Towards Information Retrieval Evaluation over Web Archives

Miguel Costa
Foundation for National Scientific Computing
Lisboa, Portugal
miguel.costa@fccn.pt

Mário Silva
University of Lisbon, Faculty of Sciences
LaSIGE
Lisboa, Portugal
mjs@di.fc.ul.pt

ABSTRACT

We present the first overview of a web archive user profile and the searching technology that supports it. Most web archives only support URL search and just a few provide full-text search in response to users' expectations. Their technology is essentially based on web search engines, which ignore the temporal dimension of collections. As consequence, the quality of results is poor. We suggest the creation of an initiative for information retrieval evaluation, meeting the needs of web archives. We believe this initiative would foster research in web archives, in resemblance with what other initiatives achieved in their domains.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search Process; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Experimentation, Measurement, Design

Keywords

web archives, ranking, evaluation

1. INTRODUCTION

All kinds of information are published on the world wide web. Part of this information is unique and historically valuable. However, since the web is too dynamic, a large amount of information is lost everyday. Several initiatives started to archive parts of the web, mainly to preserve their web heritage (see `http://www.nla.gov.au/padi/topics/92.html`). The Internet Archive is the most ambitious initiative with 150 billion documents archived since 1996. As time passes, more and more documents will be archived and their historic interest increased with age. These collections of web data offer a great potential to understand the past, but that requires the development of mechanisms to access this information in areas so diverse as sociology, history, anthropology, culture, politics or journalism.

The prevalent access in web archives is based on the search over automatically extracted metadata from web documents, specially their URLs. A URL search returns a list of the versions of that URL chronologically ordered, such as in the Internet Archive's Wayback Machine (see `http://www.archive.org/web/web.php`). However, the requirement of the user having to know the URL limits its use. A web archiving user survey indicates that full-text search is the most desired web archive functionality [6]. Users expect an interface similar to the one offered by web search engines. In conformity with this idea, a few web archives have implemented full-text search. However, all are based on the Lucene search engine, which is the core of NutchWAX, an extension of the Nutch search engine with Web Archive eXtensions. All the institutions managing these web archives are members of the International Internet Preservation Consortium (IIPC), which has the goal of aggregating efforts to produce common tools and standards. This explains the convergence to NutchWAX and was also the reason for us to adopt it in the developing of the Portuguese web archive [3]. We have indexed until now more than 200 million documents. To the best of our knowledge, the Internet Archive performed the largest indexing over parts of its collection that have close to a billion documents.

This general tendency of adapting web search engines technology to provide full-text search for web archives raises several questions. Does the technology provide good results? Cohen et al. showed that the out-of-the-box Lucene produces low quality results, a MAP of 0.154, which is less than half when compared with the best systems participating in the TREC Terabyte track [2]. We believe that the specific characteristics of web archive collections that are not handled by Lucene, degrade even more the quality of results. Being time present in all the processes and foreseen solutions over a web archive, shouldn't time be present in the ranking model to provide better results for the users? If so, which combination of temporal attributes should be used: the crawl date, creation date, last-modified date or temporal expressions extracted from text with the help of NLP and information extraction technology? Temporal information retrieval (IR) uses temporal data embedded in documents and queries, implicitly or explicitly, to improve search results. Can the rich time-based characteristics of web archive collections be explored with temporal IR? Can we take advantage from the several versions of a document or from the evolution of its links? How should the results of successive crawls from the web be fused? How many versions of a document should be returned to the user? All these questions and others require a dedicated testbed to be studied.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

2. USERS' INFORMATION NEEDS

A clear understanding of what users search is fundamental for the development of web archives search functionalities and to evaluate their performance. A shallow analysis over the top queries at PANDORA's web archive (see `http://pandora.nla.gov.au/search-trends/`) indicates that web archive queries are short like web search engines queries, which contain on average around 2 terms [4]. Unexpectedly, there isn't almost any mention to dates or temporal expressions in web archive queries. This is in conformity with Nunes et al. analysis over the AOL logs that showed that only 1.5% of the queries mention temporal expressions [5]. Our preliminary experiments with users using the Portuguese web archive revealed that they also type short queries without temporal expressions. This may be due to the dominant use of web search engines that today influences the way how users search in other systems. On the other hand, users sometimes use a date range filter incorporated in the interface to narrow the search to a specific period. This filter exists in most web archives and in some cases serves to disambiguate queries. For instance, searching for 'Iraq war' can return documents about three different wars occurring in different periods. When the documents were published during each war, the 'Iraq war' query identified unequivocally the conflict. With the accumulation of all these documents, the query is insufficient to do so.

Users try to find specific pages to see them as they were published in the past. Sometimes they browse their archived versions after that to see for instance, the oldest or youngest version. This search for specific pages is a navigational need. Users also search information about a topic, such as in a topic distillation task. The difference is that web archive users want to see what was known and written about the topic in the past, recreating an historical period. For instance, a user can find what political leaders said about the invasion of Iraq led by the U.S. when it happened in 2003.

Besides navigational and informational queries, Broder classified another query type as transactional, when the query intent is to obtain a resource available via the web (e.g. download a file or buy a product) [1]. Despite the fact that this type is significant in web search engines, we did not detect transactional queries submitted by web archive users. One of the reasons why this occurred is that the web services supporting products purchasing are mostly discontinued when trying to access these services through archived pages. However, we envision that users will use web archives to download old files, for instance, an old manual.

3. TEST COLLECTION

Web archive collections are distinct due to their temporal dimension, so time must be present in the criteria to select the test collection elements: corpus, topics and relevance judgments. The corpus should follow the same diversity of subjects, literary styles and lengths, the same heterogeneity of formats and contents, and a similar word, language and link distribution. Web archives crawl and store different snapshots of the web from different periods. Some crawls are selective, for instance focusing in one sub-domain or topic (e.g. elections). These snapshots are narrower but deeper, trying to crawl all about the topic. More general snapshots, such as country codes top-level domains (e.g. pt), are wider, but more shallow. Another aspect is that some documents, such as newspapers, have a higher change rate, while others, such as scientific articles, tend to be static for long periods. Due to this heterogeneity in crawling frequency, the number of versions of a document can be highly variable. The versions can be very similar or even duplicates, while others are totally different. These characteristics affect the ranking algorithms. For instance, link-based algorithms such as PageRank would have to handle more sparse and versioned web graphs derived from these collections.

The topics must reflect the web archive users' information needs, as described in Section 2. Despite simplistic, the general web archive user profile portrays the user performing navigational or informational queries, some times restricted with a date range or a domain name. We are presently preparing a user survey and a study over the query logs to understand this profile better. We believe that there are at least two types of users: the casual user, whose behaviour and expectations are those of a web search engine user, and the researcher, who needs to explore a topic exhaustively over a timeline. We also want to understand the taxonomy and distribution of the various types of queries to see how different they are from the web search engines queries, analyse the search trends and all critical aspects to engineer effective searching systems and representative test sets.

4. CONCLUSION

The technology used to enable search in web archives provides unsatisfactory results to web search engines and was never evaluated over web archives. Time is the main feature of web archive collections and is completely ignored. Other problems were also raised in this paper that require investigation. Being IR mostly an empirical discipline, joint evaluation initiatives are undeniably important to foster IR research and technology. The elaboration of an initiative towards the evaluation of IR over web archive collections, seems like the natural next step to study the search technology under a set of controlled conditions. It is essential to demonstrate the superior effectiveness and robustness of some retrieval approaches and to produce sustainable knowledge for future development cycles.

5. REFERENCES

[1] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[2] D. Cohen, E. Amitay, and D. Carmel. Lucene and Juru at Trec 2007: 1-million queries track. In *Proc. of the 16th Text REtrieval Conference*, 2007.

[3] D. Gomes, A. Nogueira, J. Miranda, and M. Costa. Introducing the Portuguese web archive initiative. In *Proc. of the 8th International Web Archiving Workshop*, 2008.

[4] B. Jansen and A. Spink. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing and Management*, 42(1):248–263, 2006.

[5] S. Nunes, C. Ribeiro, and G. David. Use of temporal expressions in web search. In *Proc. of the Advances in Information Retrieval, 30th European Conference on IR Research*, pages 580–584, 2008.

[6] M. Ras and S. van Bussel. Web archiving user survey. Technical report, National Library of the Netherlands (Koninklijke Bibliotheek), 2007.