Exploring Temporal Evidence in Web Information Retrieval

Sérgio Nunes
Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, s/n 4200-465 Porto PORTUGAL
sergio.nunes@fe.up.pt

Abstract: Web Information Retrieval (WebIR) is the application of Information Retrieval concepts to the World Wide Web. The most successful approaches in this field have modeled the web's structure as a directed graph and explored this concept using different approaches. Within this line of research, HITS and PageRank are two of the most well known paradigms for evaluating the importance of web documents. Most of this research has origins in the area of citation analysis, but although time is an important dimension in the citation analysis literature, it hasn't been explored in depth within WebIR. Recent studies show that the web is a highly dynamic environment, with significant changes occurring weekly. The Blogospace is a good example of this very active behavior. In this work, temporal web evidence is identified and categorized according to two classes, one based on features extracted form individual documents and the other based on features extracted from the whole web. Also, a broad survey of previous work exploring temporal evidence is presented. Finally, ideas for exploring temporal web evidence in typical web tasks are briefly discussed. The lack of suitable corpora containing temporal evidence has been a deterrent to research on this field. The recent availability of public datasets containing temporal information has raised public awareness of this topic.

Keywords: web information retrieval, temporal evidence, document ranking

1. INTRODUCTION

The World Wide Web is a vast repository of data and an increasingly important information source for millions of users worldwide. In November 2006, the Web reached a new milestone when Netcraft reported that there were more than 100 million web sites online[1]. To deal with this amount of information, search engines have become primary tools for users while online. Search is the second most common activity for Internet users, close behind email access and use.

A web search system, or search engine, typically consists of four main components, a crawler, a document indexer, a query processor and a presentation interface [1]. The crawler and indexer work offline to build an internal representation of a portion of the web prepared for fast access. The query processor and presentation interface only access this internal representation and work online in response to user queries. Baeza-Yates et al. [2] draw an analogy between web crawling and the task of an astronomer watching the sky. Web search engines users don't access the current state of the web but an image of what the crawler captured at a specific time. Currently, there is no temporal context provided to the users in standard web search engines. Results are treated and presented without distinction, despite being from a recently created web document or from a historic, stale document.

Typically, and despite the fact that the web is a very dynamic environment, web search systems do not fully explore the temporal dimension in their algorithms. Archived web documents are mostly used as a source of historical information and not exploited for improving current web tasks or meeting user's current needs. Ntoulas et al. [3] have analyzed the evolution of both content and link structure of web pages, specially focusing on aspects of potential interest to search engine designers. Crawls were performed over the course of a year, starting in late 2002. Each weekly snapshot had an average size of 65GB, resulting in a final dataset of more than 3.3TB. Both change frequency (how often) and change degree (how much) were analyzed. They found that only 20% of web pages last one year and that, after a year, 50% of the content on the web is new. Link structure was observed to be more dynamic than content, 25% new links are created every week, while only 8% new pages are created and 5% new content is produced.

To summarize, current research on web dynamics shows that the web is very active, exhibiting both high decay rates and high creation rates. Content is being created mostly through the production of new pages (rather than updating existing ones) and once created, these pages evidence very high decay rates. Nonetheless, it is important to note that recent studies have suggested that lasting contents tend to be referenced by different URL during their lifetime [4]. Although there is a strong connection between past frequency of update and future frequency of update, the correlation between frequency of update and degree of update is small. Despite the fact

[1] http://www.cnn.com/2006/TECH/internet/11/01/100millionwebsites/index.html

that more detailed studies are necessary on this topic, page templates seem to have little impact on change degree.

2. PREVIOUS WORK

Previous research on the temporal dimension of the web for information retrieval is structured according to three types, depending on the main source of temporal evidence being explored. Link-based research uses links, both in-links and out-links, in a temporal context to refine information retrieval. Content-based research examines document's contents from a temporal point of view. Finally, metadata-based research uses evidence gathered from HTTP headers and external services (e.g. PageRank values).

Using temporal evidence to improve information retrieval on the web is an area with growing popularity. First, the temporal dimension was explored with the main goal of cost reduction by focusing on crawler scheduling to optimize resource consumption [5]. Also, several works have contributed to the temporal characterization of information on the web [6, 3, 7]. More recently, attentions have turned to tasks related to result ranking [8], web mining [9], summarization [10] or question answering [11]. Overall, this is an area still in its infancy.

2.1 Link-Based Approaches

Berberich et al. [8] propose an algorithm, named T-Rank, which extends PageRank to improve page ranking by exploring the freshness and activity of both pages and links. A temporal model for link analysis is proposed where each node and edge in the web graph is annotated with timestamps. These timestamps represent different types of events (e.g. creation, modification, deletion). Also, to capture the user temporal focus, a temporal window of interest is defined based on two timestamps that limit the relevant period. Given a temporal window of interest, freshness is defined by a linear function that is maximum if timestamps occur within the user defined period and decreases linearly if they occur within the tolerance intervals. The activity of an object (page or link) is defined as the sum of all modifications occurred within the time window.

The T-Rank algorithm extends PageRank by modifying the underlying probabilities of the random walk to favor certain nodes. The transition probability from node x to node y is defined as a weighted combination of the freshness of the target node y, the freshness of the link between x and y, and the average freshness of all incoming links of y. On the other hand, the random jump probability of a target page x is a weighted combination of the freshness of x, the activity of x, the average freshness of the incoming links of x, and the average activity of the pages that link to x. Just like PageRank, T-Rank values are computed using the power iteration method. The quality of this proposal was assessed using the DBLP corpus and an Amazon's products pages.

Yu et al. [12] argue that current top search algorithms (e.g. PageRank and HITS) miss an important dimension of the web by not considering its temporal characteristics. Since these algorithms rely largely on the total number of accumulated links, older pages tend to be favored. The PageRank algorithm is adapted by weighting each citation according to the citation date. The new technique, dubbed TimedPageRank, uses an exponential decay function to weight citations. An aging factor is also included in the final formula so that node scores decline linearly with time. This algorithm is further refined by also including TimedPageRank scores for authors and journals.

2.2 Content-Based Approaches

Jatowt et al. [13] address the problem of web document summarization by exploring previous versions of the documents. This temporal web page summarization analyzes the content retrieved from temporally distributed versions of single web documents to produce a summary of the main concepts addressed over a given time period. This method is applicable to documents that are updated frequently. The reasoning behind this approach is that, by using historical versions of a document, more content becomes available, in turn leading to improvements to the final summary. To observe each term's growth, regression analysis is used to summarize the relationship between time and the frequency of a term. A formula combining the slope of the regression line, the interception point and the variance is used to rank terms. Finally, sentences are selected and ranked based on their average term scores. This work was further developed to address multi-document summarization [10].

Liebscher et al. [14] analyze content over time to identify rising or falling terms. The authors try to identify lexical dynamics so that current search results can be amplified or dampened. Dissertation abstracts and discussion board posts were used in an experimental setup. Each large corpus was split into multiple independent corpora, each associated with documents occurring within a specific time interval. Temporal trends were then computed using simple time series analysis techniques (i.e. linear models). Finally, by comparing slopes, the most dynamic topics were identified. The authors argue that this information can be exploited to improve information retrieval. The argument is that previously popular terms, belonging to documents that capture more fundamental aspects of a topic, should be amplified. On the other hand, currently popular terms that were once rare should be dampened so that new topics are not overemphasized.

To address the problem of question answering on the web, Yamamoto et al. [11] also explore web archives to assess the trustworthiness of user given statements. The rationale is that if a phrase has been continuously stated for a long time on the web, its reliability is higher. For each possible proposition, temporal profiles using a public web archive are produced representing the growth (or decline) of occurrences over time. Using time series analysis, these profiles are then used to estimate the statement's accuracy.

2.3 Metadata-Based Approaches

Berberich et al. [15] propose a method that is complementary to time-agnostic ranking algorithms. By analyzing time series of importance scores (like PageRank scores), BuzzRank identifies growth trends in these scores using generic growth models and curve fitting techniques. For instance, in a reported experiment using the DBLP bibliographic dataset, BuzzRank highlighted papers that, though not having the highest PageRank value, had the highest PageRank growth. Since PageRank scores over time are not directly comparable, graphs were normalized treating missing nodes as dangling nodes. The presented version of BuzzRank is very expensive both in terms of computing power and storage requirements because it needs to store the entire graph and perform PageRank computations for each time interval.

Amitay et al. [9] observe HTTP headers, particularly the Last-Modified field, and are able to reveal significant events and trends. The Last-Modified field is used to approximate the age of the page's content. Using this information to timestamp web resources, several interesting applications are explored. It is clearly shown that real life events can be exposed mainly due to what authors call fossilized content. In a nutshell, three phases are involved in the discovery of this fossilized content. First, by issuing one or several queries to public search engines, a topical collection is assembled. These queries are manually selected so that they characterize the desired topic. Then, a second collection is built combining the links that point to the URL in the first collection. Finally, the Last-Modified values of the pages in the second collection are gathered and a histogram is plotted. The authors report several experiments where clear patterns are visible. Also exploring the notion of timestamped web resources, the authors introduce the concept of timely authorities, opposed to simple link-based authorities. This idea is illustrated with the adaptation of the HITS and SALSA algorithms, adjusting vertices weights to include a time dependent bonus.

3. SOURCES OF TEMPORAL WEB EVIDENCE

Following the classification presented in Upstill [16], possible sources of temporal information on the web are organized in two groups, namely document-based evidence (or features) and web-based evidence. The former encompasses all features extracted from individual documents, while the latter holds features that are obtained from the whole web. Alternatively, these features might be arranged according to their temporal order. First order features are those that are direct sources of temporal information (e.g. the Last-Modified value in HTTP response headers). Second order temporal features are obtained from the observation of standard (non-temporal) features through time (e.g. a document's size evolution).

3.1 Document-Based Evidence

Document-based temporal evidence is obtained by exploring the characteristics of single web documents. These characteristics are limited in scope and, typically, user-generated. Thus, they are subject to direct manipulation and might be engineered. Document-based temporal evidence is classified according to three types: content, URL address and HTTP protocol.

3.1.1 Content

A web document's content might be explored in various ways to gather temporal evidence. Using natural language processing (NLP) techniques, specifically information extraction methods, it is possible to identify words or expressions that convey temporal meaning (e.g. "today", "a long time ago") and use these to date documents [17]. On the other hand, the evolution of a document's content through time can be viewed as a single temporal feature. This new feature encompasses multiple metrics derived from well-known IR concepts (e.g. term frequency, term count). Observing HTML markup code might also be of value to obtain temporal evidence (e.g. evolution of out-links). Also, content differences between versions of the same document can be viewed as document-based temporal evidence. Research in finding near-duplicate web pages [18] has led to the development of multiple algorithms for this task. For instance, comparing content checksums (or signatures) on web documents is a naive technique that is rarely used. Very similar documents, differing only in a word, character or number, have disparate signatures. Alternatively, the shingling technique, proposed by Broder et al. [19], is one of the most popular algorithms in WebIR. This algorithm produces a single value that can be used to compute a similarity degree. In a nutshell, documents are structured in sets of fixed length shingles (or consecutive term sequences) and then the similarity of the resulting sets is compared.

3.1.2 URL Address

All public web documents have at least one unique address. Occasionally, documents might have multiple addresses that resolve to the same web page. The different segments of an URL, namely host, path and search part, might be used as a source of temporal information. For instance, a current New York Times URL is structured in the following way - http://www.nytimes.com/2007/04/30/world/europe/30france.html. It is possible to derive the document's year, month and day of publication by parsing this URL. Based on this information, it would be possible to estimate the document's inception date.

3.1.3 HTTP Protocol

The Hypertext Transfer Protocol (HTTP) is an application level protocol used to request and transmit hypertext documents and components between user applications and online servers. Clients submit standard request to servers identifying specific resources. Servers reply to each request, sending standard headers and, if available, the requested resource (body of the message). HTTP headers are comprised of fields, part optional and part required. Of interest is the Last-Modified field representing the date and time at which the resource was last modified. This HTTP header field might be used to add temporal information to a web resource, specifically a web document. However, this field is not always available and may not return a valid date. This erratic behavior is generally attributed to incorrectly configured web servers. Several independent studies have estimated that 35% to 80% of web documents have valid Last-Updated values [6, 9, 4]. Despite this problem, it is possible to explore this information as shown by Amitay et al. [9].

3.2 Web-Based Evidence

The entire web is an important source of information about individual web documents. In other words, multiple independent sources are combined to produce information about a particular web resource. One distinct advantage of this approach is that it is hard to influence or tinker. On the web, a public document is integrated in a wider context. A document is connected to other documents and resources through a hypertext mesh (the document’s neighborhood). On the other hand, in a web information system there are inherent components that are continuously gathering information, such as external information repositories or web server logs. As shown below, each of these components might be explored as a source of web-based temporal evidence.

3.2.1 Neighbors

Links are in the nature of a hypertext system like the web. Web resources, or graph nodes, are connected by referencing other nodes or by being referenced by them. Inside this network it is possible to define the concept of a web document’s neighborhood, the set of nodes that point to a document or are pointed by it. Considering a document’s vicinity it is possible to derive multiple features. Regarding temporal evidence, a document’s in-links and out-links may be observed through time to reveal trends or unexpected patterns (e.g. link farms for spamming).

3.2.2 External Archives

There are several services with the mission of archiving public web data. The Internet Archive (IA), a non-profit organization, is the leading initiative in this field. The IA has been archiving the web since 1996 and its collection is open to public access. Typically, the data is collected through periodic web crawls and gathered using standard format archives. For each web resource found, multiple snapshots are collected through time. After acquiring old copies of specific web pages, content matching algorithms might be applied to produce temporal evidence. There are other services from where it is possible to extract dated copies of web resources, most notably search engine’s caches and large web crawls available for research.

3.2.3 Web Logs

A web information system is a combination of multiple smaller subsystems that work together to provide a seamless experience to the users. These subsystems are very heterogeneous, ranging from a user’s browser to complex HTTP server software capable of handling millions of requests per second. Most of these subsystems’ activities are recorded in a persistent fashion, usually in flat file logs. Two types of web logs can be explored as sources for temporal evidence, namely query logs and access logs. Query logs record users search intentions, while access logs record users access to web resources. Both types are filled with temporal markup that can be used to annotate web resources. The major problem of web logs research is the private nature of the data, making these logs very difficult to access and raising significant privacy issues.

4. EXPLORING TEMPORAL WEB EVIDENCE

Consider the two types of temporal evidence proposed above, document-based and web-based. An initial research approach might the exploration of both types of evidence to develop a single ranking formula. Figure 1 illustrates the concept. The final score of a web document is obtained by combining the evolution of document-based features and web-based features. As discussed above, there are ranking algorithms that take into account the temporal nature of the web. However, I’m not aware of any algorithm that has directly explored changes in web documents to improve ranking. T-Rank [8] observes changes in documents only to produce timestamps that are added to documents. The content of the changes is ignored. Jatowt et al [10] use historical content but on a summarization setting.

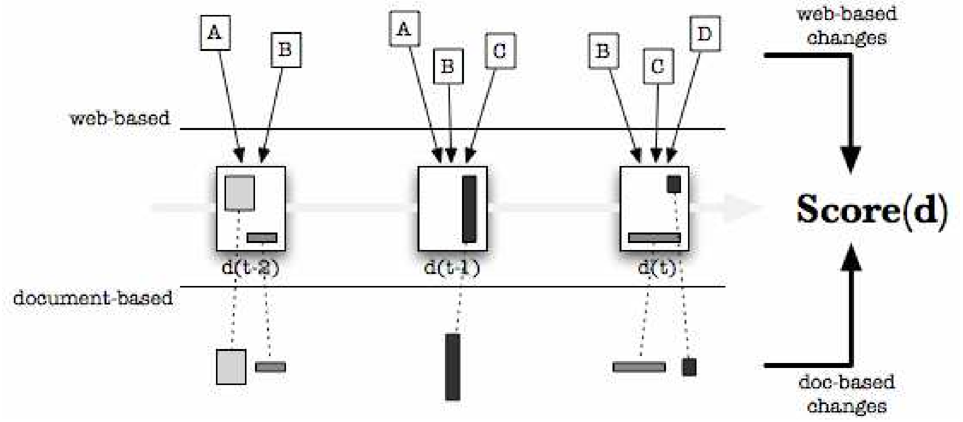

FIGURE 1: Web Document Scoring Combining Web and Document Changes.

An initial working hypothesis can be stated as follows: *The ranking of web documents in web search can be improved if temporal evidence, both document and web-based, is added to the ranking formula. This evidence may be derived directly from the documents or from their surrounding communities.*

Two complementary approaches can be used to validate this hypothesis. First, using collections containing temporal evidence, it is possible to compare a ranking based on temporal evidence with standard well-known approaches to web ranking. A standard web ranking algorithm, like PageRank, can be used to rank the documents in these collection without resorting to temporal information. In other words, the algorithm only looks at the latest snapshot within the collection. The resulting rank can be used as a baseline and represent an indication of the quality of each page. Higher ranks stand for formally higher quality documents. Then, the proposed algorithms can be used to produce an alternative ranking exploring temporal evidence from the entire collections. Finally, using rank correlation metrics, the various ranks can be compared. Based on these results, it will be possible to verify if temporal evidence contributes to identify quality web documents.

The second suggested approach is based on a standard blind user study of a web search system built over a collection containing temporal information. This system ranks results using both standard ranking algorithms and the proposed algorithms based on temporal evidence. To eliminate layout biases, the results from the two rankings being compared are presented on a single interleaved list (removing any duplicates). For each search query submitted to the system, a single rank combining the two alternative ranks is presented. Using this procedure it is possible to evaluate which ranking scheme generates more clicks. This procedure is based on the evaluation methods adopted by Das et al. [20].

Obtaining suitable corpora for these evaluation procedures might by a complex task. Existing corpora are typically static, not including any temporal information. There are two possible sources of collections containing temporal evidence, the Stanford WebBase Project and the European Archive. The Stanford WebBase [21] corpus is a large scale collection created and maintained by the InfoLab from Stanford University. This web repository holds more than 100TB (as of March 2007) of topic-focused snapshots of web sites. Of special interest is a weekly crawl of a small set of sites between January and May 2006. Around 2 million pages were crawled from 1042 top web sites during 21 weeks. This data is publicly available via a proprietary client or a standard web interface.

The UKGOV Weekly and Monthly Web corpora are two collections made available by the European Archive foundation[2]. The weekly collection contains weekly snapshots of 11 United Kingdom's governmental web sites. The monthly collection holds monthly snapshots of 49 UK's governmental web sites. The number of crawls available for each web site varies significantly. Oldest crawls start in 1997. Both datasets are publicly available via web.

5. DISCUSSION

This work presents a first approach at organizing sources of temporal web evidence. A broad survey of previous work, where the temporal dimension has been explored, is also included. As temporal corpus becomes available, a significant barrier will be removed and research on temporal web information retrieval will be more accessible. Currently, a small number of existing datasets include temporal evidence. Examples of temporally rich datasets include public archives (Internet Archive and European Archive), research projects (Stanford's WebBase project) and open projects like Wikipedia. The Blogospace, due to its very dynamic nature, has also been used as source of data for temporal exploration.

The World Wide Web is still a relatively young information system. Although of little relevance in these early days where "everything is new", time will become an increasingly important dimension to users. It is easy to imagine use

[2] http://www.europarchive.org

cases where temporal-based features play a critical role. For example: searching in a delimited temporal period or distinguishing between recently popular web resources and “all-time” popular resources. Since time is a transversal dimension, it has impact on all levels of WebIR. In this paper, the focus is on the scoring of web documents in the tasks of ad-hoc information retrieval and topic-distillation.

ACKNOWLEDGEMENTS

I would like to acknowledge Cristina Ribeiro and Gabriel David (PhD advisor and co-advisor) for reading and commenting on preliminary versions of this text. Their contributions have greatly improved the final document. I would also like to thank two anonymous reviewers, whose valuable input led to significant improvements to the final version of this document. This work was financially supported by Fundação para a Ciência e a Tecnologia (FCT) and Fundo Social Europeu (FSE – III Quadro Comunitário de Apoio), under grant SFRH/BD/31043/2006.

REFERENCES.

[1] Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 30(1-7), 107–117.
[2] Baeza-Yates, R., Castillo, C. and Saint-Jean, F (2004). Web dynamics, structure and page quality. Proceedings of Web Dynamics’04, pp.~93–109. Springer Verlag.
[3] Ntoulas, A., Cho, J. and Olston, C. (2004) What’s new on the web?: the evolution of the web from a search engine perspective. Proceedings of the 13th international Conference on World Wide Web, pp.~1–12. ACM Press.
[4] Gomes, D. and Silva, M. (2006) Modelling information persistence on the web. Proceedings of the 6th international conference on Web engineering, pp.~193–200. ACM Press.
[5] Cho, J. and Garcia-Molina, H. (2000) The evolution of the web and implications for an incremental crawler. Proceedings of the 26th International Conference on Very Large Data Bases, pp.~200–209. Morgan Kaufmann Publishers Inc.
[6] Brewington, B. and Cybenko, G. (2000) How dynamic is the web? Computer Networks, 33(1-6), 257–276.
[7] Fetterly, D., Manasse, M., Najork, M. and Wiener, J. (2004). A large-scale study of the evolution of web pages. Softw. Pract. Exper., 34(2), 213–237.
[8] Berberich, B., Vazirgiannis, M. and Weikum, G. (2004) T-rank: Time-aware authority ranking. Proceedings of the 3rd International Workshop on Algorithms and Models for the Web-graph, pp.~131–142. Springer.
[9] Amitay, E., Carmel, D., Herscovici, M., Lempel, R. and Soffer, A. (2004) Trend detection through temporal link analysis. J. Am. Soc. Inf. Sci. Technol., 55(14), 1270–1281.
[10] Jatowt, A. and Ishizuka, M. (2006) Temporal multi-page summarization. Web Intelli. and Agent Sys., 4(2), 163–180.
[11] Yamamoto, Y. Tezuka, T., Jatowt, A. and Tanaka, K.. Honto? search: Estimating trustworthiness of web information by search results aggregation and temporal analysis. Proceedings of the 9th Asia-Pacific Web Conference and the 8th International Conference on Web-Age Information Management. Springer Verlag.
[12] Yu, P., Li, X. and Liu, B. (2004) On the temporal dimension of search. Proceedings of the 13th international World Wide Web Conference, pp.~448–449, New York, ACM Press.
[13] Jatowt, A. and Ishizuka, M. (2004) Temporal web page summarization. Proceedings of WISE 2004, pp.~303–312. Springer-Verlag.
[14] Liebscher, R. and Belew, R. (2003) Lexical dynamics and conceptual change: Analyses and implications for information retrieval. Cognitive Science Online, 1.
[15] Lu, H., Luo, Q. and Shun, Y. (2003) Extending a web browser with client-side mining. Proceedings of Web Technologies and Applications: 5th Asia-Pacific Web Conference, pp.~166–177. Springer Berlin / Heidelberg.
[16] Upstill, T. (2005) Document ranking using web evidence. PhD thesis, The Australian National University.
[17] Wong, K., Xia, Y., Li, W. and Yuan, C. (2005) An overview of temporal information extraction. International Journal of Computer Processing of Oriental Languages, 18(2), 137–152.
[18] Henzinger, M. (2006) Finding near-duplicate web pages: a large-scale evaluation of algorithms. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 284–291. ACM Press.
[19] Broder, B., Glassman, S., Manasse, M. and Zweig, G. (1997) Syntactic clustering of the web. Proceedings of the 6th International Conference on World Wide Web, pp.~1157–1166. Elsevier Science Publishers Ltd.
[20] Das, A., Datar, M., Garg, A. and Rajaram, S. (2007) Google news personalization: scalable online collaborative filtering. Proceedings of the 16th International Conference on World Wide Web, pp.~271–280. ACM Press.
[21] Cho, J., Garcia-Molina, H., Haveliwala, T., Lam, W., Paepcke, A., Raghavan, S. and Wesley, G. (2006) Stanford webbase components and applications. ACM Trans. Inter. Tech., 6(2), 153–186.