DISAMBIGUATING WEB SEARCH RESULTS BY TOPIC AND TEMPORAL CLUSTERING: A PROPOSAL

Ricardo Campos
Polytechnic Institute of Tomar, Tomar, Portugal
Centre for Human Language Technology and Bioinformatics, University of Beira Interior, Covilhã, Portugal
ricardo.campos@ipt.pt, ricardo@hultig.di.ubi.pt

Gaël Dias
Centre for Human Language Technology and Bioinformatics, University of Beira Interior, Covilhã, Portugal
ddg@hultig.di.ubi.pt

Alípio Mário Jorge
LIAAD – INESC Porto L.A., Faculty of Sciences, University of Oporto, Portugal
amjorge@fc.up.pt

Keywords: Temporal Information Retrieval, Time-Based Clustering, Topic Clustering, Web Content Mining.

Abstract: With so much information available on the web, looking for relevant documents on the Internet has become a difficult task. Temporal features play an important role with the introduction of a time dimension and the possibility to restrict a search by time, recreating a particular moment of a web page set. Despite its importance, temporal information is still under-considered by current search engines, limiting themselves to the capture of the most recent snapshot of the information. In this paper, we describe the architecture of a temporal search engine which uses timelines to browse search results. More specifically, we intend to add a time measure to cluster web page results, by analyzing web page contents, supporting the search of temporal and non-temporal information embedded in web documents.

1 INTRODUCTION

Current search engines return lists of ranked URLs with their titles and their snippets. In this process, the user is required to go through the extensive list of the retrieved results, seeking for the result that best meets his needs, which is not necessarily the most recent one. Whilst traditional search engines continue to present data in a linear fashion, some commercial approaches like iBoogie, Clusty and Grokker have begun to present results in a partitioning hierarchy (Campos et al, 2008). This evolution presents an alternative mechanism to display similar documents in one page without forcing the user to go through hundreds of items (Alonso and Gertz, 2006). Although the retrieved search engines results are now more accurate, quite often they do not meet the demands of the user, especially when the query is ambiguous. Some search engines try to overcome this problem by providing some post-search features to the user. In this work we propose to do it automatically, "on the fly", without the need for user intervention, although the user can do it later by using query refinement and personalization of the results. To do so, we introduce timelines which, together with a clustering topic approach, provide a dual solution to term ambiguity. First, query terms with possible different meanings are disambiguated through the main topics they convey. Second, query terms are analysed through their display over a timeline, thus performing disambiguation through the temporal information. On Table 1 we list the combination of the 4 possible cases on term ambiguity, whether a term is ambiguous in terms of concept or over time. It seems obvious that the case that will most benefit from the introduction of a clustering and timeline approach, is the match between time ambiguity and concept ambiguity, there are however 3 other cases that will also greatly benefit from this proposal.

Table 1. Term Ambiguity.

	Concept Ambiguous	Concept Unambiguous
Time Ambiguous	Figure 4	Figure 3
Time Unambiguous	Figure 2	Figure 1

(1) Case one: query is not ambiguous nor in concept or time (see Figure 1).

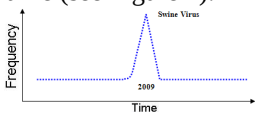

Figure 1: Swine Virus Query.

This example deals with a popular query occurring at some specific point. Despite not being an ambiguous query, nor in concept or time, its display on a timeline may improve, in an historical perspective, the understanding of the phenomenon, as the user could quickly infer when it did happen.

(2) Case two: query is ambiguous in concept, not in time (see Figure 2).

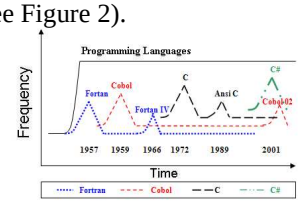

Figure 2: Programming Languages Query.

We deal with Programming Languages like they have always existed, but it has not always been like that. Its appearance is around the decade of 50 - 60 and its frequency, constant since then. Several instances have however occurred over all these years. First we had Fortran, nowadays we have C#, distinct groups related with a different time period, that will benefit from its display on a temporal clustering system in order to solve term ambiguity.

(3) Case three: query is not ambiguous in concept, only in time (see Figure 3).

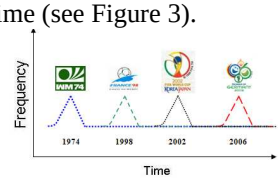

Figure 3: Football World Cup Query.

When submitted in current search engines this kind of queries will mainly retrieve references to the most recent events as opposed to the old ones. Not being ambiguous in terms of concepts, temporal clusters are enough to present the results.

(4) Case four: query is ambiguous in terms of concept and in time (see Figure 4).

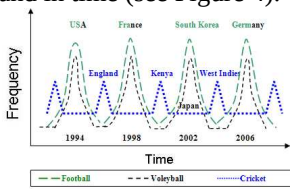

Figure 4: World Cup Query.

An attempt to make an historical perspective of some ambiguous subject, in terms of concept and in time, is able to reveal how difficult it is to perform it in current search engines. This example is particularly illustrative. With a careful observation we can note the fact that the football and the volleyball world cup occurred (except in 2002) in the same period in the same countries, so one can infer each meaning of the query term, first by using the timeline, secondly by using topic categorization. As the examples show, this is not a new problem, but a new challenge, that is particularly useful for poor and ambiguous queries. Together with a clustering approach temporal information can therefore become a crucial component for any search engine system, favouring the exploration of information in a historical perspective and leading search engines to a new level of interaction retrieving a snapshot of the entire collection and not only the latest results.

2 RELATED WORK

Temporal information is everywhere and available in every document (Alonso et al, 2007), but despite being an important dimension on information retrieval area, little research has been done till now. In fact, search engines are still limited to the most recent snapshot missing one of its most important dimensions, the temporal one, which poses two interesting problems. On the one hand, the information captured today (a price of a book or the headline of an e-news) may be gone tomorrow. On the other hand, if historical data still exists, it is lost, due to an inappropriate capture system, where temporal information is not considered. The solution to these problems may be in the introduction of timelines, through web archives and temporal search engines which, may favour the exploration of the information. In the following subsections we will present some works which try to overcome these limitations.

2.1 Web Archives

Web Archives deal with the challenges related to the constant growth of the web, allowing a multi-dimensional view of the web by storing and providing access to the past versions of web pages no longer available. As (Nunes, 2007) and (Song et al, 2008) point out, the web is a very dynamic environment where a large amount of web pages is created and removed at an impressive pace. For someone without access to historical data, the web is primarily one-dimensional, containing only one version of any page (Adar et al, 2008). Web archiving has appeared in this context to preserve these web pages, making it possible for researchers to study the evolution of the web, with emphasis on the Internet Archive (IA) which is currently the largest digital library in the world. Although historically important, IA presents some limitations for those who aim to study web page properties, as its dataset cannot be easily accessed, making it difficult to find and compare historical information.

For now, web archiving proposals still present some limitations, mainly in the search for information, as the user has to specify the URL he is looking for, and in its exhibition, which is hampered by the lack of a proper navigation through all available versions. To overcome part of these problems, (Jatowt, 2006) proposes a unified access browser that presents to the user, in response to an URL input and an initial time point definition, past versions of web pages spread over different repositories. (Jatowt, 2008**)** proposes an approach for visualizing summarized page histories based on data extracted from web archives together with a term clouding approach. (Toyoda et al, 2003) proposes a method for observing the evolution of web communities. (Song et al, 2008) present an interesting paper, where authors propose a crawling system that only stores web contents when there are differences, which is a step forward compared with the IA approach. Finally, Chronica (Denis et al, 2006) allows the user to search the Internet Archive showing the popularity of a tag over a time. As referred by (Jatowt, 2006), it is reasonable to assume that in the future, IR systems will be able to retrieve data from the live and the past web, a mixed approach yet not adopted by conventional IR systems.

2.2 Temporal Search Engines

If we have the multi-dimensional web, with temporal search engines we have access to the historical perspective of the one-dimensional web. As opposed to the web archive approach, whose aim is to store and preserve the various versions of a web page, the purpose of a temporal search engine is to retrieve a set of web pages through a timeline perspective. The introduction of this kind of search engines compared with current search mechanisms (see Figure 5), leads search engines to another level of interaction, being a step forward as opposed to a snapshot view, with the user having the chance to analyze the results through a temporal perspective.

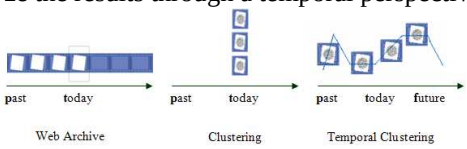

Figure 5: Three different search engine approaches.

Currently, only two works propose generic temporal search engines. (Jin et al, 2008) propose a temporal search engine towards web contents, supporting both temporal queries and keyword-based queries. (Alonso et al, 2007) propose a search engine which captures temporal expressions extracted from documents, retrieving the results by relevance based on the tf/idf measure and presenting them on a timeline made of overlapping clusters labelled by year. There are some other commercial approaches like Google Timeline which retrieves an extensive list of web pages replacing the use of web snippets through a summarization of the contents per year, or Viewzi a search engine that provides several visual search features, with emphasis on timeline, photo tag clouding, etc. Some other works are more specific, like the work presented by (Dubinko et al, 2006), which tries to understand and track the evolution of photo tags within the Flickr system over a period of time. Another interesting work is (Adar et al, 2008) which describes a semantic tagging system that extracts temporal information from news messages, while (Koen et al, 2000) introduces Time Frames for extracting time information from news articles. Other works, such as (Shaparenko et al, 2005) consider the problems of analyzing the temporal development of a document collection, trying to understand which topics are popular and how did their popularity change over time.

3 OUR APPROACH

Although the retrieved search engines results are now more accurate, quite often they do not meet the demands of the user, especially when the query is

ambiguous. Most systems provide some post-query features (query expansion and user personalization), but in the recent years clustering systems have been gaining pace as an alternative mechanism to present the results and to solve term ambiguity. Despite this fact, most search engines limit to capture the most recent snapshot of the web documents, disregarding information such as temporal dimensions. In this paper our purpose is to present a temporal search engine that presents peaks of relevant results over a clustering timeline, where the user can browse through all the retrieved results, look for years of high relevance and explore particular items, helping the user to gain a better understanding of the collection by exploring the interface. With the introduction of temporal dimension together with the display of the results in a clustering system, users are more likely to infer the knowledge they are looking for, helping to solve one of the most interesting problems of IR: term ambiguity.

Compared to other search engines, we intend to add temporal web search facilities, by capturing temporal information and focusing on the extraction of non-temporal content embedded in web pages. Like referred by (Alonso et al, 2007), temporal information is everywhere and is available in every document, either explicitly, in the form of temporal expressions, or implicitly in the form of metadata. The most obvious approach is to use temporal metadata attributes. Some works (Samia, 2003) and (Desikan et al, 2002) have also introduced the concept of temporal web mining as the application of temporal data mining and web mining techniques, to discover, extract, analyze and predict data with significant temporal information. However, well-established Natural Language Processing techniques, by applying time-entity extraction tools, are likely to produce improved results with higher coverage.

After this, the first step is to cluster the results by year, within which, reflecting the fact that a web page may contain different meanings of the query terms, documents are grouped into one or more clusters made of web pages whose relevance is determined through the application of web content mining techniques introduced in the scope of previous work (Campos et al, 2008). The final result is a time-based clustering system that, beyond helping with solving term ambiguity (through concept and time disambiguation), presents an alternative view of the retrieved results. In this regard, the system should be capable of analysing whether there is a need to visualize the results clustered by time or by topic. It seems obvious that for some queries, like swine virus, it won't make much sense to present a unique cluster named 2009.

In summary, our proposed system is an improvement of two other proposals by (Alonso, et al, 2007) and (Jin et al, 2008). Both are temporal search engines, but while the first one does not present the existing information inside each year clustered in a hierarchical manner the latter only presents them as a simple list of results. In Figure 6 we describe the architecture of our system.

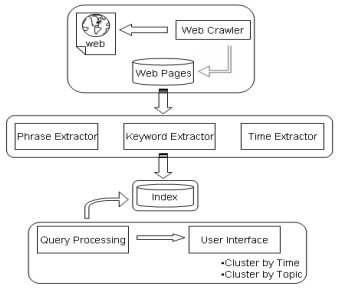

Figure 6: Architecture.

In detail it is composed of eight modular and independent components: Web Crawling; Document Parsing (Phrase Extraction; Keyword Extraction; Time Extraction); Index Constructor; Clustering by topic; Clustering by time; Cluster Labelling; Cluster Distribution Analysis; Visualization.

Specifically, it will be possible to present the results by topic or by temporal dimension. In any case, both will have some keywords able to overall describe the topic (see Figure 7) or the year (see Figure 8), which in turn are likely to be used for query refinement and personalization.

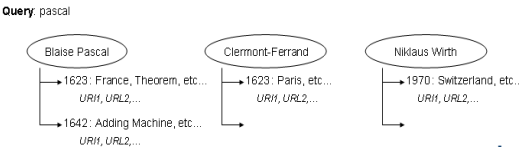

Figure 7: Example of the query Pascal clustered by topic.

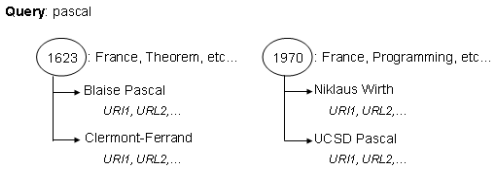

Figure 8: Example of the query Pascal clustered by time.

Overall, we think our solution is innovative, since we provide an alternative view to present the results in a time-based manner, clustered by time and topic, helping on solving the problem of term ambiguity. Moreover, we consider phrases made of contiguous terms instead of single words, web content mining techniques to represent the documents, and a clustering approach as opposed to an ordered list of relevant results.

4 CONCLUSIONS

Making time one essential feature of web contents, temporal information can become a very useful and meaningful dimension in search engines. Ideally, we would like a search engine to be aware of the temporal information embedded in documents and present the results in a time context (Alonso et al, 2007). The use of a temporal dimension would allow the user to better fit a concept within such a dynamic web context, improving its functionality, however, search engines still do not take much advantage of combining temporal aspects of web content and user experience to enhance the results.

Currently there are two approaches known as web archiving and temporal search engines. While the first one mainly focuses on the preservation of the web, the second one deals with the display of the results in a timeline perspective. In our approach we aim at providing a historical perspective of the one-dimensional web, by offering an alternative presentation of the results based on a clustering list of the web documents related with the same temporal data. With the introduction of a temporal dimension together with the exhibition of the results in a clustering system, users are more likely to infer the kind of knowledge they are seeking for, refining their query and personalizing their search results plus solving one of the most interesting problems of IR: term ambiguity.

For evaluation, we plan to execute user feedback surveys, which have been the most favoured techniques in order to evaluate the quality, precision and recall of the results. Given their difficulty in terms of logistic and subjectivity we also intend to perform a comparison between our system and other search engines, in the lines of what has been proposed by (Jin et al, 2008).

ACKNOWLEDGMENTS

This work is supported by the VIPACCESS project funded by the Portuguese Agency for Research (*Fundação para a Ciência e a Tecnologia*) with the reference PTDC/PLP/72142/2006.

REFERENCES

Adar, E., Dontcheva, M., Fogarty, J. and Weld, D., 2008. Zoetrope: interacting with the ephemeral web**.** In Proc. of 21st ACM Symp. User Interf. Soft. and Tech. USA.

Alonso, O., Baeza-Yates, R. and Gertz, M., 2007. Exploratory search using timelines. In SIGCHI Workshop on Exploratory Search and HCI Workshop.

Alonso, O. and Gertz, M., 2006. Clustering of search results using temporal attributes. Proc. of 29th SIGIR

Alonso, O., Gertz, M. and Baeza-Yates, R, 2007. On the value of temporal information in IR. In Proc. of ACM SIGIR, Vol. 41 , Issue 2, pp 35-41, ISSN:0163-5840

Campos, R., Dias, G., Nunes, C. and Nonchev, B., 2008. Clustering Web Page Search Results: A Full Text Based Approach. In International Journal of Computer and Information Science Vol 9(4). pp 29-40.

Deniz, E., Chris, F. and Terence, J., 2006. Chronica: Temporal Web Search Engine**.** In Proc. of ICWE.

Desikan, P. and Srivastava, J., 2002. Mining information from temporal behaviour of web usage. Minnesota.

Dubinko, M., Kumar, R., Magnani, J., Kovak, J., Raghavan, P. and Tomkins, A., 2006. Visualizing tags over time. In Proc. of the 15th Int. Conf. on WWW 2006, Scotland. Pp 193-202, ISBN:1-59593-323-9.

Jatowt, A., Kawai, Y., Nakamura, S., Kidawara, Y. and Tanaka, K., 2006. Journey to the past: proposal of a framework for past web browser. In Proc. of 17th Conf. on Hypertext and Hypermedia, Denmark.

Jatowt, A., Kawai, Y. and Tanaka, K., 2008). Visualizing historical content of web pages. In Proc. of 17th International Conference on WWW, pp 1221 – 1222, Beijing, China. ISBN:978-1-60558-085-2.

Jin, P., Lian, J., Zhao, X. and Wan, S., 2008. TISE: a temporal search engine for web contents**.** International Symp. on Intelligent IT Application. Shanghai, China.

Koen, D. and Bender, W., 2000. Time frames: temporal augmentation of the news**.** IBM Systems Journal, Volume 39, Issue 3-4, pp 597–616, ISSN:0018-8670.

Nunes, S., 2007. Exploring temporal evidence in web information retrieval. In Proc. of the Future Directions in Information Access, Glasgow, Scotland, pp 44 – 50.

Plachhouras, V., 2007. Temporal aspects of web search. Yahoo! Research, Barcelona.

Samia, M., 2003. Temporal web mining. In Proc. of 15th Work. on the Foundations of DB, pp 27–31, Germany.

Schilder, F. and Habel, C., 2001. From temporal expressions to temporal information: semantic tagging of news messages. In Proc, of ACL'01, pp 65–72, Toulouse, France.

Shaparenko, B., Caruana, R., Gehrke, J. and Joachims, T., 2005. Identifying temporal patterns and key players in document collections. Proc. of ICDM, 165–174, USA.

Song, S. and JaJa, J., 2008. Archiving Temporal Web Information: Organization of Web Contents for Fast Access and Compact Storage. TR Univ. of Maryland.

Toyoda, M. and Kitsuregawa, M., 2003. Extracting evolution of web communities from a series of web archives**.** Proc. of 14th ACM conference on hypertext and hypermedia, pp 28–37, ISBN: 1-58113-704-4.