Evaluation and comparison of automatic methods to identify health queries

Carla Teixeira Lopes

Faculdade de Engenharia da Universidade do Porto, Portugal
carla.lopes@fe.up.pt

Abstract. The use of the Web to find health information is a common practice nowadays. The improvement of Health Information Retrieval depends on studies that, frequently, require the identification of health-related queries. Being usually done by human assessors, this identification may turn out to be inefficient and even impracticable in some cases. To overcome this problem we propose, analyze and compare automatic methods to identify health-related queries. One type of methods uses health vocabularies and the other analyses the co-ocurrence of query terms with the word "health" in web page results. Our goal is to compare the two different strategies of automatic classification, to compare several variants in each strategy and to verify if its performance is enough to be executed without human intervention. The evaluation was done comparing the automatic classification with the classification made by a team of ten human assessors, in a pool of 20,000 queries. The use of Yahoo! to calculate the co-occurrence rate at a threshold value of 0,5 was the method with best trade-off between sensibility (73%) and specificity (79%).

Key words: Health Information Retrieval, Web Information Retrieval, Health Queries, Automatic Classifiers, Medical Vocabularies.

1 Introduction

The use of the Web to find health information has become a common practice nowadays. According to a Pew Internet & American Life Project 2006 report [5], eight in ten american users go online for health information and the typical health information session starts at a search engine. 74% of all health seekers also said that health search allowed them to make more appropriate health decisions. Jupiter Research [7] reached similar conclusions, founding that 71% of online consumers use search engines to find health-related information.

The large proliferation of health information Web search and the impact it may have on people's life accent the importance of studies in Health Information Retrieval. Usually, in these studies, one of the first steps is the identification of health-related queries in a pool of queries. A health query is a query that intends to retrieve health-related information and is related to a health information need. The most frequent classification method (as happens in [11]) involves human intervention making it a slow process and requiring 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 of health queries identification could be a useful tool.

Eysenbach and Kohler [3] 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 (a good state of the art of this area is done in the paper of Beitzel and Lewis [1]). 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 [3]. Based on the knowledge that most health queries contain terms that can be mapped to health/medical vocabularies [8, 9], we have decided to use this type of vocabularies to detect the presence of health terms in queries through several different strategies.

The vocabulary chosen is the Consumer Health Vocabulary (CHV) developed as an open source and collaborative initiative to complement the Unified Medical Language System (UMLS). This vocabulary links everyday phrases about health and technical terms used by professionals, aiming to bridge the communication gap between consumers and professionals. It is available for download on the CHV website [12] as several files. We opted for CHV instead of UMLS because the first focuses on concepts employed by consumers in health communications. As we want to analyze queries submitted to generic search engines, it's probable that most queries are submitted by non-health experts.

In short, we want to evaluate the performance of the several methods, to compare the method proposed by Eysenbach and Kohler [3] with methods that use health vocabularies and to compare the different variants of each type of methods.

The next Section of this paper describes the 14 automatic methods we propose and want to compare. This section also describes the processes of implementation and evaluation of the described methods. In Section 3 are presented the results gathered after the execution of the several methods. In Section 4 are discussed the previously presented results. Finally, conclusions are presented together with lines of future work in Section 5.

2 Methods

2.1 Automatic methods to detect health-related queries

We propose 14 automatic methods to detect health-related queries that can be grouped in two distinct categories. A first category (CHV methods), with 11 different methods, uses the CHV. The second category (co-ocurrence methods) contains 3 methods based on the idea that health-related terms should co-occur with the word "health" more often than non-health terms, as proposed by Eysenbach and Kohler [3].

While CHV methods produce a discrete class label indicating only the predicted class (health or non-health) of the query, co-occurence methods produce a continuous output to which different thresholds may be applied to predict a query's class.

CHV methods The CHV can be downloaded in 4 different flat files: concepts terms, ngrams, stop concepts and incorrect mappings. The first file contains concepts and associated terms. Each concept may have many terms and each term is listed in a separate row. The ngrams file lists terms not mapped to the UMLS but associated to medical concepts. The stop concepts file lists concepts excluded from the CHV. The last file lists incorrect combinations of concepts and terms.

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 necessity of several methods emerged from the large size of the initial concepts terms flat file (158,908 terms) and also from its contents (against initials expectations, it included several terms not specifically health-related — p.e.: rail, driver — and even stop-words). The first step involved the removal of stop-words and the replacement of characters that could be misunderstood in regular expressions (used later to parse the files). Then, 11 variants with different lists of terms (all after stop-words removal) were defined: 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 (UMLS preferred terms — with the field `UMLS_preferred_name` set to "yes"), CHV8 (CHV preferred terms — with the field `CHV_preferred_name` set to "yes"), CHV9 (UMLS or CHV preferred terms), CHV10 (6,000 more frequent terms obtained directly from the website — 5,898 terms after stop-words removal) and CHV11 (10,000 more frequent terms obtained directly from the website — 9,872 terms after stop-words removal).

The criteria behind these 11 variants 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).

Co-ocurrence methods As mentioned previously, 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 co-occurence 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:

$$cooc(Q) = \frac{\#results(terms_Q \cap health)}{\#results(terms_Q)}$$

where $terms_Q$ is the set of terms that compose the query Q. If $\#results(terms_Q) = 0$, $cooc(Q) = 0$.

This proportion is an indicator of 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. For example, the query 'diabetes symptoms' has a co-occurence rate of $\frac{478000}{929000} = 0,51$ and the query 'Pavarotti' has a co-occurence rate of $\frac{359000}{6440000} = 0,06$.

In the work of Eysenbach and Kohler [3], where this method was proposed, Google was the used search engine. Here, we have used Google and Yahoo! to determine the number of results and 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 co-ocurrence rates:

$$G_{cooc_Q} = \frac{\#google(terms_Q \cap health)}{\#google(terms_Q)} \; Y_{cooc_Q} = \frac{\#yahoo!(terms_Q \cap health)}{\#yahoo!(terms_Q)}$$

$$Y + G_{cooc_Q} = \frac{\#google(terms_Q \cap health) + \#yahoo!(terms_Q \cap health)}{\#google(terms_Q) + \#yahoo!(terms_Q)}$$

The differences detected in the number of results of both search engines (also stated in [2]) took us to combine the number of results returned by the two search engines in the third method.

After the calculation of the co-occurence 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 co-occurence rate was larger than or equal to the threshold, the query was considered to be a health-related query at that threshold.

2.2 Implementation

To evaluate the methods described previously we've used a collection of 20,000 web queries, randomly sampled from AOL Search in the Fall of 2004. This collection was used by Beitzel and Lewis in a research project [1] 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 CHV methods two other datasets were also used: one text file with a list of stop-words provided by the University of Glasgow [10] and a tab separated value (tsv) file with the CHV Concepts & Terms Flat File available at CHV website [12]. The first dataset file has one stop-word per line and the second dataset file has one line per each term and associated information.

Several Perl scripts were developed to implement the methods described previously. In each CHV method we've used two Perl scripts: one (`generateTerms-List.pl`) that generates a subset of health terms and another one (similar in all CHV methods) that classifies queries (see Figure 1). The `generateTermsList.pl` also removes stop-words and replaces special characters that may be misunderstood by regular expressions. The `classifyQueries.pl` simply checks if any of each query's terms is present in the terms list. If present, queries are classified as health-related.

In the co-occurence methods, we've developed scripts (one for each search engine) to automatically get the number of results returned for each query in Google and Yahoo (see Figure 2) through each search engine's API. Each of these scripts was then used by another script (`classifyQueries.pl`) that reads the queries collection file line by line, asks the `numberofResults.pl` for the number of results of two queries (the query read and the query plus the word "health") and writes this information in another file.

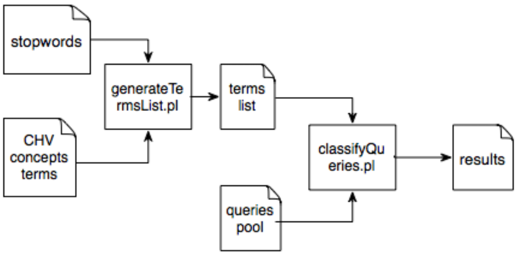


Fig. 1. CHV methods global architecture — dataset files and Perl scripts

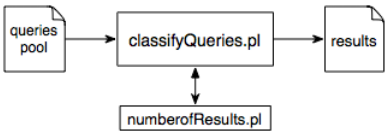


Fig. 2. Co-occurence methods global architecture — dataset files and Perl scripts

2.3 Evaluation

The evaluation of each method was done through the comparison of the classification made by the team of human assessors and the classification of each method. In the CHV methods the classification is immediately delivered after the execution of the described scripts. In the co-occurence methods, the classification only occurs after the calculation of the cooc rate and its comparison with each threshold. The best thresholds are determined after the analysis of all collected data.

3 Results

For each method, measures like sensitivity, specificity and accuracy were calculated. These can be expressed in terms of probabilities of the following events: HC_H (query is classified as health-related in a human classification), HC_NH (query is classified as non-health-related in a human classification), AC_H (query is classified as health-related in an automatic classification) and AC_NH (query is classified as non-health-related in an automatic classification).

Sensitivity (SEN) is 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: $P(AC\_H|HC\_H)$.

Specificity (SPC) is 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: $P(AC\_NH|HC\_NH)$.

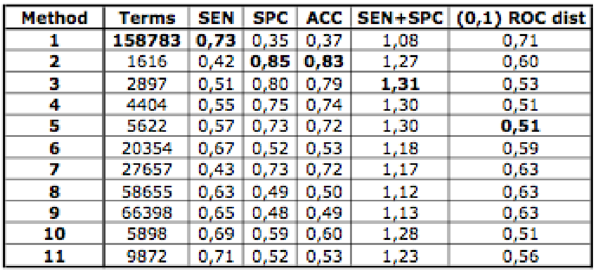
Accuracy (ACC) is the tax of correct classifications (either as health-related or as non-health-related) and is expressed by: $\frac{P(AC\_H \cap HC\_H)+P(AC\_NH \cap HC\_NH)}{P(HC\_H)+P(HC\_NH)}$

Besides the calculation 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 sensibility 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 [4].

3.1 CHV methods

Table of Figure 3 presents, for each CHV method, the number of terms used in the classification method (`Terms`), sensibility (`SEN`), specificity (`SPC`), accuracy (`ACC`), sum of sensibility and specificity (`SEN + SPC`) and the distance of each method to the optimal point in ROC space (`(0,1) ROC dist`). Each column's greatest value is highlighted in bold (except 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 sensibility 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 sensibility and specificity.

Fig. 3. Number of terms, Sensibility, Specificity, Accuracy and other Measures for CHV methods



Method	Terms	SEN	SPC	ACC	SEN+SPC	(0,1) 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

To aid the comparison of the several methods a ROC graph was drawn (Figure 4) with each method represented by a different point in the ROC space.

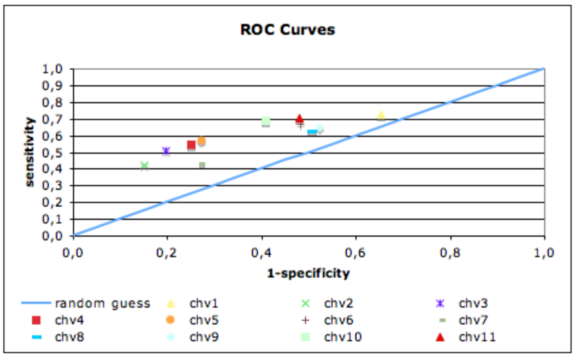


Fig. 4. CHV methods ROC graph

3.2 Co-occurence methods

As mentioned in Section 2, co-occurence methods are continuous classifiers because they produce a continuous output (co-occurence rate) that may be considered an estimate of queries health-relatedness probability. Each method has its own co-occurence rate with the distribution presented in the histograms of the Figures 5, 6 and 7. In these histograms, only co-occurence rates between 0 and 1 are represented. In the three methods were detected queries with co-occurence rates greater than 1: Google has 3,174, Yahoo! has 693 and GoogleYahoo! has 1,417 queries. Google has a co-occurence average of 0.45, Yahoo! of 0.32 and GoogleYahoo! of 0.39. The standard deviation is also greater in Google (0.305), followed by Google Yahoo! (0.243) and Yahoo! (0.228).

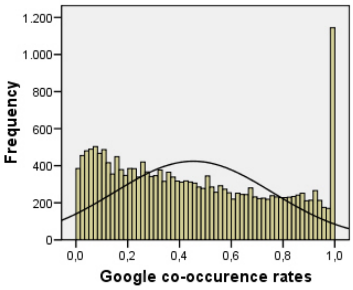


Fig. 5. Google co-occurence rate histogram

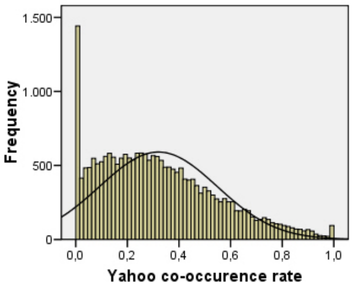


Fig. 6. Yahoo! co-occurence rate histogram

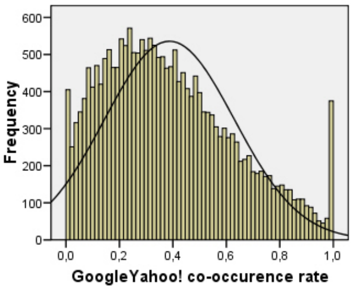
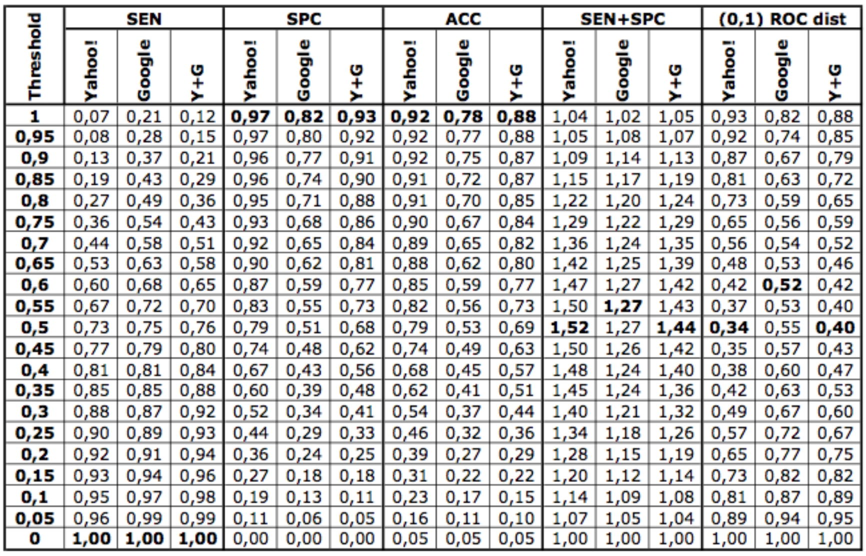


Fig. 7. GoogleYahoo! co-occurence rate histogram

To predict each query health-relatedness, this continuous output was then compared with different thresholds (ranging from 0 to 1). Sensibility, specificity, accuracy, sum of sensibility and specificity and the distance of each method to the optimal point in ROC space, for the several thresholds in each method, are presented in Table of Figure 8. Each column's greatest value is highlighted in bold (except the last column where the minimum value is the indicator of a best performance). Just as in the CHV methods, the sum of sensibility and specificity does not intend to be a single evaluation measure of the optimal threshold.

Fig. 8. Sensibility, Specificity, Accuracy and other Measures for Co-occurence methods



Threshold	**SEN**			**SPC**			**ACC**			**SEN+SPC**			**(0,1) ROC dist**		
	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

The ROC curves for each co-occurence method are represented in Figure 9. Each point in the curve corresponds to a threshold value, starting on 1 at the left side of the graph.

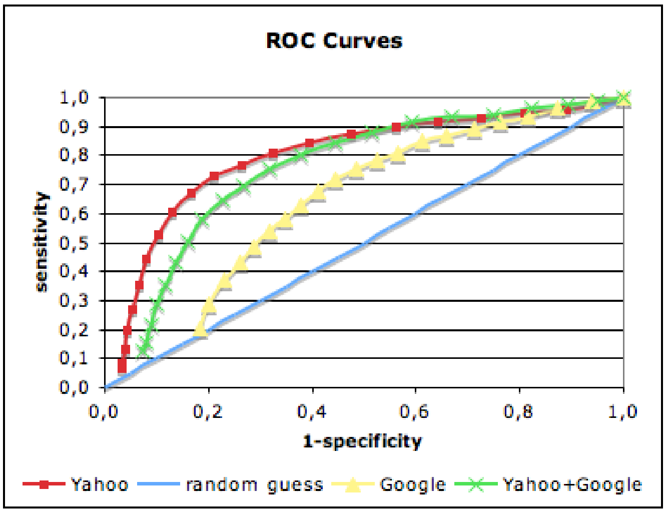


Fig. 9. Co-occurence methods ROC graph

4 Discussion

In Figure 4 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 performances 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 4, 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 sensibility 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 sensibility (51%) and specificity (80%). CHV5 is the closest to the point optimal point in ROC Space (minimum distance to (0,1)).

We can also see that the relation between the number of health terms and sensibility is not directly proportional. For example, CHV10 has less terms but higher sensibility 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 sensibility.

To begin the analysis of co-occurence methods we would like to mention the existence of co-occurence rates greater than 1. Theoretically, these values shouldn't exist because the default operator between terms in both search engines (Google and Yahoo!) is the logic "AND", what means that all terms in a query without operators should appear in results' pages. 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 larger 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 a estimate. Google Help Center [6] explains that not providing the exact count allows them to return search results faster. Yet, the high number of these cases is still surprising.

The histograms of Figures 5, 6 and 7 show that the GoogleYahoo! co-ocurrence rate is the closest to the Normal distribution, followed by the Yahoo! co-ocurrence 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, in a large number of queries' return pages, 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 of Figure 8 it's possible to verify that, as expected, sensibility is 1 at a threshold of 0 (co-occurence 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 sensibility and specificity measure has the best value at a threshold of 0.5 of the Yahoo! method (with 73% of sensibility and 79% of specificity) just as the Yahoo!Google method. The Google method has its best sensibility+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 9 it's 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 9 and Table of Figure 8, the Yahoo!Google method has an intermediate performance, being probably better than Google due to Yahoo! performance.

To test if the differences between Yahoo! (at 0.5) and Google (at 0.55 and 0.6) are significant two McNemar tests were applied: one between Yahoo! and Google (0.55) and other between Yahoo! and Google (0.6). P-value was 0 in both tests what means the differences in proportions between the best of Yahoo! methods and the two better Google methods are significant. This result encourages the use of Yahoo! to the co-occurence methods.

Google results in this sample of 20,000 queries are different from the results of Eysenbach and Kohler [3]. 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 worser 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.

We would like to emphasize that the methods indicated as optimal may be discarded when compared to others if sensibility is preferable to accuracy or vice-versa. For example, in a situation where we want to reduce 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 sensibility instead of specificity.

5 Conclusions and Future Work

We evaluated several variants of two type of classifiers: a discrete one, proposed by the author, that uses terms of health vocabularies and a continuous one, proposed by Eysenbach and Kohler [3], that evaluates the query relatedness to health through the co-ocurrence rate of query terms with the word "health" in search engines' results.

While Yahoo! demonstrated a better performance than Google in the co-occurence methods, its results were still worser than Eysenbach and Kohler's results. In their work, at a threshold of 35%, sensibility was 85.2% and specificity was 80.4%, while in our Yahoo! method, at a threshold of 0.5, sensibility was 73% and specificity was 79%. We think our results depict reality more accurately since our sample of queries is much larger (20,000 against 2,985 queries).

None of the 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).

A manual definition of a term list might improve CHV methods. Through the behavior's analysis of the best CHV methods by a human assessor it may be possible to eliminate some of the terms that produce false positives and add some terms that could reduce the number of false negatives. We also aim to define and evaluate this type of methods using the UMLS vocabulary instead of the CHV. Another line of future work in this type of methods involves the definition of a continuous output based on the number of health terms presented in the query (the methods presented in this paper only detect the presence or non-presence of health terms).

We also intend to evaluate co-occurence methods in Portuguese queries, analyzing the co-occurence rate with the "health" Portuguese word. If results in Portuguese are similar to the English results, this method has the advantage of an easier application to other languages (while the vocabularies methods require the definition of foreign languages' lists of terms). It could also be interesting to analyze the co-occurence rate with terms different from “health” or even a set of terms separated by the OR logical operator.

A specific evaluation of each query health-relatedness by a health specialist would also increase the correctness of the several methods’ performance evaluation. In fact, some human classifications of health queries used in the dataset are dubious (e.g.: “devils club” and “regedit”).

The application of these methods on other datasets would also allow to prove the validity of these results.

References

1. S. Beitzel, E. Jensen, D. Lewis, A. Chowdhury, A. Kolcz, and O. Frieder. Improving Automatic Query Classification via Semi-supervised Learning. In *The Fifth IEEE International Conference on Data Mining*, New Orleans, Louisiana, U.S.A., November 2005.
2. Ionut Alex Chitu. Google Finds Less Search Results. Available from: `http://googlesystem.blogspot.com/2007/12/google-finds-less-search-results.html` [accessed 28 December, 2007].
3. G. Eysenbach and Ch. Kohler. 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 2003 Symposium Proceedings*, pages 225–230, 2003.
4. Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, (27):861–874, 2006.
5. Susannah Fox. Online Health Search 2006. Technical report, Pew Internet & American Life Project, 2006.
6. Google. How does Google calculate the number of results? Available from: `http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=70920` [accessed 31 December 2007].
7. JupiterResearch. JupiterResearch Finds Strong Consumer Demand and Market Opportunity for Health Search Engines. Available from: `http://www.jupiterresearch.com/bin/item.pl/press:press_release/2006/id=06.07.17-health_search.html` [accessed 2th January 2008].
8. Alexa T. McCray, Russell F. Loane, Allen C. Browne, and Anantha K. Bangalore. Terminology Issues in User Access to Web-based Medical Information. In *AMIA 1998*, 1998.
9. Zeng QT, Crowell J, Plovnick RM, Kim E, Ngo L, and Dibble E. Assisting Consumer Health Information Retrieval with Query Recommendations. *J Am Med Inform Assoc*, 13(1):80–90, Jan-Feb 2006.
10. Mark Sanderson. Stop words list. Available from: `http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words` [accessed 30 December 2007].
11. Amanda Spink, Yin Yang, Jim Jansenn, Pirrko Nykanen, Daniel P. Lorence, Seda Ozmutlu, and H. Cenk Ozmutlu. A study of medical and health queries to web search engines. *Health Information and Libraries Journal*, (21):44–51, 2004.
12. Qing T. Zeng. Consumer Health Vocabulary Initiative. Available from: `http://www.consumerhealthvocab.org/` [accessed 27th December 2007].