nucleic acids research, vol. 33, no. Web Server issue, pp. W783–W786. ISSN 1362-4962. doi:10.1093/nar/gki470. URL http://dx.doi.org/10.1093/nar/gki470. Cited on p. 54.

Dourish, Paul. Feb. 2004. *What we talk about when we talk about context*. In Personal Ubiquitous Comput., vol. 8, no. 1, pp. 19–30. ISSN 1617-4909. doi:10.1007/s00779-003-0253-8. URL http://dx.doi.org/10.1007/s00779-003-0253-8. Cited on pp. 37, 38, 54, and 97.

Douyere, Magaly; Soualmia, Lina F.; Neveol, Aurelie; Rogozan, Alexandrina; Dahamna, Badisse; Leroy, Jean-Philippe; Thirion, Benoit; and Darmoni, Stefan J. 2004. *Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway*. In Health Information and Libraries Journal, vol. 21, no. 4, pp. 253–261. ISSN 1471-1834. doi: 10.1111/j.1471-1842.2004.00526.x. URL http://dx.doi.org/10.1111/j.1471-1842.2004.00526.x. Cited on pp. 31 and 34.

Dung, Tran and Kameyama, Wataru. 2007. *Ontology-Based Information Extraction and Information Retrieval in Health Care Domain*. In Data Warehousing and Knowledge Discovery, pp. 323–333. doi:10.1007/978-3-540-74553-2_30. URL http://dx.doi.org/10.1007/978-3-540-74553-2_30. Cited on p. 34.

Eerola, Johanna and Vakkari, Pertti. 2008. *How a general and a specific thesaurus cover expressions in patients' questions and physicians' answers*. In Journal of Documentation, vol. 64, no. 1, pp. 131–142. Cited on p. 148.

Efthimiadis, Efthimis N. 1996. *Query Expansion*. In Annual Review of Information Systems and Technology (ARIST), vol. 31, pp. 121–187. Cited on pp. 44, 199, and 200.

Eichmann, David; Ruiz, Miguel E.; and Srinivasan, Padmini. 1998. *Cross-language information retrieval with the UMLS metathesaurus*. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '98, pp. 72–80. ACM, New York, NY, USA. ISBN 1-58113-015-5. doi:10.1145/290941.290959. URL http://dx.doi.org/10.1145/290941.290959. Cited on p. 132.

Espanha, Rita; Mendes, Rita V.; Fonseca, Rui B.; and Correia, Tiago. 2012. *Os portugueses, a saúde e a internet*. Fundação Calouste Gulbenkian. Cited on p. 5.

Espanha, Rita and Lupiáñez Villanueva, Francisco. 2008. *Health and the Internet: Autonomy of the User*. Tech. rep., LINI - Lisbon Internet and Networks. Cited on p. 5.

Eysenbach, G. and Kohler. 2003. *What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the internet*. In AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, pp. 225–229. ISSN 1942-597X. URL http://view.ncbi.nlm.nih.gov/pubmed/14728167. Cited on pp. 31, 247, 248, 259, 260, and 273.

Eysenbach, Gunther; Powell, John; Kuss, Oliver; and Sa, Eun-Ryoung R. Feb-Sep 2002. *Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review*. In JAMA : the journal of the American Medical Association, vol. 287, no. 20, pp. 2691–2700. ISSN 0098-7484. doi:10.1001/jama.287.20.2691. URL http://dx.doi.org/10.1001/jama.287.20.2691. Cited on p. 87.

Eysenbach, Gunther and Thomson, Maria. 2007. *The FA4CT algorithm: a new model and tool for consumers to assess and filter health information on the*

Internet. In MEDINFO 2007 - Proceedings of the 12th World Congress on Health (Medical) Informatics – Building Sustainable Health Systems, vol. 129, pp. 142–146. IOS Press, Amsterdam, The Netherlands. Cited on p. 70.

Fattahi, Rahmatollah; Wilson, Concepción S.; and Cole, Fletcher. Jul. 2008. *An alternative approach to natural language query expansion in search engines: Text analysis of non-topical terms in Web documents*. In Inf. Process. Manage., vol. 44, no. 4, pp. 1503–1516. ISSN 0306-4573. doi:10.1016/j.ipm.2007.09.009. URL http://dx.doi.org/10.1016/j.ipm.2007.09.009. Cited on pp. 33 and 200.

Fawcett, Tom. Jun. 2006. *An introduction to ROC analysis*. In Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874. ISSN 01678655. doi:10.1016/j.patrec.2005.10.010. URL http://dx.doi.org/10.1016/j.patrec.2005.10.010. Cited on p. 254.

Ferguson, Tom. 2007. *e-patients - how they can help us save health care*. Tech. rep., e-patients scholars working group. Cited on p. 6.

Fernández-Luna, Juan M.; Huete, Juan F.; and Castells, Pablo. 2011. *Call for Papers for Special Issue on Personalization and Recommendation in Information Access*. URL http://www.elsevierscitech.com/dronsite/cfp/ipm-cfp.pdf. Cited on p. 42.

Finkelstein, L. E. V.; Gabrilovich, Evgeniy; Matias, Yossi; Rivlin, E. H. U. D.; Solan, Z. A. C. H.; Wolfman, G. A. D. I.; and Ruppin, Eytan. Jan. 2002. *Placing search in context: the concept revisited*. In ACM Trans. Inf. Syst., vol. 20, no. 1, pp. 116–131. ISSN 1046-8188. doi:10.1145/503104.503110. URL http://dx.doi.org/10.1145/503104.503110. Cited on p. 37.

Fitzsimmons, P. R.; Michael, B. D.; Hulley, J. L.; and Scott, G. O. Dec. 2010. *A readability assessment of online Parkinson's disease information*. In The Journal of the Royal College of Physicians of Edinburgh, vol. 40, no. 4, pp. 292–296. ISSN 2042-8189. doi:10.4997/jrcpe.2010.401. URL http://dx.doi.org/10.4997/jrcpe.2010.401. Cited on p. 124.

Fonseca, Luis C.; Menezes, Credine S.; Vicari, Rosa; and Soares, Jonatas. 2008. *An intelligent and contextual information retrieval environment for Lifelong Learning*. In 38th Annual Frontiers in Education Conference (FIE 2008)., pp. S4G–20–S4G–24. doi:10.1109/FIE.2008.4720654. URL http://dx.doi.org/10.1109/FIE.2008.4720654. Cited on p. 45.

Fox, S. Oct. 2006. *Online Health Search 2006*. Tech. rep., Pew Internet & American Life Project, Washington, USA. URL http://www.pewinternet.org/~{}/media//Files/Reports/2006/PIP_Online_Health_2006.pdf.pdf. Cited on pp. 29, 87, 95, 119, and 147.

Fox, S. and Rainie, L. 2000. *The online health care revolution: How the Web helps Americans take better care of themselves*. Tech. rep., The Pew Internet & American Life Project. URL http://www.pewinternet.org/pdfs/PIP_health_report.pdf. Cited on p. 29.

Fox, Susannah. 2011. *Health Topics*. Tech. rep., Pew Internet & American Life Project. Cited on pp. 5, 119, and 197.

Fox, Susannah and Duggan, Maeve. Nov. 2012. *Mobile Health 2012*. Tech. rep., Pew Internet & American Life Project. Cited on p. 5.

Fox, Susannah and Duggan, Maeve. Jan. 2013. *Health Online 2013*. Tech. rep., Pew Research Center's Internet & American Life Project, Washington, D.C. Cited on pp. 5, 17, 29, 65, and 82.

Fox, Susannah and Jones, Sydney. Jun. 2009. *The Social Life of Health Information*. Tech. rep., Pew Internet & American Life Project, Washington, USA. URL http://www.pewinternet.org/~{}/media//Files/Reports/2009/PIP_Health_2009.pdf. Cited on p. 87.

Freund, Luanne and Butterworth, Richard. 2008. *Tagging for use: an analysis of use-centred resource description*. In IIiX '08: Proceedings of the second international symposium on Information interaction in context, pp. 6–12. ACM, New York, NY, USA. ISBN 978-1-60558-310-5. doi:10.1145/1414694.1414699. URL http://dx.doi.org/10.1145/1414694.1414699. Cited on p. 45.

Frisse, Mark E. 1987. *Searching for information in a hypertext medical handbook*. In HYPERTEXT '87: Proceedings of the ACM conference on Hypertext, pp. 57–66. ACM, New York, NY, USA. ISBN 0-89791-340-X. doi:10.1145/317426.317433. URL http://dx.doi.org/10.1145/317426.317433. Cited on pp. 49 and 54.

Gao, Wei; Niu, Cheng; Nie, Jian Y.; Zhou, Ming; Wong, Kam F.; and Hon, Hsiao W. 2010. *Exploiting query logs for cross-lingual query suggestions*. In ACM Trans. Inf. Syst., vol. 28, no. 2, pp. 1–33. ISSN 1046-8188. doi:10.1145/1740592.1740594. URL http://dx.doi.org/10.1145/1740592.1740594. Cited on p. 200.

Gay, Clifford W.; Kayaalp, Mehmet; and Aronson, Alan R. 2005. *Semi-automatic indexing of full text biomedical articles*. In AMIA Annual Symposium Proceedings, pp. 271–275. ISSN 1942-597X. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1560666/. Cited on p. 49.

Gigerenzer, Gerd. Mar. 2003. *Calculated Risks: How to Know When Numbers Deceive You*. Simon & Schuster, 1st edn. ISBN 0743254236. URL http://www.worldcat.org/isbn/0743254236. Cited on p. 4.

Göker, Ayse and Myrhaug, Hans I. 2002. *User Context and Personalisation*. In ECCBR Workshop on Case Based Reasoning and Personalisation. Cited on pp. x, 38, 40, and 42.

González-González, A. I.; Dawes, M.; Sánchez-Mateos, J.; Riesgo-Fuertes, R.; Escortell-Mayor, E.; Sanz-Cuesta, T.; and Hernández-Fernández, T. Jul. 2007. *Information Needs and Information-Seeking Behavior of Primary Care Physicians*. In Ann Fam Med, vol. 5, no. 4, pp. 345–352. ISSN 1544-1717. doi:10.1370/afm.681. URL http://dx.doi.org/10.1370/afm.681. Cited on p. 30.

Google. Oct. 2012. *Google search result count*. Available from: http://support.google.com/webmasters/bin/answer.py?hl=en&answer=70920. Accessed: 2013-01-16. (Archived by WebCite at http://www.webcitation.org/6DiyZou3I). Cited on p. 259.

Gordon, Michael and Pathak, Praveen. Mar. 1999. *Finding information on the World Wide Web: the retrieval effectiveness of search engines*. In Information Processing & Management, vol. 35, no. 2, pp. 141–180. ISSN 03064573. doi:10.1016/s0306-4573(98)00041-7. URL http://dx.doi.org/10.1016/s0306-4573(98)00041-7. Cited on p. 68.

Graber, M. A.; Bergus, G. R.; and York, C. Jul. 1999. *Using the World Wide Web to answer clinical questions: how efficient are different methods of information retrieval?* In The Journal of Family Practice, vol. 48, no. 7, pp. 520–524. ISSN 0094-3509. URL http://view.ncbi.nlm.nih.gov/pubmed/10428249. Cited on pp. 69 and 72.

Graham, L.; Tse, T.; and Keselman, A. 2006. *Exploring user navigation during online health information seeking*. In AMIA Annu Symp Proc. 2006, pp. 299–303. Lister Hill Center, National Library of Medicine, NIH, DHHS, Bethesda, MD, USA. ISSN 1559-4076. URL http://view.ncbi.nlm.nih.gov/pubmed/17238351. Cited on p. 30.

Grefenstette, G. and Nioche, J. Apr. 2000. *Estimation of English and non-English language use on the WWW*. In Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) - RIAO 2000, 6th International Conference., pp. 237–246. CID, France. Cited on p. 131.

Gwizdka, J. and Chignell, M. H. 1999. *Towards information retrieval measures for evaluation of web search engines*. Tech. rep., Interactive Media Lab, Toronto, Ontario, Canada. URL http://peach.mie.utoronto.ca/~{}jacekg/pubs/webIR_eval1_99.pdf. Cited on p. 68.

Gyllstrom, Karl and Soules, Craig. Jan. 2008. *Seeing is retrieving: Building information context from what the user sees*. In Intelligent User Interfaces, pp. 189–198. Cited on p. 45.

Gyllstrom, Karl; Soules, Craig; and Veitch, Alistair. 2008. *Activity put in context: identifying implicit task context within the user's document interaction*. In IIiX '08: Proceedings of the second international symposium on Information interaction in context, pp. 51–56. ACM, New York, NY, USA. ISBN 978-1-60558-310-5. doi:10.1145/1414694.1414707. URL http://dx.doi.org/10.1145/1414694.1414707. Cited on p. 45.

Halkias, Daphne; Harkiolakis, Nicholas; Thurman, Paul; and Caracatsanis, Sylva. 2007. *Internet usage for health-related purposes among Greek consumers*. In ICCOMP'07: Proceedings of the 11th WSEAS International Conference on Computers, pp. 281–289. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA. URL http://portal.acm.org/citation.cfm?id=1353956.1354006. Cited on p. 29.

Harper, David J. and Kelly, Diane. 2006. *Contextual relevance feedback*. In IIiX: Proceedings of the 1st international conference on Information interaction in context, pp. 129–137. ACM Press, New York, NY, USA. ISBN 1595934820. doi:10.1145/1164820.1164847. URL http://dx.doi.org/10.1145/1164820.1164847. Cited on pp. 6, 54, and 97.

Hashmi, Z.; Zrimec, T.; and Hopkins, A. Oct. 2009. *CQGF: Context specific query generation framework from computerized clinical practice guidelines*. In Second International Conference on the Applications of Digital Information and Web Technologies - ICADIWT '09, pp. 288–293. doi: 10.1109/icadiwt.2009.5273854. URL http://dx.doi.org/10.1109/icadiwt.2009.5273854. Cited on p. 52.

Haux, R.; Ammenwerth, E.; Herzog, W.; and Knaup, P. Nov. 2002. *Health care in the information society - A prognosis for the year 2013*. In International journal of medical informatics, vol. 66, no. 1-3, pp. 3–21. ISSN 1386-5056. URL http://view.ncbi.nlm.nih.gov/pubmed/12453552. Cited on p. 4.

Hawking, David; Craswell, Nick; Bailey, Peter; and Griffihs, Kathleen. Apr. 2001. *Measuring Search Engine Quality*. In Information Retrieval, vol. 4, no. 1, pp. 33–59. ISSN 13864564. doi:10.1023/a:1011468107287. URL http://dx.doi.org/10.1023/a:1011468107287. Cited on p. 68.

Hearst, Marti A.; Divoli, Anna; Guturu, Harendra; Ksikes, Alex; Nakov, Preslav; Wooldridge, Michael A.; and Ye, Jerry. Aug. 2007a. *BioText Search Engine: beyond abstract search*. In Bioinformatics (Oxford, England), vol. 23, no. 16, pp. 2196–2197. ISSN 1367-4811. doi:10.1093/bioinformatics/btm301. URL http://dx.doi.org/10.1093/bioinformatics/btm301. Cited on p. 53.

Hearst, Marti A.; Divoli, Anna; Ye, Jerry; and Wooldridge, Michael A. 2007b. *Exploring the efficacy of caption search for bioscience journal search interfaces*. In BioNLP '07: Proceedings of the Workshop on BioNLP 2007, pp. 73–80. Association for Computational Linguistics, Morristown, NJ, USA. URL http://portal.acm.org/citation.cfm?id=1572392.1572406. Cited on p. 53.

Hemminger, Bradley M.; Lu, Dihui; Vaughan, K. T. L.; and Adams, Stephanie J. Dec. 2007. *Information seeking behavior of academic scientists*. In J. Am. Soc. Inf. Sci. Technol., vol. 58, no. 14, pp. 2205–2225. ISSN 1532-2882. doi: 10.1002/asi.v58:14. URL http://dx.doi.org/10.1002/asi.v58:14. Cited on p. 30.

Hersh, W.; Price, S.; and Donohoe, L. 2000. *Assessing thesaurus-based query expansion using the UMLS Metathesaurus*. In AMIA Annual Symposium Proceedings, pp. 344–348. ISSN 1531-605X. URL http://view.ncbi.nlm.nih.gov/pubmed/11079902. Cited on p. 53.

Hersh, W. R. and Donohoe, L. C. 1998. *SAPHIRE International: a tool for cross-language information retrieval*. In AMIA Annual Symposium proceedings, pp. 673–677. AMIA, USA. ISSN 1531-605X. URL http://view.ncbi.nlm.nih.gov/pubmed/9929304. Cited on p. 132.

Hersh, William. Nov. 2008a. *Information Retrieval: A Health and Biomedical Perspective (Health Informatics)*. Springer, New York, NY, USA, 3rd edn. ISBN 038778702X. URL http://www.worldcat.org/isbn/038778702X. Cited on pp. 51, 60, 65, 68, 131, and 132.

Hersh, William. Dec. 2008b. *Ubiquitous but unfinished: grand challenges for information retrieval*. In Health Information & Libraries Journal, vol. 25,

no. s1, pp. 90–93. ISSN 1471-1842. doi:10.1111/j.1471-1842.2008.00815.x. URL http://dx.doi.org/10.1111/j.1471-1842.2008.00815.x. Cited on p. 6.

Hersh, William R. Dec. 2002. *Information Retrieval - A Health and Biomedical Perspective*. Springer. ISBN 0387955224. URL http://www.worldcat.org/isbn/0387955224. Cited on pp. 17, 18, 19, and 23.

Hersh, William R. and Hickam, David H. Oct. 1998. *How Well Do Physicians Use Electronic Information Retrieval Systems?: A Framework for Investigation and Systematic Review*. In JAMA, vol. 280, no. 15, pp. 1347–1352. ISSN 0098-7484. doi:10.1001/jama.280.15.1347. URL http://dx.doi.org/10.1001/jama.280.15.1347. Cited on p. 30.

Hersh, William R.; Muller, Henning; Jensen, Jeffery R.; Yang, Jianji; Gorman, Paul N.; and Ruch, Patrick. Sep. 2006. *Advancing Biomedical Image Retrieval: Development and Analysis of a Test Collection*. In J Am Med Inform Assoc, vol. 13, no. 5, pp. 488–496. doi:10.1197/jamia.m2082. URL http://dx.doi.org/10.1197/jamia.m2082. Cited on p. 34.

Hliaoutakis, Angelos; Varelas, Giannis; Petrakis, Euripides; and Milios, Evangelos. 2006a. *MedSearch: A Retrieval System for Medical Information Based on Semantic Similarity*. In Research and Advanced Technology for Digital Libraries, vol. 4172/2006, pp. 512–515. doi:10.1007/11863878_56. URL http://dx.doi.org/10.1007/11863878_56. Cited on p. 34.

Hliaoutakis, Angelos; Zervanou, Kalliopi; Petrakis, Euripides G. M.; and Milios, Evangelos E. 2006b. *Automatic document indexing in large medical collections*. In HIKM '06: Proceedings of the international workshop on Healthcare information and knowledge management, pp. 1–8. ACM Press, New York, NY, USA. ISBN 1595935282. doi:10.1145/1183568.1183570. URL http://dx.doi.org/10.1145/1183568.1183570. Cited on p. 31.

Houston, Andrea L.; Chen, Hsinchun; Schatz, Bruce R.; Hubbard, Susan M.; Sewell, Robin R.; and Ng, Tobun D. Dec. 2000. *Exploring the use of concept spaces to improve medical information retrieval.* In Decis. Support Syst., vol. 30, no. 2, pp. 171–186. ISSN 0167-9236. URL http://portal.acm.org/citation.cfm?id=364616.364626. Cited on p. 31.

Humphreys, B. L.; Lindberg, D. A.; Schoolman, H. M.; and Barnett, G. O. 1998. *The Unified Medical Language System: an informatics research collaboration.* In Journal of the American Medical Informatics Association : JAMIA, vol. 5, no. 1, pp. 1–11. ISSN 1067-5027. URL http://www.jamia.org/cgi/content/abstract/5/1/1. Cited on p. 24.

Humphreys, Betsy L. and Schuyler, Peri L. 1993. *The Unified Medical Language System: moving beyond the vocabulary of bibliographic retrieval.* In High performance medical libraries: Advances in information management for the virtual era, pp. 31–44. URL http://portal.acm.org/citation.cfm?id=165920.165928. Cited on p. 25.

Hung, Peter W.; Johnson, Stephen B.; Kaufman, David R.; and Mendonça, Eneida A. Apr. 2008. *A multi-level model of information seeking in the clinical domain*. In Journal of Biomedical Informatics, vol. 41, no. 2, pp. 357–370.

ISSN 15320464. doi:10.1016/j.jbi.2007.09.005. URL http://dx.doi.org/10.1016/j.jbi.2007.09.005. Cited on p. 34.

Huuskonen, S. and Vakkari, P. 2008. *Students' search process and outcome in Medline in writing an essay for a class on evidence-based medicine*. In Journal of Documentation, vol. 64, no. 2, pp. 287–303. doi:10.1108/00220410810858065. URL http://dx.doi.org/10.1108/00220410810858065. Cited on p. 29.

Ide, N. C.; Loane, R. F.; and Demner-Fushman, D. 2007. *Essie: a concept-based search engine for structured biomedical text.* In Journal of the American Medical Informatics Association : JAMIA, vol. 14, no. 3, pp. 253–263. ISSN 1067-5027. doi:10.1197/jamia.m2233. URL http://dx.doi.org/10.1197/jamia.m2233. Cited on p. 33.

Ilic, D.; Bessell, T. L.; Silagy, C. A.; and Green, S. Mar. 2003. *Specialized medical search-engines are no better than general search-engines in sourcing consumer information about androgen deficiency*. In Human Reproduction, vol. 18, no. 3, pp. 557–561. doi:10.1093/humrep/deg154. URL http://dx.doi.org/10.1093/humrep/deg154. Cited on pp. 70 and 73.

Ingwersen, P.; Jelin, K.; and Belkin, N., eds. Aug. 2005. *Proceedings of the ACM SIGIR 2005 Workshop on Information Retrieval in Context (IRiX)*. Royal School of Library and Information Science. Denmark. Cited on pp. 6, 42, and 97.

Ingwersen, P.; Rijsbergen, K.; and Belkin, N., eds. 2004. *Proceedings of the ACM SIGIR 2004 Workshop on Information Retrieval in Context (IRiX)*. Sheffield, UK. Cited on p. 42.

Ingwersen, Peter. 2006. *Context in information interaction revisited 2006*. Presentation at Prolissa 2006: Proceedings of the Fourth Biennial DISSAnet Conference. Cited on p. 38.

Ingwersen, Peter. 2009. *The User in Interactive Information Retrieval Evaluation*. Presentation. Cited on pp. 13 and 66.

Ingwersen, Peter and Järvelin, Kalervo. Sep. 2005. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer, Dordrecht, The Netherlands, 1st edn. ISBN 140203850X. URL http://www.worldcat.org/isbn/140203850X. Cited on pp. 38, 41, 43, 45, 66, 98, 99, and 100.

Inskip, Charlie; Macfarlane, Andy; and Rafferty, Pauline. 2008. *Content or context?: searching for musical meaning in task-based interactive information retrieval*. In IIiX '08: Proceedings of the second international symposium on Information interaction in context, pp. 72–74. ACM, New York, NY, USA. ISBN 978-1-60558-310-5. doi:10.1145/1414694.1414711. URL http://dx.doi.org/10.1145/1414694.1414711. Cited on p. 45.

Jansen, Bernard J. Jul. 2005. *Seeking and implementing automated assistance during the search process*. In Information Processing & Management, vol. 41, no. 4, pp. 909–928. ISSN 03064573. doi:10.1016/j.ipm.2004.04.017. URL http://dx.doi.org/10.1016/j.ipm.2004.04.017. Cited on p. 198.

Jansen, Bernard J. and McNeese, Michael D. 2005. *Evaluating the effectiveness of and patterns of interactions with automated searching assistance*. In J. Am. Soc. Inf. Sci., vol. 56, no. 14, pp. 1480–1503. doi:10.1002/asi.20242. URL http://dx.doi.org/10.1002/asi.20242. Cited on pp. 198 and 229.

Jansen, Bernard J. and Pooch, Udo. Feb. 2001. *A review of web searching studies and a framework for future research*. In J. Am. Soc. Inf. Sci. Technol., vol. 52, no. 3, pp. 235–246. ISSN 1532-2882. doi:10.1002/1097-4571(2000)9999:9999\%3C::aid-asi1607\%3E3.3.co;2-6. URL http://dx.doi.org/10.1002/1097-4571(2000)9999:9999%3C::aid-asi1607%3E3.3.co;2-6. Cited on pp. 98, 102, and 114.

Järvelin, Kalervo and Kekäläinen, Jaana. Oct. 2002. *Cumulated gain-based evaluation of IR techniques*. In ACM Transactions on Information Systems, vol. 20, no. 4, pp. 422–446. ISSN 1046-8188. doi:10.1145/582415.582418. URL http://dx.doi.org/10.1145/582415.582418. Cited on p. 67.

Johnson. Sep. 2003. *On contexts of information seeking*. In Information Processing & Management, vol. 39, no. 5, pp. 735–760. ISSN 03064573. doi:10.1016/s0306-4573(02)00030-4. URL http://dx.doi.org/10.1016/s0306-4573(02)00030-4. Cited on p. 38.

Johnson, Pamela T.; Chen, Jennifer K.; Eng, John; Makary, Martin A.; and Fishman, Elliot K. Sep. 2008. *A comparison of World Wide Web resources for identifying medical information*. In Academic Radiology, vol. 15, no. 9, pp. 1165–1172. ISSN 1076-6332. doi:10.1016/j.acra.2008.02.010. URL http://dx.doi.org/10.1016/j.acra.2008.02.010. Cited on p. 69.

Joho, Hideo; Hopfgartner, Frank; Jose, Joemon M.; and van Rijsbergen, C. J. 2009. *AIR 2008: second international workshop on adaptive information retrieval*. In SIGIR Forum, vol. 43, no. 1, pp. 63–65. ISSN 0163-5840. doi:10.1145/1670598.1670611. URL http://dx.doi.org/10.1145/1670598.1670611. Cited on p. 42.

Jones, Dee and Timm, Donna F. 2008. *Consumer Health Search Engines Comparison*. In Journal of Hospital Librarianship, vol. 8, no. 4, pp. 418–432. doi:10.1080/15323260802391936. URL http://dx.doi.org/10.1080/15323260802391936. Cited on pp. 69, 70, and 73.

Jones, Karen S. 1972. *A statistical interpretation of term specificity and its application in retrieval*. In Journal of Documentation, vol. 28, no. 1, pp. 11–21. Cited on p. 4.

Jones, Karen S. 2006. *What's the value of TREC: is there a gap to jump or a chasm to bridge?* In SIGIR Forum, vol. 40, no. 1, pp. 10–20. ISSN 0163-5840. doi:10.1145/1147197.1147198. URL http://dx.doi.org/10.1145/1147197.1147198. Cited on p. 66.

Jose, J.; Joho, H.; and Vanrijsbergen, C. Nov. 2008. *Adaptive information retrieval: Introduction to the special topic issue of information processing and management*. In Information Processing & Management, vol. 44, no. 6, pp. 1819–1821. ISSN 03064573. doi:10.1016/j.ipm.2008.08.002. URL http://dx.doi.org/10.1016/j.ipm.2008.08.002. Cited on p. 42.

Jydstrup, R. A. and Gross, M. J. 1966. *Cost of information handling in hospitals.* In Health services research, vol. 1, no. 3, pp. 235–271. ISSN 0017-9124. URL http://view.ncbi.nlm.nih.gov/pubmed/5971636. Cited on p. 17.

Kelly, Diane. 2009. *Methods for Evaluating Interactive Information Retrieval Systems with Users*. In Foundations and Trends® in Information Retrieval, vol. 3, no. 1-2, pp. 1–224. Cited on p. 66.

Kelly, Diane and Cool, Colleen. 2002. *The effects of topic familiarity on information search behavior*. In Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, JCDL '02, pp. 74–75. ACM, New York, NY, USA. ISBN 1-58113-513-0. doi:10.1145/544220.544232. URL http://dx.doi.org/10.1145/544220.544232. Cited on pp. 151 and 152.

Kelly, Liadh; Chen, Yi; Fuller, Marguerite; and Jones, Gareth J. F. 2008. *A study of remembered context for information access from personal digital archives*. In Proceedings of the second international symposium on Information interaction in context, IIiX '08, pp. 44–50. ACM, New York, NY, USA. ISBN 978-1-60558-310-5. doi:10.1145/1414694.1414706. URL http://dx.doi.org/10.1145/1414694.1414706. Cited on p. 45.

Keselman, Alla; Massengale, Lisa; Ngo, Long; Browne, Allen; and Zeng, Qing. 2006. *The Effect of User Factors on Consumer Familiarity with Health Terms: Using Gender as a Proxy for Background Knowledge About Gender-Specific Illnesses*. In Biological and Medical Data Analysis, pp. 472–481. doi:10.1007/11946465_43. URL http://dx.doi.org/10.1007/11946465_43. Cited on pp. 32 and 149.

Kingsland, L. C.; Harbourt, A. M.; Syed, E. J.; and Schuyler, P. L. Apr. 1993. *Coach: applying UMLS knowledge sources in an expert searcher environment*. In Bull Med Libr Assoc., vol. 81, no. 2, pp. 178–183. Cited on p. 52.

Kipp, Margaret E. I. 2007. *Tagging for health information organisation and retrieval*. In JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, p. 485. ACM, New York, NY, USA. ISBN 978-1-59593-644-8. doi:10.1145/1255175.1255284. URL http://dx.doi.org/10.1145/1255175.1255284. Cited on p. 31.

Kleinsorge, Rachel and Willis, Jan. Sep. 2007. *Unified Medical Language System (UMLS) Basics*. Presentation. URL http://umlsinfo.nlm.nih.gov/UMLS_Basics.pdf. Cited on pp. 24 and 25.

Knight, Dallas; Holt, Alec; and Warren, Jim. 2009. *Search engines: a study of nine search engines in four categories*. In Journal of Health Informatics in Developing Countries, vol. 3, no. 1, pp. 1–8. Cited on pp. 70 and 73.

Kodagoda, Neesha and Wong, B. L. William. 2008. *Effects of low & high literacy on user performance in information search and retrieval*. In Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 1, BCS-HCI '08, pp. 173–181. British Computer Society, Swinton, UK, UK. ISBN 978-1-906124-04-5. URL http://portal.acm.org/citation.cfm?id=1531514.1531538. Cited on p. 184.

Kodagoda, Neesha; Wong, B. L. William; Rooney, Chris; and Khan, Nawaz. 2012. *Interactive visualization for low literacy users: from lessons learnt to design*. In Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems, CHI '12, pp. 1159–1168. ACM, New York, NY, USA. ISBN 978-1-4503-1015-4. doi:10.1145/2208516.2208565. URL http://dx.doi.org/10.1145/2208516.2208565. Cited on pp. 185 and 192.

Kogan, S.; Zeng, Q.; Ash, N.; and Greenes, R. A. 2001. *Problems and challenges in patient information retrieval: a descriptive study.* In Proceedings / AMIA ... Annual Symposium. AMIA Symposium, pp. 329–333. ISSN 1531-605X. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243602/. Cited on pp. 147 and 197.

Kriewel, Sascha and Fuhr, Norbert. 2010. *Evaluation of an adaptive search suggestion system*. In 32nd European Conference on Information Retrieval Research (ECIR 2010), pp. 544–555. Springer. Cited on pp. 197 and 199.

Kumar, Aseem. May 2005. *Health Search Tool Evaluation*. Master's thesis, Oregon Health & Science University. Cited on pp. 69, 70, and 73.

Kumaran, Giridhar and Allan, James. 2008. *Adapting information retrieval systems to user queries*. In Inf. Process. Manage., vol. 44, no. 6, pp. 1838–1862. ISSN 0306-4573. doi:10.1016/j.ipm.2007.12.006. URL http://dx.doi.org/10.1016/j.ipm.2007.12.006. Cited on p. 45.

Kumaran, Giridhar; Jones, Rosie; and Madani, Omid. 2005. *Biasing web search results for topic familiarity*. In Proceedings of the 14th ACM international conference on Information and knowledge management, CIKM '05, pp. 271–272. ACM, New York, NY, USA. ISBN 1-59593-140-6. doi:10.1145/1099554.1099622. URL http://dx.doi.org/10.1145/1099554.1099622. Cited on p. 152.

Kushniruk, A. W.; Kan, M. Y.; Mckeown, K.; Klavans, J.; Jordan, D.; Laflamme, M.; and Patel, V. L. 2002. *Usability evaluation of an experimental text summarization system and three search engines: implications for the reengineering of health care interfaces.* In AMIA Annu Symp Proc. 2002, pp. 420–424. Department of Mathematics and Statistics, York University, Toronto, Ontario M3J 1P3, Canada. ISSN 1531-605X. URL http://view.ncbi.nlm.nih.gov/pubmed/12463858. Cited on p. 34.

Kutner, Mark; Greenberg, Elizabeth; Jin, Ying; and Paulsen, Christine. Sep. 2006. *The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy*. Tech. rep., National Center for Education Statistics. Cited on pp. 9, 127, 150, 208, and 273.

Lambert, Sylvie D. and Loiselle, Carmen G. Oct. 2007. *Health Information Seeking Behavior*. In Qual Health Res, vol. 17, no. 8, pp. 1006–1019. ISSN 1049-7323. doi:10.1177/1049732307305199. URL http://dx.doi.org/10.1177/1049732307305199. Cited on p. 30.

Lee, Shoou-Yih Y.; Bender, Deborah E.; Ruiz, Rafael E.; and Cho, Young Ik I. Aug. 2006. *Development of an easy-to-use Spanish Health Literacy test.* In Health services research, vol. 41, no. 4 Pt 1, pp. 1392–1412. ISSN

0017-9124. doi:10.1111/j.1475-6773.2006.00532.x. URL http://dx.doi.org/10.1111/j.1475-6773.2006.00532.x. Cited on p. 125.

Leroy, G.; Eryilmaz, E.; and Laroya, B. T. 2006. *Health information text characteristics*. In AMIA Annu Symp Proc. 2006, pp. 479–483. Claremont Graduate University, Claremont, California, USA. ISSN 1942-597X. URL http://view.ncbi.nlm.nih.gov/pubmed/17238387. Cited on p. 32.

Leroy, G.; Miller, T.; Rosemblat, G.; and Browne, A. Jul. 2008. *A balanced approach to health information evaluation: A vocabulary-based naïve Bayes classifier and readability formulas*. In Journal of the American Society for Information Science and Technology, vol. 59, no. 9. doi:10.1002/asi.20837. URL http://www3.interscience.wiley.com/journal/118903525/abstract?CRETRY=1&SRETRY=0. Cited on p. 32.

Lewandowski, Dirk. 2008. *The retrieval effectiveness of web search engines: considering results descriptions*. In Journal of Documentation, vol. 64, no. 6, pp. 915–937. ISSN 0022-0418. Cited on p. 68.

Lewis, M. Paul, ed. 2009. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 16th edn. URL http://www.ethnologue.com/. Cited on p. 131.

Li, Yuelin and Belkin, Nicholas J. Nov. 2008. *A faceted approach to conceptualizing tasks in information seeking*. In Information Processing & Management, vol. 44, no. 6, pp. 1822–1837. ISSN 03064573. doi:10.1016/j.ipm.2008.07.005. URL http://dx.doi.org/10.1016/j.ipm.2008.07.005. Cited on p. 45.

Lin, Jimmy and Fushman, Dina D. 2005. *Representation of Information Needs and the Elements of Context: A Case Study in the Domain of Clinical Medicine*. In ACM SIGIR 2005 Workshop on Information Retrieval in Context (IRiX). Cited on pp. 45, 49, and 97.

Lindberg, D. A.; Humphreys, B. L.; and Mccray, A. T. Aug. 1993. *The Unified Medical Language System*. In Methods of information in medicine, vol. 32, no. 4, pp. 281–291. ISSN 0026-1270. URL http://view.ncbi.nlm.nih.gov/pubmed/8412823. Cited on p. 24.

Liu; Zhenyu; Chu; and Wesley. Apr. 2007. *Knowledge-based query expansion to support scenario-specific retrieval of medical free text*. In Information Retrieval, vol. 10, no. 2, pp. 173–202. ISSN 1386-4564. doi:10.1007/s10791-006-9020-6. URL http://dx.doi.org/10.1007/s10791-006-9020-6. Cited on pp. 45 and 53.

Liu, Jingjing and Belkin, Nicholas J. 2010. *Personalizing information retrieval for people with different levels of topic knowledge*. In Proceedings of the 10th annual joint conference on Digital libraries, JCDL '10, pp. 383–384. ACM, New York, NY, USA. ISBN 978-1-4503-0085-8. doi:10.1145/1816123.1816191. URL http://dx.doi.org/10.1145/1816123.1816191. Cited on pp. 151 and 152.

Liu, Ling and Özsu, M. Tamer, eds. Sep. 2009. *Encyclopedia of Database Systems*. Springer-Verlag. ISBN 0387355448. URL http://tomgruber.org/writing/ontology-definition-2007.htm. Cited on p. 27.

Lorence, D. and Park, H. Feb. 2007. *Study of education disparities and health information seeking behavior.* In Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society, vol. 10, no. 1, pp. 149–151. ISSN 1094-9313. doi:10.1089/cpb.2006.9977. URL http://dx.doi.org/10.1089/cpb.2006.9977. Cited on p. 30.

Lu, Wen-Hsiang; Lin, Shih-Jui; Chan, Yi-Che; and Chen, Kuan-Hsi. 2005. *Semi-automatic construction of the Chinese-English MeSH using Web-based term translation method.* In AMIA Annual Symposium proceedings, pp. 475–479. ISSN 1942-597X. URL http://view.ncbi.nlm.nih.gov/pubmed/16779085. Cited on p. 133.

Lu, Wen-Hsiang H.; Shih-Jui, Ray; Chan, Yi-Che C.; and Chen, Kuan-Hsi H. 2006. *Overcoming terminology barrier using Web resources for cross-language medical information retrieval.* In AMIA Annual Symposium proceedings, pp. 519–523. ISSN 1942-597X. URL http://view.ncbi.nlm.nih.gov/pubmed/17238395. Cited on p. 148.

Luhn, H. P. Oct. 1957. *A statistical approach to mechanized encoding and searching of literary information*. In IBM Journal of Research and Development, vol. 1, no. 4, pp. 309–317. URL http://www.research.ibm.com/journal/rd/014/ibmrd0104D.pdf. Cited on p. 4.

Luo, Gang. 2009. *Design and Evaluation of the iMed Intelligent Medical Search Engine*. In Proceedings of the 2009 IEEE International Conference on Data Engineering, ICDE '09, pp. 1379–1390. IEEE Computer Society, Washington, DC, USA. ISBN 978-0-7695-3545-6. doi:10.1109/icde.2009.10. URL http://dx.doi.org/10.1109/icde.2009.10. Cited on pp. 148, 149, and 200.

Luo, Gang and Tang, Chunqiang. 2008. *On iterative intelligent medical search*. In SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 3–10. ACM, New York, NY, USA. ISBN 9781605581644. doi:10.1145/1390334.1390338. URL http://dx.doi.org/10.1145/1390334.1390338. Cited on pp. 51, 53, and 200.

Luo, Gang; Tang, Chunqiang; Yang, Hao; and Wei, Xing. 2008. *MedSearch: a specialized search engine for medical information retrieval.* In CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge mining, pp. 143–152. ACM, New York, NY, USA. ISBN 978-1-59593-991-3. doi:10.1145/1458082.1458104. URL http://dx.doi.org/10.1145/1458082.1458104. Cited on pp. 148 and 149.

Lyman, Peter and Varian, Hal R. 2003. *How much information*. Available from: http://www.sims.berkeley.edu/how-much-info-2003 [cited 2008-07-10]. Cited on p. 4.

Ma, Zhongrui; Chen, Yu; Song, Ruihua; Sakai, Tetsuya; Lu, Jiaheng; and Wen, Ji R. 2012. *New assessment criteria for query suggestion*. In Proceedings of the

35th international ACM SIGIR conference on Research and development in information retrieval, SIGIR '12, pp. 1109–1110. ACM, New York, NY, USA. ISBN 978-1-4503-1472-5. doi:10.1145/2348283.2348493. URL http://dx.doi.org/10.1145/2348283.2348493. Cited on p. 199.

Macevičiūtė, Elena and Wilson, T. D. Dec. 2008. *Proceedings of the 7th conference on Information Seeking in Context*. In Information Research, vol. 13, no. 4. URL http://informationr.net/ir/13-4/isic08.html. Cited on p. 42.

Mamlin, J. J. and Baker, D. H. 1973. *Combined time-motion and work sampling study in a general medicine clinic.* In Medical care, vol. 11, no. 5, pp. 449–456. ISSN 0025-7079. URL http://view.ncbi.nlm.nih.gov/pubmed/4744980. Cited on p. 17.

Manning, Christopher D.; Raghavan, Prabhakar; and Schütze, Hinrich. Jul. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 1st edn. ISBN 0521865719. URL http://www.worldcat.org/isbn/0521865719. Cited on pp. 3, 4, 43, 44, 65, 66, 199, and 200.

Mansourian, Yazdan. 2008. *Contextualization of web searching: a grounded theory approach.* In The Electronic Library, vol. 26, no. 2, pp. 202–214. URL http://www.emeraldinsight.com/10.1108/02640470810864091. Cited on pp. x and 41.

Marchionini, Gary. Mar. 1997. *Information Seeking in Electronic Environments (Cambridge Series on Human-Computer Interaction)*. Cambridge University Press. ISBN 0521586747. URL http://www.worldcat.org/isbn/0521586747. Cited on p. 38.

Marcus, R. S. Nov. 1983. *An experimental comparison of the effectiveness of computers and humans as search intermediaries.* In Journal of the American Society for Information Science. American Society for Information Science, vol. 34, no. 6, pp. 381–404. ISSN 0002-8231. URL http://view.ncbi.nlm.nih.gov/pubmed/10299379. Cited on p. 53.

Martinez, R.; Cebrian, M.; de Borja Rodriguez, F.; and Camacho, D. 2008. *Contextual information retrieval based on algorithmic information theory and statistical outlier detection.* In IEEE Information Theory Workshop (ITW '08)., pp. 292–297. doi:10.1109/ITW.2008.4578672. URL http://dx.doi.org/10.1109/ITW.2008.4578672. Cited on p. 45.

Martins, Diogo S.; Santana, Luiz H. Z.; Biajiz, Mauro; Antonio; and de Souza, Wanderley L. 2008. *Context-aware information retrieval on a ubiquitous medical learning environment*. In SAC '08: Proceedings of the 2008 ACM symposium on Applied computing, pp. 2348–2349. ACM, New York, NY, USA. ISBN 9781595937537. doi:10.1145/1363686.1364243. URL http://dx.doi.org/10.1145/1363686.1364243. Cited on pp. 45, 47, 49, 50, and 54.

Maviglia, Saverio M.; Yoon, Catherine S.; Bates, David W.; and Kuperman, Gilad. Jan. 2006. *KnowledgeLink: Impact of Context-Sensitive Information Retrieval on Clinicians' Information Needs*. In J Am Med Inform Assoc, vol. 13, no. 1, pp. 67–73. Cited on p. 52.

Mccray, A. T. 2003. *An upper-level ontology for the biomedical domain*. In Comparative and Functional Genomics, pp. 80–84. ISSN 1532-6268. doi:10.1002/cfg.255. URL http://dx.doi.org/10.1002/cfg.255. Cited on p. 28.

McCray, A. T.; Loane, R. F.; Browne, A. C.; and Bangalore, A. K. 1999. *Terminology issues in user access to Web-based medical information.* In Proceedings / AMIA ... Annual Symposium. AMIA Symposium, pp. 107–111. ISSN 1531-605X. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2232498/. Cited on p. 247.

McCray, Alexa T. Mar. 2005. *Promoting Health Literacy*. In Journal of the American Medical Informatics Association, vol. 12, no. 2, pp. 152–163. ISSN 1067-5027. doi:10.1197/jamia.m1687. URL http://dx.doi.org/10.1197/jamia.m1687. Cited on p. 150.

McCray, Alexa T. and Tse, Tony. 2003. *Understanding search failures in consumer health information systems.* In AMIA Annual Symposium Proceedings, pp. 430–434. ISSN 1942-597X. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1479930/. Cited on pp. 147 and 197.

Mcdonald, C. J.; Huff, S. M.; Suico, J. G.; Hill, G.; Leavelle, D.; Aller, R.; Forrey, A.; Mercer, K.; Demoor, G.; Hook, J.; Williams, W.; Case, J.; and Maloney, P. Apr. 2003. *LOINC, a universal standard for identifying laboratory observations: a 5-year update.* In Clin Chem, vol. 49, no. 4, pp. 624–633. ISSN 0009-9147. doi:10.1373/49.4.624. URL http://dx.doi.org/10.1373/49.4.624. Cited on p. 23.

Mendonça, E. A.; Cimino, J. J.; Johnson, S. B.; and Seol, Y. H. Apr. 2001. *Accessing heterogeneous sources of evidence to answer clinical questions.* In Journal of biomedical informatics, vol. 34, no. 2, pp. 85–98. ISSN 1532-0464. doi:10.1006/jbin.2001.1012. URL http://dx.doi.org/10.1006/jbin.2001.1012. Cited on p. 52.

Miller, R. A.; Gieszczykiewicz, F. M.; Vries, J. K.; and Cooper, G. F. 1992. *CHARTLINE: providing bibliographic references relevant to patient charts using the UMLS Metathesaurus Knowledge Sources*. In Proceedings of the Annual Symposium on Computer Application in Medical Care., pp. 86–90. ISSN 0195-4210. URL http://view.ncbi.nlm.nih.gov/pubmed/1483014. Cited on p. 51.

Miller, Trudi; Leroy, Gondy; Chatterjee, Samir; Fan, Jie; and Thoms, Brian. 2007. *A Classifier to Evaluate Language Specificity of Medical Documents*. In HICSS '07: Proceedings of the 40th Annual Hawaii International Conference on System Sciences. IEEE Computer Society, Washington, DC, USA. doi:10.1109/hicss.2007.6. URL http://dx.doi.org/10.1109/hicss.2007.6. Cited on p. 32.

Mooers, Calvin E. 1950. *Coding, Information Retrieval, and the Rapid Selector*. In American Documentation, vol. 1, no. 4, pp. 225–229. Cited on p. 3.

Muresan, G.; Cole, M.; Smith, C. L.; Liu, Lu; and Belkin, N. J. 2006. *Does Familiarity Breed Content? Taking Account of Familiarity with a Topic in Personalizing Information Retrieval.* In Proceedings of the 39th Annual Hawaii

International Conference on System Sciences, 2006. HICSS '06., p. 53c. doi: 10.1109/hicss.2006.130. URL http://dx.doi.org/10.1109/hicss.2006.130. Cited on pp. 152 and 179.

Musen, M. A. 2002. *Medical informatics: searching for underlying components.* In Methods of information in medicine, vol. 41, no. 1, pp. 12–19. ISSN 0026-1270. URL http://view.ncbi.nlm.nih.gov/pubmed/11933757. Cited on p. 28.

Mylonas, P.; Vallet, D.; Castells, P.; Fernandez, M.; and Avrithis, Y. Mar. 2008. *Personalized information retrieval based on context and ontological knowledge.* In The Knowledge Engineering Review, vol. 23, no. Special Issue 01, pp. 73–100. ISSN 0269-8889. doi:10.1017/s0269888907001282. URL http://dx.doi.org/10.1017/s0269888907001282. Cited on p. 45.

Névéol, Aurélie; Pereira, Suzanne; Soualmia, Lina F.; Thirion, Benoît; and Darmoni, Stéfan J. 2006. *A method of cross-lingual consumer health information retrieval.* In Studies in health technology and informatics, vol. 124, pp. 601–608. ISSN 0926-9630. URL http://view.ncbi.nlm.nih.gov/pubmed/17108583. Cited on p. 133.

NFIL. 2013. *What is the NFIL?* Available from: http://infolit.org/about-the-nfil/what-is-the-nfil/ [cited 2013-03-04]. URL http://infolit.org/about-the-nfil/what-is-the-nfil/. Cited on p. 184.

NLM. Nov. 2001. *Relationships in Medical Subject Headings (MeSH).* Available from: http://www.nlm.nih.gov/mesh/meshrels.html [cited 2013-01-07]. Cited on p. 22.

NLM. Sep. 2009. *UMLS Reference Manual.* Bethesda (MD): National Library of Medicine (US). URL http://www.ncbi.nlm.nih.gov/books/NBK9676/. Cited on pp. 25, 26, and 27.

NLM. 2011. *MEDLINE Fact Sheet.* Available from: http://www.nlm.nih.gov/pubs/factsheets/medline.html [cited 2013-01-07]. Cited on p. 18.

NLM. 2012a. *2012AA Consumer Health Vocabulary Source Information.* URL http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CHV/index.html. Cited on p. 23.

NLM. Nov. 2012b. *The Medical Subject Headings Vocabulary.* Available from: http://www.nlm.nih.gov/pubs/factsheets/mesh.html [cited 2013-01-07]. Cited on pp. 22 and 23.

NLM. Aug. 2012c. *MeSH Tree Structures.* Available from: http://www.nlm.nih.gov/mesh/trees.html [cited 2013-01-07]. Cited on p. 23.

NLM. Oct. 2012d. *Principles of MEDLINE Subject Indexing.* Available from: http://www.nlm.nih.gov/bsd/disted/mesh/indexprinc.html [cited 2013-01-07]. Cited on p. 23.

NLM. May 2012e. *UMLS Metathesaurus Fact Sheet*. Available from: http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html [cited 2013-01-07]. Cited on p. 25.

NLM. Aug. 2012f. *Use of MeSH in Online Retrieval*. Available from: http://www.nlm.nih.gov/mesh/intro_retrieval.html [cited 2013-01-07]. Cited on p. 23.

Pandey, Gaurav and Luxenburger, Julia. 2008. *Exploiting Session Context for Information Retrieval - A Comparative Study*. In Macdonald, Craig; Ounis, Iadh; Plachouras, Vassilis; Ruthven, Ian; and White, Ryen W., eds., Advances in Information Retrieval, vol. 4956 of *Lecture Notes in Computer Science*, chap. 73, pp. 652–657. Springer Berlin, Berlin, Heidelberg. ISBN 978-3-540-78645-0. doi:10.1007/978-3-540-78646-7_73. URL http://dx.doi.org/10.1007/978-3-540-78646-7_73. Cited on p. 45.

Parker, Ruth; Baker, David; Williams, Mark; and Nurss, Joanne. Oct. 1995. *The test of functional health literacy in adults*. In Journal of General Internal Medicine, vol. 10, no. 10, pp. 537–541. ISSN 0884-8734. doi:10.1007/bf02640361. URL http://dx.doi.org/10.1007/bf02640361. Cited on p. 125.

Patrick, T. B.; Monga, H. K.; Sievert, M. E.; Houston Hall, J.; and Longo, D. R. 2001. *Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes.* In Journal of medical Internet research, vol. 3, no. 3. ISSN 1438-8871. doi:10.2196/jmir.3.3.e24. URL http://dx.doi.org/10.2196/jmir.3.3.e24. Cited on pp. 148 and 157.

Pedersen, Ted; Pakhomov, Serguei V. S.; Patwardhan, Siddharth; and Chute, Christopher G. Jun. 2007. *Measures of semantic similarity and relatedness in the biomedical domain*. In J. of Biomedical Informatics, vol. 40, no. 3, pp. 288–299. ISSN 1532-0464. doi:10.1016/j.jbi.2006.06.004. URL http://dx.doi.org/10.1016/j.jbi.2006.06.004. Cited on p. 272.

Pellegrin, P. 1986. *Aristotle's Classification of Animals: Biology and the Conceptual Unity of the Aristotelian Corpus*. Berkeley: University of California Press. Cited on p. 21.

Petrock, Victoria. Aug. 2010. *Cyberchondriacs Becoming Empowered Health Information Seekers*. Available from: http://www.emarketer.com/blog/index.php/cyberchondriacs-empowered-health-seekers/ [cited 2011-05-11] (Archived by WebCite at http://www.webcitation.org/5ym7xLoYp). URL http://www.emarketer.com/blog/index.php/cyberchondriacs-empowered-health-seekers/. Cited on p. 147.

Pirkola, Ari. 1998. *The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval.* In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '98, pp. 55–63. ACM, New York, NY, USA. ISBN 1-58113-015-5. doi:10.1145/290941.290957. URL http://dx.doi.org/10.1145/290941.290957. Cited on p. 132.

Plovnick, Robert M. and Zeng, Qing T. Sep. 2004. *Reformulation of consumer health queries with professional terminology: a pilot study.* In Journal of medical Internet research, vol. 6, no. 3. ISSN 1438-8871. doi:10.2196/jmir.6.3.e27. URL http://dx.doi.org/10.2196/jmir.6.3.e27. Cited on pp. 148 and 157.

Powsner, Seth M. and Miller, Perry L. Nov. 1989. *Linking Bibliographic Retrieval to Clinical Reports: PsychTopix*. In Proc Annu Symp Comput Appl Med Care. Cited on p. 51.

Pratt, Wanda; Hearst, Marti A.; and Fagan, Lawrence M. 1999. *A knowledge-based approach to organizing retrieved documents*. In AAAI '99/IAAI '99: Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence, pp. 80–85. American Association for Artificial Intelligence, Menlo Park, CA, USA. ISBN 0-262-51106-1. URL http://portal.acm.org/citation.cfm?id=315232. Cited on p. 54.

Price, S. L. and Delcambre, L. M. 2005. *Using concept relations to improve ranking in information retrieval.* In AMIA Annu Symp Proc. 2005, pp. 619–623. Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, USA. ISSN 1559-4076. URL http://view.ncbi.nlm.nih.gov/pubmed/16779114. Cited on p. 33.

Price, S. L. and Hersh, W. R. 1999. *Filtering Web pages for quality indicators: an empirical approach to finding high quality consumer health information on the World Wide Web.* In AMIA Annu Symp Proc. 1999, pp. 911–915. Division of Medical Informatics and Outcomes Research, Oregon Health Sciences University, Portland, USA. ISSN 1531-605X. URL http://view.ncbi.nlm.nih.gov/pubmed/10566493. Cited on p. 33.

Price, Susan L.; Delcambre, Lois M.; and Nielsen, Marianne L. 2006. *Using semantic components to express clinical questions against document collections*. In HIKM '06: Proceedings of the international workshop on Healthcare information and knowledge management, pp. 9–16. ACM Press, New York, NY, USA. ISBN 1595935282. doi:10.1145/1183568.1183571. URL http://dx.doi.org/10.1145/1183568.1183571. Cited on pp. 33 and 34.

Price, Susan L.; Hersh, William R.; Olson, Daniel D.; and Embi, Peter J. 2002. *SmartQuery: context-sensitive links to medical knowledge sources from the electronic patient record.* In Proc AMIA Symp, pp. 627–631. ISSN 1531-605X. URL http://view.ncbi.nlm.nih.gov/pubmed/12463899. Cited on p. 52.

Price, Susan L.; Nielsen, Marianne L.; Delcambre, Lois M. L.; and Vedsted, Peter. 2007. *Semantic components enhance retrieval of domain-specific documents*. In CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 429–438. ACM, New York, NY, USA. ISBN 9781595938039. doi:10.1145/1321440.1321502. URL http://dx.doi.org/10.1145/1321440.1321502. Cited on pp. 33 and 34.

Purcell, Gretchen P.; Rennels, Glenn D.; and Shortliffe, Edward H. Dec. 1997. *Development and evaluation of a context-based document representation for searching the medical literature*. In International Journal on Digital Libraries,

vol. 1, no. 3, pp. 288–296. doi:10.1007/s007990050023. URL http://dx.doi.org/10.1007/s007990050023. Cited on pp. 49 and 50.

Qu, Peng; Liu, Chang; and Lai, Maosheng. 2010. *The effect of task type and topic familiarity on information search behaviors*. In Proceeding of the third symposium on Information interaction in context, IIiX '10, pp. 371–376. ACM, New York, NY, USA. ISBN 978-1-4503-0247-0. doi:10.1145/1840784.1840841. URL http://dx.doi.org/10.1145/1840784.1840841. Cited on p. 151.

Quellec, G.; Lamard, M.; Bekri, L.; Cazuguel, G.; Cochener, B.; and Roux, C. 2007. *Multimedia medical case retrieval using decision trees*. In IEEE Engineering in Medicine and Biology Society, vol. 2007, pp. 4536–4539. ENST Bretagne, GET-ENST, Brest, F-29200 France. gwenole.quellec@enst-bretagne.fr. ISSN 1557-170X. doi:10.1109/iembs.2007.4353348. URL http://dx.doi.org/10.1109/iembs.2007.4353348. Cited on pp. 49 and 50.

Rahurkar, Mandar and Cucerzan, Silviu. 2008. *Predicting when browsing context is relevant to search*. In Myaeng, Sung H.; Oard, Douglas W.; Sebastiani, Fabrizio; Chua, Tat S.; Leong, Mun K.; Myaeng, Sung H.; Oard, Douglas W.; Sebastiani, Fabrizio; Chua, Tat S.; and Leong, Mun K., eds., SIGIR, pp. 841–842. ACM, New York, NY, USA. ISBN 978-1-60558-164-4. doi:10.1145/1390334.1390532. URL http://dx.doi.org/10.1145/1390334.1390532. Cited on p. 45.

Rains, Stephen A. 2007. *Perceptions of Traditional Information Sources and Use of the World Wide Web to Seek Health Information: Findings From the Health Information National Trends Survey*. In Journal of Health Communication, vol. 12, no. 7, pp. 667–680. doi:10.1080/10810730701619992. URL http://dx.doi.org/10.1080/10810730701619992. Cited on p. 29.

Rawson, Katherine A.; Gunstad, John; Hughes, Joel; Spitznagel, Mary Beth B.; Potter, Vanessa; Waechter, Donna; and Rosneck, James. Jan. 2010. *The METER: a brief, self-administered measure of health literacy*. In Journal of general internal medicine, vol. 25, no. 1, pp. 67–71. ISSN 1525-1497. doi:10.1007/s11606-009-1158-7. URL http://dx.doi.org/10.1007/s11606-009-1158-7. Cited on p. 207.

Rector, A. L. and Nowlan, W. A. Oct. 1994. *The GALEN project*. In Computer methods and programs in biomedicine, vol. 45, no. 1-2, pp. 75–78. ISSN 0169-2607. URL http://view.ncbi.nlm.nih.gov/pubmed/7889770. Cited on p. 28.

Reddy, Madhu C. and Spence, Patricia R. Jan. 2008. *Collaborative information seeking: A field study of a multidisciplinary patient care team*. In Information Processing & Management, vol. 44, no. 1, pp. 242–255. ISSN 03064573. doi:10.1016/j.ipm.2006.12.003. URL http://dx.doi.org/10.1016/j.ipm.2006.12.003. Cited on p. 30.

Revere, Debra; Turner, Anne M.; Madhavan, Ann; Rambo, Neil; Bugni, Paul F.; Kimball, AnnMarie; and Fuller, Sherrilynne S. Aug. 2007. *Understanding the information needs of public health practitioners: a literature review to inform design of an interactive digital knowledge management system*.

In Journal of biomedical informatics, vol. 40, no. 4, pp. 410–421. ISSN 1532-0480. doi:10.1016/j.jbi.2006.12.008. URL http://dx.doi.org/10.1016/j.jbi.2006.12.008. Cited on p. 30.

Rijsbergen, C. 1979. *Information Retrieval*. Butterworths, London. URL http://www.dcs.gla.ac.uk/Keith/Preface.html. Cited on p. 3.

Rijsbergen, K.; Jose, J.; Urban, J.; Villa, R.; and Joho, H., eds. 2006. *Proceedings of the First International Workshop on Adaptive Information Retrieval (AIR) at String Processing and Information Retrieval (SPIRE)*. Glasgow, UK. Cited on p. 42.

Ritchie, Anna; Robertson, Stephen; and Teufel, Simone. 2008. *Comparing citation contexts for information retrieval*. In Proceedings of the 17th ACM conference on Information and knowledge management, CIKM '08, pp. 213–222. ACM, New York, NY, USA. ISBN 978-1-59593-991-3. doi:10.1145/1458082.1458113. URL http://dx.doi.org/10.1145/1458082.1458113. Cited on p. 45.

Robertson, S. E. and Jones, Sparck K. 1976. *Relevance weighting of search terms*. In Journal of the American Society for Information Science, vol. 27, no. 3, pp. 129–146. Cited on p. 4.

Robertson, Stephen E.; Kanoulas, Evangelos; and Yilmaz, Emine. Jul. 2010. *Extending average precision to graded relevance judgments*. In Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '10, pp. 603–610. ACM, New York, NY, USA. ISBN 978-1-4503-0153-4. doi:10.1145/1835449.1835550. URL http://dx.doi.org/10.1145/1835449.1835550. Cited on pp. 67 and 74.

Rosemblat, G.; Logan, R.; Tse, T.; and Graham, L. 2006. *Test Features and Readability: Expert Evaluation of Consumer Health Text*. In MEDNET. Cited on p. 32.

Rosemblat, Graciela; Gemoets, Darren; Browne, Allen C.; and Tse, Tony. 2003. *Machine translation-supported cross-language information retrieval for a consumer health resource*. In AMIA Annual Symposium proceedings, pp. 564–568. ISSN 1942-597X. URL http://view.ncbi.nlm.nih.gov/pubmed/14728236. Cited on p. 132.

Sakji, S.; Dibad, A. D. D.; Kergourlay, I.; Darmoni, S.; and Joubert, M. 2009. *Information retrieval in context using various health terminologies*. In Research Challenges in Information Science, 2009. RCIS 2009. Third International Conference on, pp. 453–458. doi:10.1109/rcis.2009.5089310. URL http://dx.doi.org/10.1109/rcis.2009.5089310. Cited on pp. 49 and 50.

Salton, G. 1968. *Automatic Information Organization and Retrieval*. McGraw-Hill, New York, NY, USA. Cited on p. 3.

Salton, G. 1971. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. URL http://portal.acm.org/citation.cfm?id=1102022. Cited on p. 4.

Sanderson, Mark. 2010. *Test Collection Based Evaluation of Information Retrieval Systems*. In Foundations and Trends in Information Retrieval, vol. 4, no. 4, pp. 247–375. Cited on pp. 13, 65, and 66.

Santana, Silvina and Pereira, Sousa A. 2007. *Da Utilização da Internet para Questões de Saúde e Doença em Portugal*. In Acta Med Port, vol. 20, pp. 47–57. URL http://www.actamedicaportuguesa.com/pdf/2007-20/1/47-58.pdf. Cited on p. 29.

Saracevic, Tefko. 1975. *Relevance: A Review of and a framework for the thinking on the notion in information science*. In Journal of the American Society for Information Science, vol. 26, no. 6, pp. 321–343. Cited on p. 99.

Saracevic, Tefko. Oct. 1996. *Relevance reconsidered*. In Information science: Integration in perspectives.Proceedings of the Second Conference on Conceptions of Library and Information Science, pp. 201–218. Royal School of Librarianship, Copenhagen. Cited on pp. 13, 99, 100, 123, 127, 132, 143, 155, 208, and 210.

Saracevic, Tefko. Nov. 2007a. *Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance*. In J. Am. Soc. Inf. Sci. Technol., vol. 58, no. 13, pp. 2126–2144. ISSN 1532-2882. doi:10.1002/asi.v58:13. URL http://dx.doi.org/10.1002/asi.v58:13. Cited on p. 99.

Saracevic, Tefko. 2007b. *Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance*. In J. Am. Soc. Inf. Sci. Technol., vol. 58, no. 13, pp. 2126–2144. ISSN 1532-2882. Cited on pp. 99, 108, 115, 166, and 232.

Sato, K. Jun. 2003. *Context Sensitive Interactive Systems Design: A Framework for Representation of contexts*. In Proceedings of the 10th International Conference on Human-Computer Interaction, vol. 3, pp. 1323–1327. Lawrence-Erlbaum Associates, Crete, Greece. URL http://www.google.com/search?client=safari&rls=en-us&q=Context+sensitive+interactive+systems+design&ie=UTF-8&oe=UTF-8. Cited on p. 38.

Schamber, L. Nov. 1990. *A re-examination of relevance: toward a dynamic, situational definition*. In Inf. Process. Manage., vol. 26, no. 6, pp. 755–776. ISSN 0306-4573. doi:10.1016/0306-4573(90)90050-c. URL http://dx.doi.org/10.1016/0306-4573(90)90050-c. Cited on p. 99.

Schamber, Linda. 1994. *Relevance and Information Behavior*. In Annual Review of Information Science and Technology (ARIST), pp. 3–48. Cited on p. 99.

Schembri, G. and Schober, P. Apr. 2009. *The Internet as a diagnostic aid: the patients' perspective*. In Int J STD AIDS, vol. 20, no. 4, pp. 231–233. doi:10.1258/ijsa.2008.008339. URL http://dx.doi.org/10.1258/ijsa.2008.008339. Cited on p. 87.

Scott-Wright, A.; Crowell, J.; Zeng, Q.; Bates, D.; and Greenes, R. 2006. *Analysis of information needs of users of MEDLINEplus, 2002 - 2003*. In AMIA

Annu Symp Proc. 2006, pp. 699–703. Decision Systems Group, Brigham & Women's Hospital, Harvard Medical School, Boston, MA, USA. ISSN 1559-4076. URL http://view.ncbi.nlm.nih.gov/pubmed/17238431. Cited on p. 29.

Shang, Yi and Li, Longzhuang. 2002. *Precision Evaluation of Search Engines*. In World Wide Web, vol. 5, no. 2, pp. 159–173. ISSN 1386-145X. doi:10.1023/a:1019679624079. URL http://dx.doi.org/10.1023/a:1019679624079. Cited on pp. 67 and 68.

Sharit, Joseph; Hernández, Mario A.; Czaja, Sara J.; and Pirolli, Peter. May 2008. *Investigating the Roles of Knowledge and Cognitive Abilities in Older Adult Information Seeking on the Web*. In ACM Trans. Comput.-Hum. Interact., vol. 15, no. 1, pp. 1–25. ISSN 1073-0516. doi:10.1145/1352782.1352785. URL http://dx.doi.org/10.1145/1352782.1352785. Cited on p. 29.

Sherrilynne, Hsinchun;. Jun. 2005. *Medical Informatics: Knowledge Management and Data Mining in Biomedicine (Integrated Series in Information Systems)*. Springer, 2005th edn. ISBN 038724381X. URL http://www.worldcat.org/isbn/038724381X. Cited on pp. 25 and 28.

Shiri, Ali. 2005. *Topic familiarity and its effects on term selection and browsing in a thesaurus-enhanced search environment*. In Library Review, vol. 54, no. 9, pp. 514–518. Cited on pp. 151 and 184.

Shortliffe, Edward H. and Cimino, James J., eds. 2000. *Medical Informatics: Computer Applications in Health Care and Biomedicine*. Springer-Verlag, 2nd edn. Cited on pp. 18, 19, 21, and 22.

Shtykh, Roman Y. and Jin, Qun. 2008. *Capturing User Contexts: Dynamic Profiling for Information Seeking Tasks*. In 3rd International Conference on Systems and Networks Communications (ICSNC '08)., pp. 365–370. doi:10.1109/ICSNC.2008.55. URL http://dx.doi.org/10.1109/ICSNC.2008.55. Cited on p. 45.

Sihvonen, Anne and Vakkari, Pertti. 2004. *Subject knowledge, thesaurus-assisted query expansion and search success*. In Proceedings of the RIAO 2004 Conference. Cited on pp. 151 and 184.

Silva, Juan M. and Favela, Jesus. 2006. *Context Aware Retrieval of Health Information on the Web*. In LA-WEB '06: Proceedings of the Fourth Latin American Web Congress (LA-WEB'06), pp. 135–146. IEEE Computer Society, Washington, DC, USA. ISBN 0769526934. doi:10.1109/la-web.2006.10. URL http://dx.doi.org/10.1109/la-web.2006.10. Cited on pp. 6, 53, and 54.

Skov, M.; Larsen, B.; and Ingwersen, P. Sep. 2008. *Inter and intra-document contexts applied in polyrepresentation for best match IR*. In Information Processing & Management, vol. 44, no. 5, pp. 1673–1683. ISSN 03064573. doi:10.1016/j.ipm.2008.05.006. URL http://dx.doi.org/10.1016/j.ipm.2008.05.006. Cited on p. 45.

Smith, Barry and Rosse, Cornelius. 2004. *The role of foundational relations in the alignment of biomedical ontologies*. In Studies in health technology

and informatics, vol. 107, no. Pt 1, pp. 444–448. ISSN 0926-9630. URL http://view.ncbi.nlm.nih.gov/pubmed/15360852. Cited on p. 22.

Souden, Maria and Rubenstein, Ellen L. 2010. *Listening to patients: how understanding health information use can contribute to health literacy constructs*. In Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47, ASIS&T '10. American Society for Information Science, Silver Springs, MD, USA. URL http://portal.acm.org/citation.cfm?id=1920331.1920434. Cited on p. 183.

Spink, A.; Yang, Y.; Jansen, J.; Nykanen, P.; Lorence, D. P.; Ozmutlu, S.; and Ozmutlu, H. C. Mar. 2004. *A study of medical and health queries to web search engines*. In Health Information & Libraries Journal, vol. 21, no. 1, pp. 44–51. ISSN 1471-1834. doi:10.1111/j.1471-1842.2004.00481.x. URL http://dx.doi.org/10.1111/j.1471-1842.2004.00481.x. Cited on pp. 30 and 247.

Srinivasan, P. 1996. *Retrieval feedback in MEDLINE*. In Journal of the American Medical Informatics Association : JAMIA, vol. 3, no. 2, pp. 157–167. ISSN 1067-5027. URL http://view.ncbi.nlm.nih.gov/pubmed/8653452. Cited on p. 52.

Stapley, B. J. and Benoit, G. 2000. *Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts*. In Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, pp. 529–540. ISSN 1793-5091. URL http://view.ncbi.nlm.nih.gov/pubmed/10902200. Cited on p. 35.

Stichele, Robert V. Dec. 1995. *Multilingual Glossary of technical and popular medical terms in nine European Languages*. Tech. rep., Heymans Institute of Pharmacology, University of Gent, Gent. URL http://users.ugent.be/~{}rvdstich/eugloss/welcome.html. Cited on pp. 24 and 120.

Stuckenschmidt, H.; van Harmelen, F.; de Waard, A.; Scerri, T.; Bhogal, R.; van Buel, J.; Crowlesmith, I.; Fluit, C.; Kampman, A.; Broekstra, J.; and van Mulligent, E. May 2004. *Exploring Large Document Repositories with RDF Technology: The DOPE Project*. In IEEE Intelligent Systems, vol. 19, no. 3, pp. 34–40. ISSN 1541-1672. doi:10.1109/mis.2004.9. URL http://dx.doi.org/10.1109/mis.2004.9. Cited on p. 34.

Su, Louise T. Nov. 2003a. *A comprehensive and systematic model of user evaluation of web search engines: I. theory and background*. In Journal of the American Society for Information Science and Technology, vol. 54, no. 13, pp. 1175–1192. ISSN 1532-2882. doi:10.1002/asi.10303. URL http://dx.doi.org/10.1002/asi.10303. Cited on p. 68.

Su, Louise T. 2003b. *A comprehensive and systematic model of user evaluation of web search engines: II. an evaluation by undergraduates*. In Journal of the American Society for Information Science and Technology, vol. 54, no. 13, pp. 1193–1223. ISSN 1532-2882. doi:10.1002/asi.10334. URL http://dx.doi.org/10.1002/asi.10334. Cited on p. 68.

Summers, Kathryn and Summers, Michael. Jan. 2005. *Reading and navigational strategies of Web users with lower literacy skills*. In Proc. Am. Soc.

Info. Sci. Tech., vol. 42, no. 1, p. n/a. doi:10.1002/meet.1450420179. URL http://dx.doi.org/10.1002/meet.1450420179. Cited on pp. 150, 184, 185, 191, and 275.

Tang, H. and Ng, J. H. Dec. 2006. *Googling for a diagnosis–use of Google as a diagnostic aid: internet based study.* In BMJ (Clinical research ed.), vol. 333, no. 7579, pp. 1143–1145. ISSN 1756-1833. doi:10.1136/bmj.39003.640567.ae. URL http://dx.doi.org/10.1136/bmj.39003.640567.ae. Cited on p. 30.

Tang, M. C. and Sun, Y. 2003. *Evaluation of Web-based search engines using user-effort measures*. In : Library and Information Science Research Electronic Journal, vol. 13, no. 2. URL http://libres.curtin.edu.au/libres13n2/index.htm. Cited on p. 68.

Tang, Thanh T.; Craswell, Nick; Hawking, David; Griffiths, Kathy; and Christensen, Helen. Mar. 2006. *Quality and relevance of domain-specific search: A case study in mental health*. In Information Retrieval, vol. 9, no. 2, pp. 207–225. ISSN 1386-4564. doi:10.1007/s10791-006-7150-5. URL http://dx.doi.org/10.1007/s10791-006-7150-5. Cited on pp. 34, 69, 70, and 73.

Toms, Elaine G. and Latter, Celeste. Sep. 2007. *How consumers search for health information.* In Health informatics journal, vol. 13, no. 3, pp. 223–235. ISSN 1460-4582. doi:10.1177/1460458207079901. URL http://dx.doi.org/10.1177/1460458207079901. Cited on pp. 147, 167, and 197.

Torres, Sergio D.; Hiemstra, Djoerd; Weber, Ingmar; and Serdyukov, Pavel. 2012. *Query recommendation for children*. In Proceedings of the 21st ACM international conference on Information and knowledge management, CIKM '12, pp. 2010–2014. ACM, New York, NY, USA. ISBN 978-1-4503-1156-4. doi:10.1145/2396761.2398562. URL http://dx.doi.org/10.1145/2396761.2398562. Cited on p. 199.

Tran, Tuan D.; Garcelon, Nicolas; Burgun, Anita; and Le Beux, Pierre. 2004. *Experiments in cross-language medical information retrieval using a mixing translation module*. In Studies in Health Technology and Informatics, vol. 107, no. Pt 2, pp. 946–949. ISSN 0926-9630. URL http://view.ncbi.nlm.nih.gov/pubmed/15360952. Cited on p. 133.

Turtle, Howard and Croft, W. Bruce. Jul. 1991. *Evaluation of an inference network-based retrieval model.* In ACM Trans. Inf. Syst., vol. 9, no. 3, pp. 187–222. ISSN 1046-8188. doi:10.1145/125187.125188. URL http://dx.doi.org/10.1145/125187.125188. Cited on p. 4.

Twose, Claire; Swartz, Patricia; Bunker, Edward; Roderer, Nancy K.; and Oliver, Kathleen B. Mar. 2008. *Public health practitioners information access and use patterns in the Maryland (USA) public health departments of Anne Arundel and Wicomico Counties*. In Health information and libraries journal, vol. 25, no. 1, pp. 13–22. ISSN 1471-1834. doi:10.1111/j.1471-1842.2007.00738.x. URL http://dx.doi.org/10.1111/j.1471-1842.2007.00738.x. Cited on p. 30.

Ukkonen, Antti; Castillo, Carlos; Donato, Debora; and Gionis, Aristides. 2008. *Searching the wikipedia with contextual information*. In CIKM

'08: Proceeding of the 17th ACM conference on Information and knowledge management, pp. 1351–1352. ACM, New York, NY, USA. ISBN 978-1-59593-991-3. doi:10.1145/1458082.1458274. URL http://dx.doi.org/10.1145/1458082.1458274. Cited on p. 45.

Urquhart, C.; Turner, J.; Durbin, J.; and Ryan, J. Jan. 2007. *Changes in information behavior in clinical teams after introduction of a clinical librarian service.* In Journal of the Medical Library Association : JMLA, vol. 95, no. 1, pp. 14–22. ISSN 1558-9439. URL http://view.ncbi.nlm.nih.gov/pubmed/17252062. Cited on p. 30.

USA Department of Health and Human Services. 2000. *Healthy People 2010*. Washington, DC. Cited on pp. 9, 150, and 183.

USA Department of Health and Human Services. Dec. 2010. *Healthy People 2020 objectives.* Tech. rep., U.S. Department of Health and Human Services, Washington, DC. Cited on p. 6.

Vaughan, Liwen. Jul. 2004. *New measurements for search engine evaluation proposed and tested*. In Information Processing and Management: an International Journal, vol. 40, no. 4, pp. 677–691. ISSN 0306-4573. doi:10.1016/s0306-4573(03)00043-8. URL http://dx.doi.org/10.1016/s0306-4573(03)00043-8. Cited on p. 68.

Volk, Martin; Ripplinger, Bärbel; Vintar, Spela; Buitelaar, Paul; Raileanu, Diana; and Sacaleanu, Bogdan. Dec. 2002. *Semantic annotation for concept-based cross-language medical information retrieval.* In International Journal of Medical Informatics, vol. 67, no. 1-3, pp. 97–112. ISSN 1386-5056. URL http://view.ncbi.nlm.nih.gov/pubmed/12460635. Cited on p. 132.

Voorhees, Ellen M. 2008. *On test collections for adaptive information retrieval.* In Inf. Process. Manage., vol. 44, no. 6, pp. 1879–1885. ISSN 0306-4573. doi:10.1016/j.ipm.2007.12.011. URL http://dx.doi.org/10.1016/j.ipm.2007.12.011. Cited on p. 66.

W3Techs. May 2013. *[duplicate] Usage of content languages for websites.* URL http://w3techs.com/technologies/overview/content_language/all.Accessed:2012-07-23.(ArchivedbyWebCiteathttp://www.webcitation.org/6GJge2VfQ). Cited on pp. 131 and 274.

Wang, Yunli and Liu, Zhenkai. 2005. *Personalized Health Information Retrieval System.* In AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium. ISSN 1942-597X. URL http://view.ncbi.nlm.nih.gov/pubmed/16779435. Cited on p. 150.

Weber, Ingmar and Castillo, Carlos. 2010. *The demographics of web search.* In Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '10, pp. 523–530. ACM, New York, NY, USA. ISBN 978-1-4503-0153-4. doi:10.1145/1835449.1835537. URL http://dx.doi.org/10.1145/1835449.1835537. Cited on p. 273.

Wen, Lei; Ruthven, Ian; and Borlund, Pia. 2006. *The Effects on Topic Familiarity on Online Search Behaviour and Use of Relevance Criteria*. In

Lalmas, Mounia; MacFarlane, Andy; Rüger, Stefan; Tombros, Anastasios; Tsikrika, Theodora; and Yavlinsky, Alexei, eds., Advances in Information Retrieval, vol. 3936 of *Lecture Notes in Computer Science*, chap. 40, pp. 456–459. Springer Berlin / Heidelberg, Berlin, Heidelberg. ISBN 978-3-540-33347-0. doi:10.1007/11735106_40. URL http://dx.doi.org/10.1007/11735106_40. Cited on p. 151.

Westbrook, Johanna I.; Coiera, Enrico W.; and Gosling, A. Sophie. 2005. *Do online information retrieval systems help experienced clinicians answer clinical questions?* In Journal of the American Medical Informatics Association : JAMIA, vol. 12, no. 3, pp. 315–321. ISSN 1067-5027. doi:10.1197/jamia.m1717. URL http://dx.doi.org/10.1197/jamia.m1717. Cited on p. 30.

White, Ryen; Capra, Rob; Golovchinsky, Gene; Kules, Bill; Russell, Dan; Smith, Catherine; and Tunkelang, Daniel. 2012. *Call for Papers for Special Issue on Human-Computer Information Retrieval.* URL http://research.microsoft.com/en-us/um/people/ryenw/docs/HCIR_IPM_SpecialIssue_CFP.pdf. Cited on p. 42.

White, Ryen W.; Dumais, Susan; and Teevan, Jaime. 2008. *How medical expertise influences web search interaction*. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '08, pp. 791–792. ACM, New York, NY, USA. ISBN 978-1-60558-164-4. doi:10.1145/1390334.1390506. URL http://dx.doi.org/10.1145/1390334.1390506. Cited on p. 148.

White, Ryen W.; Dumais, Susan T.; and Teevan, Jaime. 2009. *Characterizing the influence of domain expertise on web search behavior*. In Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM '09, pp. 132–141. ACM, New York, NY, USA. ISBN 978-1-60558-390-7. doi:10.1145/1498759.1498819. URL http://dx.doi.org/10.1145/1498759.1498819. Cited on pp. 148 and 167.

White, Ryen W. and Horvitz, Eric. Nov. 2009. *Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search*. In ACM Transactions on Information Systems (ACM TOIS), vol. 27, no. 4, pp. 23:1–23:37. Cited on p. 76.

Wildemuth, Barbara M. Feb. 2004. *The effects of domain knowledge on search tactic formulation*. In J. Am. Soc. Inf. Sci. Technol., vol. 55, no. 3, pp. 246–258. ISSN 1532-2882. doi:10.1002/asi.10367. URL http://dx.doi.org/10.1002/asi.10367. Cited on pp. 151, 184, 190, and 197.

Winograd, Terry. 2001. *Architectures for context*. In Hum.-Comput. Interact., vol. 16, no. 2, pp. 401–419. ISSN 0737-0024. URL http://portal.acm.org/citation.cfm?id=1463126. Cited on p. 38.

Wu, G. and Li, J. Oct. 1999. *Comparing Web search engine performance in searching consumer health information: evaluation and recommendations.* In Bulletin of the Medical Library Association, vol. 87, no. 4, pp. 456–461. ISSN 0025-7338. URL http://view.ncbi.nlm.nih.gov/pubmed/10550031. Cited on pp. 69, 70, and 73.

Yan, Xin; Song, Dawei; and Li, Xue. 2006. *Concept-based document readability in domain specific information retrieval*. In CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management, pp. 540–549. ACM, New York, NY, USA. ISBN 1595934332. doi:10.1145/1183614.1183692. URL http://dx.doi.org/10.1145/1183614.1183692. Cited on p. 31.

Yang, Shuang-Hong; Crain, Steven; and Hongyuan. 2011. *Bridging the Language Gap: Topic Adaptation for Documents with Different Technicality*. In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 823–831. Cited on pp. 272, 273, and 275.

Yoo, Eun-Young and Robbins, Louise S. Feb. 2008. *Understanding middle-aged women's health information seeking on the web: A theoretical approach*. In J. Am. Soc. Inf. Sci. Technol., vol. 59, no. 4, pp. 577–590. ISSN 1532-2882. doi:10.1002/asi.v59:4. URL http://dx.doi.org/10.1002/asi.v59:4. Cited on p. 29.

Yu, Hong and Kaufman, David. Jan. 2007. *A cognitive evaluation of four online search engines for answering definitional questions posed by physicians*. In Pacific Symposium on Biocomputing, pp. 328–339. University of Wisconsin-Milwaukee, Department of Health Sciences, 2400 E. Hartford Avenue, PO Box 413, Milwaukee, WI 53210, USA. ISSN 1793-5091. URL http://psb.stanford.edu/psb-online/proceedings/psb07/yu.pdf. Cited on p. 69.

Zarro, Michael and Lin, Xia. Oct. 2011. *Using Social Tags and Controlled Vocabularies As Filters for Searching and Browsing: A Health Science Experiment*. In The Fifth Workshop on Human-Computer Interaction and Information Retrieval. Cited on p. 200.

Zeng, Q.; Kogan, S.; Ash, N.; Greenes, R. A.; and Boxwala, A. A. 2002. *Characteristics of consumer terminology for health information retrieval*. In Methods of information in medicine, vol. 41, no. 4, pp. 289–298. ISSN 0026-1270. URL http://view.ncbi.nlm.nih.gov/pubmed/12425240. Cited on pp. 32, 148, 157, and 183.

Zeng, Q. T. and Crowell, J. 2006. *Semantic Classification of Consumer Health Content*. In MedNet. Cited on p. 31.

Zeng, Q. T.; Tse, T.; Crowell, J.; Divita, G.; Roth, L.; and Browne, A. C. 2005a. *Identifying consumer-friendly display (CFD) names for health concepts*. In AMIA Annu Symp Proc. 2005, pp. 859–863. DSG, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. ISSN 1559-4076. URL http://view.ncbi.nlm.nih.gov/pubmed/16779162. Cited on p. 32.

Zeng, Qing and Cimino, James J. 1997. *Linking a clinical system to heterogeneous information resources*. In Proc AMIA Annu Fall Symp., pp. 553–557. Cited on p. 51.

Zeng, Qing; Kim, Eunjung; Crowell, Jon; and Tse, Tony. 2005b. *A Text Corpora-Based Estimation of the Familiarity of Health Terminology*. In Biological and Medical Data Analysis, pp. 184–192. doi:10.1007/11573067_19. URL http://dx.doi.org/10.1007/11573067_19. Cited on p. 32.

Zeng, Qing T.; Crowell, Jonathan; Plovnick, Robert M.; Kim, Eunjung; Ngo, Long; and Dibble, Emily. 2006. *Assisting consumer health information retrieval with query recommendations.* In Journal of the American Medical Informatics Association : JAMIA, vol. 13, no. 1, pp. 80–90. ISSN 1067-5027. doi:10.1197/jamia.m1820. URL http://dx.doi.org/10.1197/jamia.m1820. Cited on pp. 33, 148, 149, 150, 197, 200, and 247.

Zeng-Treitler, Qing; Goryachev, Sergey; Tse, Tony; Keselman, Alla; and Boxwala, Aziz. May 2008. *Estimating Consumer Familiarity with Health Terminology: A Context-based Approach*. In J Am Med Inform Assoc, vol. 15, no. 3, pp. 349–356. doi:10.1197/jamia.m2592. URL http://dx.doi.org/10.1197/jamia.m2592. Cited on pp. 32 and 149.

Zhang, Y. Nov. 2008. *Complex adaptive filtering user profile using graphical models*. In Information Processing & Management, vol. 44, no. 6, pp. 1886–1900. ISSN 03064573. doi:10.1016/j.ipm.2008.08.001. URL http://dx.doi.org/10.1016/j.ipm.2008.08.001. Cited on p. 45.

Zhang, Yan. 2010. *Contextualizing consumer health information searching: an analysis of questions in a social Q&A community*. In Proceedings of the 1st ACM International Health Informatics Symposium, pp. 210–219. Cited on pp. 147 and 197.

Zhang, Yan. Oct. 2011. *A Review of Search Interfaces in Consumer Health Websites*. In The Fifth Workshop on Human-Computer Interaction and Information Retrieval. Cited on p. 197.

Zielstorff, R. Oct. 2003. *Controlled vocabularies for consumer health*. In Journal of Biomedical Informatics, vol. 36, no. 4-5, pp. 326–333. ISSN 15320464. doi:10.1016/j.jbi.2003.09.015. URL http://dx.doi.org/10.1016/j.jbi.2003.09.015. Cited on pp. 147 and 199.

Zou, Jie; Le, Daniel; and Thoma, George R. 2007. *Structure and content analysis for HTML medical articles: a hidden markov model approach*. In DocEng '07: Proceedings of the 2007 ACM symposium on Document engineering, pp. 199–201. ACM, New York, NY, USA. ISBN 9781595937766. doi:10.1145/1284420.1284468. URL http://dx.doi.org/10.1145/1284420.1284468. Cited on p. 31.

APPENDICES

INITIAL QUESTIONNAIRE OF USER EXPERIMENT 1

This is an English translated version of the questionnaire users answered before the task assessments of User Experiment 1. Users answered this questionnaire using a web form.

A.1 PERSONAL INFORMATION

User ID: _____

Age: _____

Gender: ( ) Male ( ) Female

Nationality: [ ] portuguese [ ] other: _____

Do you consider yourself healthy?

	1	2	3	4	5	
Not healthy	( )	( )	( )	( )	( )	**Very healthy**

A.2 WEB SEARCHES

For how many years have you been searching on the Web? _____

Do you carry out web searches at home or work?
( ) Yes ( ) No

How frequently do you search on the Web?
( ) Once a year ( ) Once a month ( ) Once a week ( ) Once a day
( ) More often

During web searches, how frequently do you find what you are searching for?

	1	2	3	4	5	
Never	( )	( )	( )	( )	( )	**Always**

A.3 HEALTH WEB SEARCHES

Have you ever conducted a web search about health subjects?
( ) Yes ( ) No

How frequently do you search on the Web about health subjects?
( ) Once a year ( ) Once a month ( ) Once a week ( ) Once a day
( ) More often

During web searches about health subjects, how frequently do you find what you are searching for?

	1	2	3	4	5	
Never	( )	( )	( )	( )	( )	**Always**

Do you feel any reluctance about conducting health searches in specific places?
( ) Yes ( ) No

What type of health searches do you usually conduct? [1 - Never; 5 - Frequently]

	1	2	3	4	5
Fact search (e.g.: clinician phone)	( )	( )	( )	( )	( )
Decision search (e.g.: best treatment for a condition)	( )	( )	( )	( )	( )
Exploratory search (e.g.: know about allergies)	( )	( )	( )	( )	( )

When you conduct web searches about health subjects, what difficulties do you have?

A.4 FIRST INFORMATION NEED

Which was the first selected information need? _____

Have you ever searched for the topic related to the first information need?
( ) Yes ( ) No

Describe what you already know about the topic related to the first information need.

Write queries that could be used to search for information to satisfy the first information need. [Write one query per line.]

From the above queries, which one do you prefer?

A.5 SECOND INFORMATION NEED

Which was the second selected information need? _____

Have you ever searched for the topic related to the second information need?
( ) Yes ( ) No

Describe what you already know about the topic related to the second information need.

Write queries that could be used to search for information to satisfy the second information need. [Write one query per line.]

From the above queries, which one do you prefer?

FINAL QUESTIONNAIRE OF USER EXPERIMENT 1

This is an English translated version of the questionnaire users answered after completing the assessment tasks of User Experiment 1. Users answered this questionnaire using a web form.

B.1 PERSONAL INFORMATION

User ID: _____

B.2 SEARCH ENGINES

Which were the 4 search engines you selected?
[ ] Bing (http://www.bing.com)
[ ] Google (http://www.google.com)
[ ] MedlinePlus (http://medlineplus.gov)
[ ] Sapo (http://www.sapo.pt)
[ ] Sapo Saúde (http://saude.sapo.pt)
[ ] WebMD (http://webmd.com)
[ ] Yahoo! (http://www.yahoo.com)

B.3 FIRST INFORMATION NEED

I chose this information need because:
[ ] it was interesting
[ ] it was familiar
[ ] it was easy
[ ] no reason
[ ] other: _____

The task associated with this information need was:

	1	2	3	4	5	
Unclear	( )	( )	( )	( )	( )	**Clear**
Easy	( )	( )	( )	( )	( )	**Complex**
Unfamiliar	( )	( )	( )	( )	( )	**Familiar**

I had an exact idea of the type of information I wanted.

	1	2	3	4	5	
Disagree	( )	( )	( )	( )	( )	**Agree**

I believe I have succeeded in this task.

	1	2	3	4	5	
Disagree	( )	( )	( )	( )	( )	**Agree**

I think there is better information available than the one I have found.

	1	2	3	4	5	
Disagree	( )	( )	( )	( )	( )	**Agree**

To know the queries that other users have formulated for this information need would have helped.

	1	2	3	4	5	
Disagree	( )	( )	( )	( )	( )	**Agree**

To know the queries that other users have formulated for the work task associated with this information need would have helped.

	1	2	3	4	5	
Disagree	( )	( )	( )	( )	( )	**Agree**

What difficulties did you have on the task associated with this information need?

_____

_____

_____

B.4 SECOND INFORMATION NEED

I chose this information need because:
[ ] it was interesting
[ ] it was familiar
[ ] it was easy
[ ] no reason
[ ] other: _____

The task associated with this information need was:

	1	2	3	4	5	
Unclear	( )	( )	( )	( )	( )	**Clear**
Easy	( )	( )	( )	( )	( )	**Complex**
Unfamiliar	( )	( )	( )	( )	( )	**Familiar**

I had an exact idea of the type of information I wanted

	1	2	3	4	5	
Disagree	( )	( )	( )	( )	( )	**Agree**

I believe I have succeeded in this task

	1	2	3	4	5	
Disagree	( )	( )	( )	( )	( )	**Agree**

I think there is better information available than the one I have found

	1	2	3	4	5	
Disagree	( )	( )	( )	( )	( )	**Agree**

To know the queries that other users have formulated for this information need would have helped.

	1	2	3	4	5	
Disagree	( )	( )	( )	( )	( )	**Agree**

To know the queries that other users have formulated for the work task associated with this information need would have helped.

	1	2	3	4	5	
Disagree	( )	( )	( )	( )	( )	**Agree**

What difficulties did you have on the task associated with this information need?

STATISTICAL DETAILS OF THE COMPARATIVE EVALUATION OF SEARCH ENGINES IN HIR

In this appendix we present the statistical details of the significant differences found in the study described in Chapter 5. The following tables contain the observed difference, its level of significance (** for α=0.01 and * for α=0.05), the test value and, when different from 0, the p-value. The applied tests vary depending on each specific case. In cases where the assumptions to apply a parametric test have been met, we applied a t-test in the pairwise comparison (t(df)=test-value). When these assumptions have not been met, we applied the Mann-Whitney test (w=test-value). When there were ties, we had to use an Exact Mann-Whitney test (z=test-value). In the situations in which we applied the Tukey test, we present the confidence interval (lower, upper). The level of significance was always adjusted in the pairwise comparison except in the case of Tukey tests where this value is already adjusted.

Table C.1: Statistical Results on the Overall Analysis

GAP	AP	gP@5	P@5	gP@10	P@10
			Search Engine Type (NH-Non-Health; H-Health)		
NH>H**	NH>H**	NH>H**	NH>H**	NH>H**	NH>H**
t(260.83)=-3.50	t(264.29)=-3.85	t(250.15)=-3.55	t(234.60)=-4.22	t(251.02)=-4.15	t(233.20)=-4.51
			Search Engine (B-Bing; G-Google; M-MedlinePlus; S-Sapo; SS-Sapo Saúde; W-WebMD; Y-Yahoo!)		
		B>SS**		B>SS**	B>SS**
		t(96.57)=3.21		t(96.33)=3.77	t(94.80)=4.02
G>B**	G>B**	G>B**	G>B**	G>B**	G>B**
t(93.93)=-3.69	t(89.64)=-3.74	t(100.47)=-4.19	t(85.08)=-4.90	t(98.56)=-4.09	t(85.39)=-5.20
G>M*	G>M**	G>M**	G>M**	G>M**	G>M**
t(80.91)=3.46	t(76.42)=3.63	t(81.00)=3.52	t(66.41)=4.08	t(76.66)=3.91	t(63.63)=4.52
		G>S**	G>S**	G>S**	G>S**
		w=1932	t(36.63)=5.89	w=1848.5	w=1904
G>SS**	G>SS**	G>SS**	G>SS**	G>SS**	G>SS**
t(90.72)=5.21	t(93.12)=6.06	t(94.31)=7.64	t(74.03)=8.05	t(104.32)=8.70	t(75.31)=9.36
G>W*	G>W*		G>W**		G>W**
w=1633	w=1672.5		w=1709.5		w=1650.5
		G>Y**	G>Y**	G>Y*	G>Y**
		t(69.46)=3.70	t(53.92)=4.04	t(59.52)=3.24	t(50.57)=4.15
		M>SS**		M>SS*	
		t(90.40)=3.23		t(83.46)=3.06	
		Y>SS**		Y>SS*	
		t(81,26)=-3.09		t(67.4)=-3.18	
			Clinical Question (O-Overview; D/S-Diagnosis/Symptoms; T-Treatment; P/S-Prevention/Screening; DM-Disease Management)		
		O>P/S*	O>P/S*		O>P/S**
		w=3458.5	t(141.57)=3.39		(-0.28, -0.02)
					p=0.01
		D/S>P/S*	D/S>P/S*		
		w=4829.5	t(169.97)=3.11		
			Specialty (D-Dermatology; G-Gynaecology; P-Psichiatry; U-Urology)		
		D<G*			
		w=1270.5			
		p=0.01			
		D<P**	D<P**	D<P*	D<P*
		w=3258	t(112.46)=-3.76	(0.01, 0.23)	(0.01, 0.23)
				p=0.02	p=0.03
			Severity (S-Severe; NS-Non-Severe)		
S>NS**	S>NS**	S>NS**	S>NS**	S>NS**	S>NS**
t(89.22)=2.71	t(96.22)=2.95	w=9838	t(91.94)=3.61	t(97.33)=3.08	t(91.16)=2.77

Table C.2: Statistical results in the clinical question analysis by search engine type

	GAP	AP	gP@5	P@5	gP@10	P@10
	Clinical Question (O-Overview; D/S-Diagnosis/Symptoms; T-Treatment; P/S-Prevention/Screening; DM-Disease Management)					
NH				O>P/S*	O>P/S*	O>P/S*
				(-0.34,-0.02)	(-0.31,-0.02)	(-0.31,-0.02)
				p=0.02	p=0.02	p=0.02
NH				O>T*	O>T*	O>T*
				(-0.36,-0.01)	(-0.34,-0.02)	(-0.34,-0.02)
				p=0.04	p=0.01	p=0.01
	Search Engine Type (NH-Non-Health; H-Health)					
O	NH>H**	NH>H**	NH>H**	NH>H**	NH>H**	NH>H**
	t(62.96)=3.52	t(62.66)=3.68	t(50.87)=2.83	t(43.23)=3.18	t(43.60)=2.86	t(43.60)=2.86
D/S	NH>H**	NH>H**	NH>H**	NH>H**	NH>H**	NH>H**
	t(88.17)=4.07	t(86.05)=4.07	t(74.45)=2.46	w=1644.5	w=1725	w=1725
			p=0.01			
P/S		NH>H*	NH>H*	NH>H**	NH>H*	NH>H**
		z=1.75	z=2.17	z=2.97	z=1.98	t(55.34)=2.86
		p=0.04	p=0.01		p=0.02	

Table C.3: Statistical results in the clinical question analysis by search engine

	GAP	AP	gP@5	P@5	gP@10	P@10
Clinical Question (O-Overview; D/S-Diagnosis/Symptoms; T-Treatment; P/S-Prevention/Screening; DM-Disease Management)						
Y						O>T*
						z=2.69
Search Engine (B-Bing; G-Google; M-MedlinePlus; S-Sapo; SS-SapoSaúde; W-WebMD; Y-Yahoo!)						
O	G>SS*	G>SS**	G>SS**	G>SS**	G>SS**	G>SS**
	z=2.97	(-0.63,-0.08)	z=3.35	z=3.51	z=3.47	z=3.76
D/S	G>SS**	G>SS**	G>SS**	G>SS**	G>SS**	G>SS**
	t(37.80)=6.74	z=4.77	z=3.62	z=3.87	z=4.46	z=4.65
D/S		G>M**				
		w=323				
P/S				G>S**		G>S**
				z=3.43		(-0.66, -0.07)
P/S			G>SS**		G>SS*	G>SS**
			(-0.76,0.16)		z=2.85	(-0.59,-0.06)
P/S				G>W**	G>W**	G>W*
				z=3.41	z=3.17	(-0.65,-0.05)
T						G>Y**
						z=3.41

Table C.4: Statistical results in the medical specialty analysis by search engine type

	GAP	AP	gP@5	P@5	gP@10	P@10
	Specialty (D-Dermatology; G-Gynaecology; P-Psichiatry; U-Urology)					
NH			G>D*	G>D**		G>D*
			(-0.04,0.42)	(0.04,0.41)		(0.03,0.36)
			p=0.01			
	Search Engine Type (NH-Non-Health; H-Health)					
D			NH>H*	NH>H**	NH>H**	NH>H*
			z=2.13	z=2.41	z=2.36	t(46.06)=2.01
			p=0.02	p=0.01	p=0.01	p=0.03
G			NH>H**	NH>H**	NH>H**	NH>H**
			z=3.54	z=3.45	z=3.07	t(47.31)=4.44
P	NH>H*	NH>H*			NH>H*	NH>H*
	t(120.14)=1.75	t(121.82)=1.87			t(115.22)=1.82	t(107.32)=1.85
	p=0.04	p=0.03			p=0.04	p=0.03
U	NH>H*	NH>H*			NH>H*	NH>H*
	t(30.31)=1.98	t(30.42)=2.19			z=1.68	z=1.98
	p=0.03	p=0.02			p=0.05	p=0.02

Table C.5: Statistical results in the medical specialty analysis by search engine

	GAP	AP	gP@5	P@5	gP@10	P@10
				Specialty (D-Dermatology; G-Gynaecology; P-Psichiatry; U-Urology)		
M			P>D* (0.01,0.68) p=0.04			
W			P>D* (0.05,0.75) p=0.02			
				Search Engine (B-Bing; G-Google; M-MedlinePlus; S-Sapo; SS-SapoSaúde; W-WebMD; Y-Yahoo!)		
D				G>M* (-0.65,-0.02)		
D				G>S** (-0.86,-0.15)		
D				G>SS* (-0.60,-0.02)		
D				G>W** (-0.73,-0.11)		
G						G>M* (-0.62,-0.05)
G	G>SS** z=2.88	G>SS** z=2.93	G>SS** z=3.42	G>SS** z=3.62	G>SS** z=3.73	G>SS** (-0.79,-0.23)
G						Y>SS* (0.01,0.73)
P			G>S** z=3.75	G>S** z=3.38	G>S** z=2.94	G>S** z=3.39
P	G>SS** z=3.69	G>SS** z=4.27	G>SS** z=4.44	G>SS** z=4.33	G>SS** z=4.59	G>SS** z=5.09
U						G>SS* z=2.76
U					G>S** z=2.95	

Table C.6: Statistical results in the severity analysis by search engine type

	GAP	AP	gP@5	P@5	gP@10	P@10
				Severity (S-Severe; NS-Non-Severe)		
NH	S>NS**	S>NS**	S>NS**	S>NS**	S>NS*	S>NS*
	t(55.76)=-2.89	t(60.30)=-3.03	t(54.42)=-2.64	t(49.88)=-2.54	t(54.87)=-2.20	t(52.33)=-2.33
			p=0.01	p=0.01	p=0.02	p=0.01
W			S>NS*	S>NS*	S>NS*	
			w=838	w=825	w=883.5	
			p=0.02	p=0.02	p=0.04	
				Search Engine Type (NH-Non-Health; H-Health)		
NS			NH>H*	NH>H**	NH>H**	NH>H*
			z=2.13	z=2.41	z=2.35	t(46.06)=2.01
			p=0.01	p=0.01	p=0.01	p=0.03
S	NH>H**	NH>H**	NH>H**	NH>H**	NH>H**	NH>H**
	t(214.70)=3.57	t(216.58)=3.78	t(195.53)=2.93	t(181.33)=3.59	t(198.48)=3.54	t(184.85)=4.06

Table C.7: Statistical results in the severity analysis by search engine

	GAP	AP	gP@5	P@5	gP@10	P@10
				Severity (S-Severe; NS-Non-Severe)		
B	S>NS*	S>NS*				
	w=137	w=132				
	p=0.02	p=0.01				
G	S>NS*	S>NS*				
	w=312	t(23.65)=-2.28				
	p=0.02	p=0.02				
M			S>NS**	S>NS*		
			z=-2.44	z=-1.95		
				p=0.03		
S	S>NS*			S>NS*		
	t(11.24)=-2.04			t(8.32)=-2.65		
	p=0.03			p=0.01		
W	S>NS**	S>NS**	S>NS*	S>NS*	S>NS*	
	t(18.28)=-2.73	t(19.24)=-2.81	z=-2.15	z=-2.2	z=-1.82	
	p=0.01		p=0.02	p=0.01	p=0.04	
			Search Engine (B-Bing; G-Google; M-MedlinePlus; S-Sapo; SS-SapoSaúde; W-WebMD; Y-Yahoo!)			
NS			G>S*	G>S*	G>S*	
			z=2.61	t(8.56)=4.60	(-0.61,0.02)	
			p=0.05		p=0.03	
NS					G>SS*	
					(-0.56,0.01)	
					p=0.04	
NS				G>W**		
				t(14.90)=4.3		
S			G>B**	G>B**		G>B**
			w=1786	w=717		w=1856.5
S			B>SS**		B>SS**	B>SS**
			z=3.61		z=3.52	z=3.65
S	G>M**	G>M**			G>M**	G>M**
	w=1833	t(64.75)=3.54			w=1748.5	w=1773.5
S			G>S**	G>S**	G>S**	G>S**
			w=1281.5	w=1249.5	w=1222.5	w=1263.5
S	G>SS**	G>SS**	G>SS**	G>SS**	G>SS**	G>SS**
	w=2168	t(80.69)=6.53	w=2271	w=2272	w=2319	w=2419.5
S	G>Y*			G>Y**		G>Y*
	w=1568			w=1633.5		w=1633
S			M>SS**			M>SS*
			z=3.44			z=3.37
S			W>SS*		W>SS*	W>SS*
			z=-3.07		z=2.96	z=3.12
S			Y>SS**		Y>SS*	Y>SS*
			z=-3.57		z=3.25	z=3.11

INITIAL QUESTIONNAIRE OF USER EXPERIMENT 2

This is an English translated version of the questionnaire users answered before the task assessments of User Experiment 2. Users answered this questionnaire using a web form.

D.1 PERSONAL INFORMATION

User ID: _____

Age: _____

Gender: ( ) Male ( ) Female

Nationality: [ ] portuguese [ ] other: _____

Do you consider yourself healthy?

	1	2	3	4	5	
Not healthy	( )	( )	( )	( )	( )	**Very healthy**

D.2 WEB SEARCHES

For how many years have you been searching on the Web? _____

How frequently do you search on the Web?
( ) Once a year ( ) Once a month ( ) Once a week ( ) Once a day
( ) More often

During web searches, how frequently do you find what you are searching for?

	1	2	3	4	5	
Never	( )	( )	( )	( )	( )	**Always**

D.3 HEALTH WEB SEARCHES

Have you ever conducted a web search about health subjects?
( ) Yes ( ) No

How frequently do you search on the Web about health subjects?
( ) Once a year ( ) Once a month ( ) Once a week ( ) Once a day
( ) More often

During web searches about health subjects, how frequently do you find what you are searching for?

	1	2	3	4	5	
Never	( )	( )	( )	( )	( )	**Always**

When you conduct web searches about health subjects, what difficulties do you have?

How frequently do you conduct your health web searches with the following languages? [1 - Never; 5 - Frequently]

	1	2	3	4	5
Portuguese	( )	( )	( )	( )	( )
English	( )	( )	( )	( )	( )
Other language	( )	( )	( )	( )	( )

During web searches about health subjects, how frequently do you use medico-scientific terminology (e.g.: using *pyrosis* instead of the lay term *heartburn*)?

	1	2	3	4	5	
Never	( )	( )	( )	( )	( )	**Always**

On the Internet, how do you usually satisfy your health information needs? [1 - Never; 5 - Frequently]

	1	2	3	4	5
Web pages	( )	( )	( )	( )	( )
Blogs	( )	( )	( )	( )	( )
Foruns	( )	( )	( )	( )	( )
Social networks	( )	( )	( )	( )	( )
Chat	( )	( )	( )	( )	( )
Newsletters	( )	( )	( )	( )	( )
RSS feeds	( )	( )	( )	( )	( )

D.4 TOPIC FAMILIARITY

Do you know the meaning of the following concepts?

	Yes	No
alopecia	( )	( )
artralgia	( )	( )
disúria	( )	( )
eritema	( )	( )
estomatite	( )	( )
hiperuricemia	( )	( )
pirose	( )	( )
prurido	( )	( )

D.5 SEARCH QUERIES

Formulate a query for the following health information situations.

About 3 days ago, I started having a burning feeling every time I urinated. How should I treat this?

For the past 5 days my head has been very itchy and I don't have lice. What can I do to stop the itching?

I have high uric acid (8.0 mg/dL) with reference units 3.6 - 7.7. How can I lower my uric acid level?

I am suffering with an inflammation on my lips and mouth area for more than a year. I have difficulties eating. What can I do to treat it?

My father got bit by a dog and is in the hospital with a bone infection. How is this treated?

I frequently get heartburn even when I stay away from spicy stuff. What can I do to prevent it?

I have been noticing lots of hair coming out from my head. Usually I only comb my hair once a day. What can I do to stop losing my hair?

I'm on the computer all day so I type a lot and use the mouse. My right pointing finger is starting to give me some joint pain. How I can treat my finger?

TASK QUESTIONNAIRE OF USER EXPERIMENT 2

This is an English translated version of the questionnaire users answered after each task of User Experiment 2. Users answered this questionnaire using a web form.

User ID: _____

Task ID: _____

Have you ever searched bout this topic before: ( ) Yes ( ) No

I had an exact idea of the type of information I wanted.

	1	2	3	4	5	
Disagree	( )	( )	( )	( )	( )	**Agree**

The task associated with this information need was:

	1	2	3	4	5	
Unclear	( )	( )	( )	( )	( )	**Clear**
Easy	( )	( )	( )	( )	( )	**Complex**
Unfamiliar	( )	( )	( )	( )	( )	**Familiar**

I believe I have succeeded in this task.

	1	2	3	4	5	
Disagree	( )	( )	( )	( )	( )	**Agree**

Formulate a query for the information need associated with this task.

_____

Formulate another query for the information need associated with this task.

_____

What treatments did you found for the condition associated with this task?

_____

_____

_____

What difficulties did you felt during this task?

QUIZ TO EVALUATE ENGLISH PROFICIENCY IN USER EXPERIMENT 2

F.1 ENGLISH GRAMMAR I

1. Juan _____ in the library this morning.
    a) is study
    b) studying
    c) is studying
    d) are studying
2. Maria _____ never late for work.
    a) am
    b) is
    c) are
    d) were
3. The company will upgrade _____ computer information systems next month.
    a) there
    b) their
    c) it's
    d) its
4. Cheryl likes apples, _____ she does not like oranges.
    a) for
    b) so
    c) but
    d) or
5. You were _____ the New York office before 2 p.m.
    a) supposed calling
    b) supposed to call

c) supposed call

d) suppose call

6. When I graduate from college next June, I _____ a student here for five years.

   a) has been

   b) will have been

   c) will have

   d) have been

7. Ms. Guth _____ rather not invest that money in the stock market.

   a) would

   b) could

   c) must

   d) has to

8. Alicia, _____ the windows please. It's too hot in here.

   a) opens

   b) opened

   c) open

   d) will opened

9. The movie was _____ the book.

   a) as

   b) as good

   c) as good as

   d) good as

10. Eli's hobbies include jogging, swimming, and _____.

    a) climbing mountains

    b) climb mountains

    c) to climb mountains

    d) to climb

11. Mr. Hawkins requests that someone _____ the data by fax immediately.

    a) send

    b) sends

    c) to send

    d) sent

12. Who is _____, Marina or Sachiko?

a) taller
b) tallest
c) the tallest
d) tall

13. The concert will begin _____fifteen minutes.

a) in
b) on
c) about
d) with

14. I have only a _____ Christmas cards left to write.

a) fewer
b) less
c) few
d) little

15. Each of the Olympic athletes _____ for months, even years.

a) were training
b) has been training
c) have been training
d) been training

F.2 ENGLISH GRAMMAR II

Select the one underlined word or phrase that is incorrect.

1. He goes never to the company softball games.

a) never
b) games
c) softball
d) the

2. The majority to the news is about violence or scandal.

a) violence
b) news
c) The
d) to

3. When our vacation, we plan to spend three days scuba diving.

a) days

b) plan
c) diving
d) When

4. Mr. Olsen is telephoning a American Red Cross for help.

a) Red
b) is
c) for
d) a

5. Each day after school, Jerome run five miles.

a) miles
b) run
c) Each
d) after

6. I had a enjoyable time at the party last night.

a) time
b) last
c) a
d) at

7. Do you know the student who books were stolen?

a) know
b) Do
c) were
d) who

8. Frederick used work for a multinational corporation when he lived in Malaysia.

a) multinational
b) lived in
c) when
d) used work

9. Jean-Pierre will spend his vacation either in Singapore nor the Bahamas.

a) will
b) nor
c) his
d) Bahamas

10. I told the salesman that I was not interesting in buying the latest model.

a) that
b) interesting
c) told
d) buying

11. Takeshi swimmed one hundred laps in the pool yesterday.

a) swimmed
b) in
c) yesterday
d) hundred

12. Mr. Feinauer does not take critical of his work very well.

a) does
b) well
c) critical
d) his

13. Yvette and Rinaldo send e-mail messages to other often.

a) other
b) often
c) and
d) send

14. The doctor him visited the patient's parents.

a) The
b) him
c) patient's
d) visited

15. Petra intends to starting her own software business in a few years.

a) software
b) intends
c) starting
d) few

F.3 ENGLISH VOCABULARY

1. We were _____ friends in that strange but magical country.

a) toward
b) among
c) in addition to

d) upon

2. The hurricane caused _____ damage to the city.

   a) extension
   b) extensive
   c) extend
   d) extended

3. Many cultures have special ceremonies to celebrate a person's _____ of passage into adulthood.

   a) right
   b) writ
   c) rite
   d) write

4. Do you _____ where the nearest grocery store is?

   a) know
   b) no
   c) now
   d) not

5. Jerry Seinfeld, the popular American comedian, has his audiences _____.

   a) putting too many irons in the fire
   b) rolling in the aisles
   c) keeping their noses out of someone's business
   d) going to bat for someone

6. The rate of _____ has been fluctuating wildly this week.

   a) exchange
   b) bills
   c) money
   d) coins

7. The bus _____ arrives late during bad weather.

   a) yesterday
   b) later
   c) always
   d) every week

8. The chairperson will _____ members to the subcommittee.

   a) appoint
   b) disappoint

c) disappointed

d) appointment

9. The critics had to admit that the ballet _____ was superb.

   a) performance

   b) pathology

   c) procrastinate

   d) psychosomatic

10. Peter says he can't _____ our invitation to dinner tonight.

    a) almost

    b) angel

    c) accept

    d) across

F.4 ENGLISH READING COMPREHENSION

1. The B&B Tour
   Spend ten romantic days enjoying the lush countryside of southern England. The counties of Devon, Dorset, Hampshire, and Essex invite you to enjoy their castles and coastline, their charming bed and breakfast inns, their museums and their cathedrals. Spend lazy days watching the clouds drift by or spend active days hiking the glorious hills. These fields were home to Thomas Hardy, and the ports launched ships that shaped world history. Bed and breakfasts abound, ranging from quiet farmhouses to lofty castles. Our tour begins August 15. Call or fax us today for more information 1-800-222-XXXX. Enrollment is limited, so please call soon.

   How many people can go on this tour?

   a) a limited number

   b) 2-8

   c) 10

   d) an unlimited number

   What can we infer about this area of southern England?

   a) The region has lots of vegetation.

   b) The coast often has harsh weather.

   c) The sun is hot and the air is dry.

   d) The land is flat.

   Which of the following counties is not included in the tour?

   a) Cornwall

   b) Hampshire

c) Essex

d) Devon

2. Directions to Erik's house:

   Leave Interstate 25 at exit 7S. Follow that road (Elm Street) for two miles. After one mile, you will pass a small shopping center on your left. At the next set of traffic lights, turn right onto Maple Drive. Erik's house is the third house on your left. It's number 33, and it's white with green trim.

   Which is closest to Erik's house?

   a) a greenhouse

   b) the traffic lights

   c) exit 7S

   d) the shopping center

   What is Erik's address?

   a) 2 Elm Street

   b) 13 Erika Street

   c) Interstate 25

   d) 33 Maple Drive

3. Date: May 16, 1998 To: Megan Fallerman From: Steven Roberts Subject: Staff Meeting Please be prepared to give your presentation on the monthly sales figures at our upcoming staff meeting. In addition to the accurate accounting of expenditures for the monthly sales, be ready to discuss possible reasons for fluctuations as well as possible trends in future customer spending. Thank you.

   Who will give the presentation?

   a) Steven Roberts

   b) Megan Fallerman

   c) the company president

   d) future customers

   The main focus of the presentation will be _____.

   a) monthly salary figures

   b) monthly sales figures

   c) the company president

   d) staff meeting presentations

4. Anna Szewcyzk, perhaps the most popular broadcaster in the news media today, won the 1998 Broadcasting Award. She got her start in journalism as an editor at the Hollsville County Times in Missouri. When the newspaper went out of business, a colleague persuaded her to enter the field of broadcasting. She moved to Oregon to begin a master's degree in broadcast journalism at Atlas University. Following graduation,

she was able to begin her career as a local newscaster with WPSU-TV in Seattle, Washington, and rapidly advanced to national television. Noted for her quick wit and trenchant commentary, her name has since become synonymous with Good Day, America! Accepting the award at the National Convention of Broadcast Journalism held in Chicago, Ms. Szewcyzk remarked, "I am so honored by this award that I'm at a total loss for words!" Who would ever have believed it?

What is the purpose of this announcement?

a) to encourage college students to study broadcasting
b) to advertise a job opening at the Hollsville County Times
c) to invite people to the National Convention of Broadcast Journalism
d) to recognize Ms. Szewcyzk's accomplishments

What was Ms. Szewczyk's first job in journalism?

a) She was a newscaster in Oregon.
b) She was a T.V. announcer in Washington.
c) She was a talk show host in Chicago.
d) She was an editor for a newspaper in Missouri.

The expression "to become synonymous with" means

a) to be the opposite of.
b) to be discharged from.
c) to be in sympathy with.
d) to be the same as.

TRANSLATED VERSION OF SAHLSA

Table G.1: Translated version of SAHLSA. In parentheses, the original Spanish concept.

Stem	**Key or Distracter**	
próstata (próstata)	glândula (glándula)	circulação (circulación)
emprego (empleo)	trabalho (trabajo)	educação (educación)
menstrual (menstrual)	mensal (mensual)	diário (diario)
gripe (gripe)	saudável (sano)	doente (enfermo)
avisar (avisar)	medir (medir)	dizer (decir)
refeição (comidas)	jantar (cena)	passeio (paseo)
alcoolismo (alcoholismo)	vício (adicción)	lazer (recreo)
gordura (grasa)	laranja (naranja)	manteiga (manteca)
asma (asma)	respirar (respirar)	pele (piel)
cafeína (cafeína)	energia (energía)	água (agua)
osteoporose (osteoporosis)	osso (hueso)	músculo (músculo)
depressão (depresión)	apetite (apetito)	sentimentos (sentimientos)
obstipação (estreñimiento)	bloqueado (bloqueado)	solto (suelto)
gravidez (embarazo)	parto (parto)	infância (niñez)
incesto (incesto)	família (familia)	vizinhos (vecinos)
pílula (pastilla)	comprimido (tableta)	bolacha (galleta)
testículo (testículo)	óvulo (óvulo)	esperma (esperma)
retal (rectal)	chuveiro (egadera)	sanita (inodoro)
olho (ojo)	ouvir (oír)	ver (ver)
irritação (irritación)	tensão (rígido)	sem dor (adolorido)
anormal (abnormal)	diferente (diferente)	similar (similar)
stress (estrés)	preocupação (preocupación)	felicidade (feliz)
aborto espontâneo (aborto espontáneo)	perda (pérdida)	casamento (matrimonio)
icterícia (ictericia)	amarelo (amarillo)	branco (blanco)
papanicolau (papanicolaou)	teste (prueba)	vacina (vacuna)
impetigo (impétigo)	cabelo (pelo)	pele (piel)
indicação (indicado)	instrução (instrucción)	decisão (decisión)
ataque (ataque)	ferida (herida)	são (sano)
menopausa (menopausia)	mulher (señoras)	menina (niñas)
apêndice (apéndice)	coçar (rascar)	dor (dolor)
comportamento (comportamiento)	pensamento (pensamiento)	conduta (conducta)
nutrição (nutrición)	saudável (saludable)	refrigerante (gaseosa)
diabetes (diabetes)	açúcar (azúcar)	sal (sal)
sífilis (sífilis)	contracetivo (anticonceptivo)	preservativo (condón)
inflamatório (inflamatorio)	inchaço (hinchazón)	suor (sudor)
hemorróides (hemorroides)	veias (venas)	coração (corazón)
herpes (herpes)	ar (aire)	sexo (sexo)
alérgico (alérgico)	resistência (resistencia)	reação (reacción)
rim (riñón)	urina (orina)	febre (fiebre)
calorias (calorías)	alimentos (alimentos)	vitaminas (vitaminas)
medicamento (medicamento)	instrumento (instrumento)	tratamento (tratamiento)
anemia (anemia)	sangue (sangre)	nervo (nervio)
intestinos (intestinos)	digestão (digestión)	suor (sudor)
potássio (potasio)	mineral (mineral)	proteína (proteína)
colite (colitis)	intestino (intestino)	bexiga (vejiga)
obesidade (obesidad)	peso (pes)o	altura (altura)
hepatite (hepatitis)	pulmão (pulmó)n	fígado (hígado)
vesícula biliar (vesícula biliar)	artéria (arteria)	órgão (órgano)
convulsões (convulsiones)	tontura (mareado)	tranquilo (tranquilo)
artrite (artritis)	estômago (estómago)	articulação (articulación)

INITIAL QUESTIONNAIRE OF USER EXPERIMENT 3

This is an English translated version of the questionnaire users answered before the task assessments of User Experiment 3. Users answered this questionnaire using a web form.

H.1 PERSONAL INFORMATION

User ID: _____

Age: _____

Gender: ( ) Male ( ) Female

Nationality: [ ] portuguese [ ] other: _____

Do you consider yourself healthy?

	1	2	3	4	5	
Not healthy	( )	( )	( )	( )	( )	**Very healthy**

H.2 WEB SEARCHES

For how many years have you been searching on the Web? _____

How frequently do you search on the Web?
( ) Once a year ( ) Once a month ( ) Once a week ( ) Once a day
( ) More often

During web searches, how frequently do you find what you are searching for?

	1	2	3	4	5	
Never	( )	( )	( )	( )	( )	**Always**

H.3 HEALTH WEB SEARCHES

Have you ever conducted a web search about health subjects?
( ) Yes ( ) No

How frequently do you search on the Web about health subjects?
( ) Once a year ( ) Once a month ( ) Once a week ( ) Once a day
( ) More often

During web searches about health subjects, how frequently do you find what you are searching for?

	1	2	3	4	5	
Never	( )	( )	( )	( )	( )	**Always**

When you conduct web searches about health subjects, what difficulties do you have?

How frequently do you conduct your health web searches with the following languages? [1 - Never; 5 - Frequently]

	1	2	3	4	5
Portuguese	( )	( )	( )	( )	( )
English	( )	( )	( )	( )	( )
Other language	( )	( )	( )	( )	( )

During web searches about health subjects, how frequently do you use medico-scientific terminology (e.g.: using *pyrosis* instead of the lay term *heartburn*)?

	1	2	3	4	5	
Never	( )	( )	( )	( )	( )	**Always**

On the Internet, how do you usually satisfy your health information needs? [1 - Never; 5 - Frequently]

	1	2	3	4	5
Web pages	( )	( )	( )	( )	( )
Blogs	( )	( )	( )	( )	( )
Foruns	( )	( )	( )	( )	( )
Social networks	( )	( )	( )	( )	( )
Chat	( )	( )	( )	( )	( )
Newsletters	( )	( )	( )	( )	( )
RSS feeds	( )	( )	( )	( )	( )

H.4 QUERY SUGGESTION

How useful do you find to be suggested with translated queries? [1 - Not useful at all; 5 - Extremely useful; as example, *pyrosis* is the medico-scientific translation of the lay term *heartburn*]

	1	2	3	4	5
from lay terminology to medico-scientific terminology	( )	( )	( )	( )	( )
from medico-scientific terminology to lay terminology	( )	( )	( )	( )	( )
from Portuguese to English	( )	( )	( )	( )	( )

H.5 PRE-SEARCH KNOWLEDGE

Please answer the best you can to the following questions without consulting external sources of information.

Your mother has just been diagnosed with breast cancer but, shocked with the news, she was not able to ask everything she wanted to the doctor. She only remembers hearing about ductal breast cancer and she is interested in knowing more about her treatment options. Help her and find 4 types of treatments for ductal breast cancer. Give a comprehensive answer, indicating what circumstances may condition each treatment (for example: type of cancer and tumor size).

Since the summer, your brother has been feeling a severe pain in the leg, from the buttocks to the knee. A friend has told him that he had the same symptom in the past and was diagnosed with a sciatica. Help your brother knowing how this pain can be treated pointing 3 ways to reduce his symptoms.

Two weeks ago, someone from your family has been diagnosed with shingles. To understand what characterizes this disease you decided to what are its causes and symptoms. Find out what causes the disease and identify two common symptoms.

You suffer from the Irritable Bowel Syndrome. Point out 4 possible ways to alleviate the symptoms. Give a complete answer, indicating what constrains each form of reducing the symptoms.

Your younger brother is 2 years old and was diagnosed with atopic dermatitis. Indicate 4 ways to reduce or treat the symptoms associated with this condition.

You have just been painfully stung by an insect and, not knowing what insect it was, you want to know how to proceed. Point out 3 possible signs or symptoms to which you should be aware and how you should act on their presence. What is the most serious problem associated with insect bites?

You have been diagnosed with hypothyroidism. Indicate 5 symptoms usually associated with this disease.

You have been felling shortness of breath. Investigate what may be behind this symptom, indicating 5 possible causes for shortness of breath.

TASK QUESTIONNAIRE OF USER EXPERIMENT 3

This is an English translated version of the questionnaire users answered after each task of User Experiment 3. Users answered this questionnaire using a web form.

User ID: _____

Information Situation ID: _____

Search Engine ID: ( ) 1 ( ) 2

Have you ever searched bout this topic before: ( ) Yes ( ) No

I had an exact idea of the type of information I wanted.

	1	2	3	4	5	
Disagree	( )	( )	( )	( )	( )	**Agree**

The task associated with this information situation was:

	1	2	3	4	5	
Unclear	( )	( )	( )	( )	( )	**Clear**
Easy	( )	( )	( )	( )	( )	**Complex**

The topic associated with this information situation was:

	1	2	3	4	5	
Extremely unfamiliar	( )	( )	( )	( )	( )	**Extremely familiar**

Evaluate your feeling of success with the iterations associated with this task.

		1	2	3	4	5	
First Iteration	**Extremely unsuccessful**	( )	( )	( )	( )	( )	**Extremely successful**
Second Iteration	**Extremely unsuccessful**	( )	( )	( )	( )	( )	**Extremely successful**
Third Iteration	**Extremely unsuccessful**	( )	( )	( )	( )	( )	**Extremely successful**

I believe I have succeeded in this task.

How useful were the queries suggested by the system?

	1	2	3	4	5	
Disagree	( )	( )	( )	( )	( )	**Agree**

	Not useful	Neutral	Useful
In the first iteration	( )	( )	( )
In the second iteration	( )	( )	( )

If you have used the query suggested by the system in the formulation of your second query, explain why did you found them useful.

If you have used the query suggested by the system in the formulation of your third query, explain why did you found them useful.

Please answer to what is asked in the information situation.

What difficulties did you felt during this task?

QUIZ TO EVALUATE ENGLISH PROFICIENCY IN USER EXPERIMENT 3

1. Who are _____ men at the table over there?
    a) this
    b) that
    c) those
    d) these
2. What's the difference _____ good and excellent?
    a) for
    b) between
    c) from
    d) among
3. Last Saturday I went _____ foot from Espinho to Porto.
    a) on
    b) with
    c) in
    d) by
4. Frank will be fine in the race. He's got _____ experience.
    a) lot of
    b) many
    c) much
    d) lots of
5. Anne's really enjoyed herself _____ her husband died.
    a) since
    b) when
    c) during
    d) for

6. A 1,000 euro fine for speeding is ridiculous! No, that _____ be right!

    a) mustn't do
    b) needn't
    c) can't
    d) wouldn't

7. I offered to do the washing up but Peter had _____ done it.

    a) yet
    b) even
    c) already
    d) still

8. I _____ better ring the airport to confirm the flight time.

    a) should
    b) would
    c) did
    d) had

9. She wishes she _____ more time to take a long holiday.

    a) will have
    b) would have
    c) has
    d) had

10. Barbara said she'll see you _____ Sunday night.

    a) in
    b) to
    c) at
    d) on

11. You have to complete the report _____ Monday.

    a) by
    b) till
    c) in
    d) until

12. The new law means you _____ drink alcohol in public parks.

    a) are not allowed to
    b) can't be
    c) doesn't have to

d) could not to

13. You look _____ a teacher.

a) like
b) as
c) as like
d) same as

14. We'd be on the beach now, if only the car _____ broken down.

a) haven't
b) hadn't
c) isn't
d) didn't

15. How old is Helen?

a) She's 80
b) She have 80
c) She has 80
d) She's 80 years

16. Jack's kids want _____ married again.

a) him to get
b) him get
c) that he get
d) that he gets

17. Although Lucy is _____ university teacher she doesn't earn much.

a) a
b) the
c) one
d) an

18. There's the man _____ took your mobile phone.

a) which
b) whose
c) whom
d) that

19. Do you _____ where the nearest grocery store is?

a) know
b) no

c) now

d) not

20. In May my mother went to Fatima _____ the Pope.

a) for see

b) to see

c) for to see

d) for seeing

21. _____ David manage to get tickets for the concert?

a) Did

b) Should

c) Have

d) Do

22. Read the text below and think of a word which best fits each space. Use only **one word** in each space.

If you want to work in advertising, _____ are three areas you can work in. The first is the Creative Department, which invents _____ the advertisements. Workers in _____ department are known as "Creatives" and they always work _____ pairs. A creative job, _____ outsiders, _____ might not should very stressful, the pressure to create original work is intense. "Creatives" have to keep up to _____ with the latest films, cartoons, videos, books and fashions to discover new techniques that could _____ used to sell a product.

The second area is the Accounts Department. This does _____ deal with financial accounts but with the companies that the agency produces advertisements for. Account Executives have to _____ sure that the "Creatives" fully understand _____ the client requires. Account Executives need to keep up _____ the Creative team _____ the client happy. It's a job that requires a lot of diplomacy, as _____ as a very good memory and excellent organizational skills.

The third is the media, which involves placing advertisements in magazines, _____ radio or TV, or in public areas. The Media Department carries _____ research into people's habits, to find _____, for example, _____ radio stations long-distance lorry drivers prefer. Then it advises clients _____ which medium would be _____ appropriate for its advertisement.

TRANSLATED VERSION OF METER

The following list contains some real medical words. For example, some of the words have to do with body parts or functions, kinds of diseases, or things that can make your health better or worse. The list also contains some items that may look or sound like medical words but that are not actually real words.

As you read through the list, put an "X" next to the items that you know are real words. You should not guess. Only put an "X" next to an item if you're sure it's a real word.

Table K.1: Translated version of METER. In parentheses, the original English concept.

_____	Irritice (Irrity)	_____	Diagnóstico (Diagnosis)
_____	Artrite (Arthritis)	_____	Depreção (Depretion)
_____	Obesidade (Obesity)	_____	Icterícia (Jaundice)
_____	Gripe (Flu)	_____	Vesícula (Gallbladder)
_____	Comportamentose (Behaviose)	_____	Aborto (Miscarriage)
_____	Sífilis (Syphilis)	_____	Apêndice (Appendix)
_____	Potássio (Potassium)	_____	Vasular (Fam)
_____	Hormonas (Hormones)	_____	Infarte (Infarth)
_____	Nervos (Nerves)	_____	Dose (Dose)
_____	Meite (Pilk)	_____	Hemorróides (Hemorrhoids)
_____	Reção (Rection)	_____	Testículo (Testicle)
_____	pirrose (Blout)	_____	Olho (Eye)
_____	Hemofícia (Boweling)	_____	Diafagma (Midlocation)
_____	Exercício (Exercise)	_____	Insoniar (Insomniate)
_____	Pústula (Pustule)	_____	Conceptivo (Bloodgatten)
_____	Inlesto (Inlest)	_____	Hepatite (Hepatitis)
_____	Pólema (Pollent)	_____	Astringente (Astiringe)
_____	Malorias (Malories)	_____	Nutro (Nutral)
_____	Cancro (Cancer)	_____	Asma (Asthma)
_____	Alcooliose (Alcoheliose)	_____	Inflamatório (Inflammatory)
_____	Antibióticos (Antibiotics)	_____	Anemia (Anemia)
_____	Antirepressivo (Antiregressant)	_____	Alargénico (Allagren)
_____	Colite (Colitis)	_____	Gravitez (Prognincy)
_____	Diabetes (Diabetes)	_____	Stress (Stress)
_____	Occitital (Occipitent)	_____	Elérgico (Ellargic)
_____	Náuseo (Nausion)	_____	Sexualmente (Sexually)
_____	Impetigo (Impetigo)	_____	Pélvice (Pelvince)
_____	Menstrual (Menstrual)	_____	Pacritite (Vaccilly)
_____	Basso (Abghorral)	_____	Prescrição (Prescription)
_____	Ataque (Seizure)	_____	Germes (Germs)
_____	Cerpes (Cerpes)	_____	Gonorréia (Gonorrhea)
_____	Rim (Kidney)	_____	Abdómico (Tumic)
_____	Emergência (Emergency)	_____	Fadiga (Fatigue)
_____	Pociente (Potient)	_____	Osteoporose (Osteoporosis)
_____	Menopausa (Menopause)	_____	Obstipação (Constipation)