Use of Temporal Expressions in Web Search

Sérgio Nunes[1], Cristina Ribeiro[1,2], and Gabriel David[1,2]

[1] Faculdade de Engenharia da Universidade do Porto
[2] INESC-Porto
Rua Dr. Roberto Frias, s/n 4200-465 Porto Portugal
{sergio.nunes,mcr,gtd}@fe.up.pt

Abstract. While trying to understand and characterize users' behavior online, the temporal dimension has received little attention by the research community. This exploratory study uses two collections of web search queries to investigate the use of temporal information needs. Using state-of-the-art information extraction techniques we identify temporal expressions in these queries. We find that temporal expressions are rarely used (1.5% of queries) and, when used, they are related to current and past events. Also, there are specific topics where the use of temporal expressions is more visible.

1 Introduction

Query log analysis is currently an active topic in information retrieval. There is a significant and growing number of contributions to the understanding of online user behavior. However, work in this field has been somewhat limited due to the lack of real user data and the existence of important ethical issues [2]. The recent availability of large datasets has specially contributed to a growing interest in this topic. On the other hand, temporal information extraction has reached a point of significant maturity. Current algorithms and software tools are able to extract temporal expressions from free text with a high degree of accuracy. This paper contributes to the characterization of the use of temporal expressions in web search queries, by combining work from these two areas. Our main goal is to provide a better understanding of how users formulate their information needs using standard web search systems. Our focus is on a particular facet of this behavior, namely the use of temporal expressions.

In the following section is presented the experimental setup, including details about the datasets and software used. Section 3 describes the experiments and highlights the main results. An overview of related work is included in Section 4, followed by the conclusions in Section 5

2 Experimental Setup

We used two publicly available datasets containing web search queries. The first dataset includes a collection of manually classified web search queries collected

C. Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 580–584, 2008.

from the AOL search engine [4]. Each one of the 23,781 queries has been manually classified using a set of predefined topics by a team of human editors. The classification breaks down as follows: Autos (2.9%), Business (5.1%), Computing (4.5%), Entertainment (10.6%), Games (2.0%), Health (5.0%), Holidays (1.4%), Home & Garden (3.2%), News & Society (4.9%), Organizations (3.7%), Personal Finances (1.4%), Places (5.2%), Porn (6.0%), Research (5.7%), Shopping (8.6%), Sports (2.8%), Travel (2.6%), URL (5.7%), *Misspellings* (5.5%) and *Other* (13.2%).

The second dataset is also from AOL [8] and includes more than 30 million (non-unique) web queries collected from more than 650,000 users over a three month period. This dataset is sorted by user ID and sequentially ordered. For each request there is also information about the time at which the query was issued and, when users follow a link, the rank and the URL of the link. An important feature of this second dataset is the availability of the query issuing time, making possible the positioning of temporal expressions. For instance, we are able to determine the specific date of a search for *"a week ago"* because we have access to this information. However, and unlike the first dataset, this one isn't classified.

Temporal expressions were extracted from each query using free, publicly available, Natural Language Processing (NLP) software. First, text was tagged using Aaron Coburn's `Lingua::EN::Tagger` [1], a Part of Speech tagger for English. Then, the output was redirected to TempEx [7], a text tagger that is able to identify a large number of temporal expressions. This tagger covers most of the types of time expressions contained in the 2001 TIMEX2 standard [5]. In Table 1 several examples of this process are presented, showing that TempEx is able to detect a wide range of temporal expressions (e.g. explicit dates, implicit dates, periods).

Table 1. Examples of Tagged Search Queries

`olympics 2004`
`⇒ olympics <TIMEX2 TYPE="DATE" VAL="2004">2004</TIMEX2>`
`easter 2005`
`⇒ <TIMEX2 TYPE="DATE" ALT_VAL="20050327">easter 2005</TIMEX2>`
`monday night football`
`⇒ <TIMEX2 TYPE="DATE">monday night</TIMEX2> football`
`us weekly`
`⇒ us <TIMEX2 TYPE="DATE" SET="YES" PERIODICITY="F1W">weekly</TIMEX2>`

3 Use of Temporal Expressions

First, we investigate how temporal expressions are distributed within distilled web queries. For this task we used the first dataset since it is manually annotated with classes. The topics containing the higher percentage of queries with

[1] `http://search.cpan.org/~acoburn/Lingua-EN-Tagger`

temporal expressions are: Autos (7.8%), Sports (5.2%), News & Society (3.9%) and Holidays (2.5%). Examples of queries containing temporal expressions are: *"1985 ford ranger engine head"* (Autos), *"chicago national slam 2003"* (Sports) and *"los angeles times newspaper april 1946"* (News & Society). Manual inspection reveals that the higher number of temporal expressions in the Autos class is mostly due to searches for vintage cars.

In a second experiment, we analyzed the overall distribution of temporal expressions in web search queries. Our first finding is that the use of these expressions is relatively rare. In the first AOL dataset the total number of queries including temporal expression was 347 (1.5%). Remarkably, on the second dataset, we found 532,989 temporal expressions resulting in an equal percentage of 1.5%. Removing duplicate queries results in a small increase in these percentages, specifically 1.6% for the first dataset and 1.9% for the second.

To evaluate the quality of the TempEx tagger when applied to web search queries, we manually classified a random subset of 1,000 queries from the large AOL corpus (including duplicates). We compared this classification with an automatic classification performed by the TempEx tagger. Standard IR measures were computed: accuracy (0.99), precision (0.92) and recall (0.63). The low recall value indicates that the tagger is being conservative, missing some temporal expressions. The non-parametric McNemar test was performed to evaluate the homogeneity of the two classifications. The test confirms that the automatic classification is equivalent to the human classification ($p > 0.05$).

Restricting our analysis to the second AOL dataset, we performed additional measurements. Since query issuing time is available, we are able to precisely date a large fraction of the temporal expressions found. Using this information we measured the number of expressions referencing past events, present events and future events. TempEx automatically detects some of these expressions. This module was able to identify generic references to the past (e.g. *"once"*, *"the past"*) (1.25% of the temporal expressions), to the present (e.g. *"now"*, *"current"*) (4%) and to the future (e.g. *"future"*) (0.82%). The vast majority of temporal expressions (94%) do not include explicit references like these.

We've extracted the year from all dated expressions and counted the occurrences. Taking into account that all queries were issued between March and May 2006, we see that the majority of temporal expressions are related to current events. The frequency distribution is positively skewed with a long tail toward past years. Summarizing, in all temporal expressions identified, 42.5% indicate a date from 2006, 49.9% are from dates prior to 2006, and 4.2% are from dates after 2006.

To better understand which temporal expressions were being used, we manually inspected a list of the top 100 more common expressions. These expressions account for slightly more than 80% of the queries containing temporal expressions. We grouped all references to a single year and to a single month in two generic expressions (i.e. <*Year*> and <*Month*>). The top 10 expressions used in queries are: <*Year*> (45.7%), *easter* (5.6%), *daily* (5.4%), <*Month*> (4.6%), *now* (2.3%), *today* (2.1%), *mothers day* (1.8%), *current* (1.2%), *christmas* (1.1%) and

weekly (0.8%). It is important to note that seasonal results (e.g. "Easter") are artificially inflated in the 3 month of data (March to May).

When starting this research one of our hypothesis was that temporal expressions were regularly used to improve initial queries. To investigate on this hypothesis we did a rough analysis on query refinements within the AOL dataset. Our algorithm is very simple and only identifies trivial query reformulations. In a nutshell, since the dataset is ordered by user and issuing time, we simply compare each query with the previous one to see if there is an expansion of the terms used. For instance, *"easter holidays"* is considered a reformulation of *"easter"*. With this algorithm we found 1,512,468 reformulated queries (4.2%). We then counted the presence of temporal expressions in this subset and verified that only 1.4% of these queries contained temporal expressions.

4 Related Work

We found no previous work on the specific topic of identifying and characterizing the use of temporal expressions in web search queries. Thus, the related work presented here is divided in the two parent topics: *temporal expression extraction* and *query log analysis*. In recent years, Temporal Information Extraction emerged from the broader field of Information Extraction [9]. Most work in this field is focused on the study of temporal expressions within semi-structured documents [6]. In our work we apply these techniques in processing short text segments that represent information needs.

Query log analysis has been the focus of increasing interest in recent years. Most work in this area has been devoted to the classification and characterization of queries. An example of a detailed work in this area is from Beitzel et al. [3]. In this work the authors perform a detailed characterization of web search queries through time using a large topically classified dataset. Our work differs from this since we are interested in how temporal expressions are used within queries.

5 Conclusions

Contrary to our initial expectations, the use of temporal expressions in web queries is relatively scarce. Using two different datasets, we've found that temporal expressions are used in approximately 1.5% of the queries. We speculate on three reasons that might explain this situation: (1) information needs of web users are mostly focused on current events; (2) users are generally happy with the results obtained using short text queries; (3) users resort to more advanced interfaces when they have dated information needs. Investigating these hypotheses is left for future work. Focusing on the small subset of temporal expressions extracted, we've found that most temporal expressions reference current dates (within the same year) and past dates (exhibiting a long tailed behavior). Future dates are rarely used. Finally, we've shown that these expressions are more frequently used in topics such as: Autos, Sports, News and Holidays.

Although temporal expressions appear in only a small fraction of all queries, the scale of the Web translates this percentage into a large number of users. Temporal expression extraction might be used in public search engines to improve ranking or result clustering. As the web grows older, and more content is accumulated in archives (e.g. Internet Archive), we think that the need for dated information will rise [1]. Search engine designers can respond to this challenge by incorporating temporal information extraction algorithms or by developing specialized search interfaces. As an example, Google has recently launched a prototype that provides date-based navigation in search results using timelines [2].

Acknowledgments. Sérgio Nunes was financially supported by the Fundação para a Ciência e a Tecnologia (FCT) and Fundo Social Europeu (FSE - III Quadro Comunitário de Apoio), under grant SFRH/BD/31043/2006. We would like to thank the ECIR 2008 reviewers, whose comments have contributed to improve the final version of this paper.

References

1. Alonso, O., Gertz, M., Baeza-Yates, R.: On the value of temporal information in information retrieval. ACM SIGIR Forum 41(2), 35–41 (2007)
2. Bar-Ilan, J.: Access to query logs - an academic researcher's point of view. In: Query Log Analysis: Social And Technological Challenges Workshop. 16th International World Wide Web Conference (WWW 2007) (May 2007)
3. Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., Frieder, O.: Hourly analysis of a very large topically categorized web query log. In: SIGIR 2004: Proceedings of the 27th annual international conference on Research and development in information retrieval, pp. 321–328. ACM Press, New York (2004)
4. Beitzel, S.M., Jensen, E.C., Frieder, O., Lewis, D.D., Chowdhury, A., Kolcz, A.: Improving automatic query classification via semi-supervised learning. In: ICDM 2005: Proceedings of the Fifth IEEE International Conference on Data Mining, Houston, Texas, USA, November 2005, pp. 42–49 (2005)
5. Ferro, L., Mani, I., Sundheim, B., Wilson, G.: TIDES temporal annotation guidelines (v. 1.0.2). Technical report, The MITRE Corporation, Virginia (June 2001)
6. Mani, I., Pustejovsky, J., Sundheim, B.: Introduction to the special issue on temporal information processing. ACM Transactions on Asian Language Information Processing (TALIP) 3(1), 1–10 (2004)
7. Mani, I., Wilson, G.: Robust temporal processing of news. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000), Hong Kong, pp. 69–76 (2000)
8. Pass, G., Chowdhury, A., Torgeson, C.: A picture of search. In: InfoScale 2006: Proceedings of the 1st international conference on Scalable information systems, ACM Press, New York (2006)
9. Wong, K.-F., Xia, Y., Li, W., Yuan, C.: An overview of temporal information extraction. International Journal of Computer Processing of Oriental Languages 18(2), 137–152 (2005)

[2] `http://www.google.com/experimental`