Summarization of Changes in Text Collections

PhD Proposal for the MAP-i Program*

Sérgio Nunes and Cristina Ribeiro
{ssn,mcr}@fe.up.pt
INESC TEC &
DEI, FEUP, University of Porto

December 18, 2012

Abstract

Information Retrieval is the Informatics field primarily focused on all problems and challenges related to information storage and access. The large majority of works in this area are based on static collections of documents. However, many of these collections are dynamic, and have evolved over time with documents being added, edited or simply removed at different times. Even in highly dynamic environments such as the World Wide Web, research tends to be centered on the most recent version of the documents and all past information is normally discarded. It is recognized that retrieval over dynamic collections introduces new and relevant research challenges. This proposal addresses this opportunity to investigate a problem that gains relevance in this context – the summarization of changes. The goal is to develop new methods and techniques that are able to automatically produce a summary given a temporal period.

1 Introduction

One of the central motivations for the initial development of the World Wide Web was to foster the sharing of information between researchers. Since its first version, launched more than 20 years ago, the World Wide Web has largely surpassed all expectations to become a global communication medium with unique characteristics. An important distinctive aspect

*This proposal was prepared for the student Manika Kar.

of modern information systems is the strong dynamic nature of the information being managed. The web is a prime example of a collection where information evolves and changes significantly over time. Several studies have observed and documented this dynamic nature of the web [5, 3, 1]. A study by Adar et al. [1] has shown that popular pages exhibit a very high change rate, with approximately 40% of the pages in the sample changing nearly every hour. In a previous study, Ntoulas et al. [5] found that, after one year, 50% of the content on the web is new, reflecting a high degree of change.

Wikipedia, the collaboratively edited encyclopedia available on the web, is a major example of an information system where data is dynamic by nature. In this case, millions of users collaborate to build and maintain a repository of millions of documents, either by adding new information or revising existing content. Online social networks, such as Facebook or Twitter, are also an example of information systems where the temporal aspect of information is central. Time is a key aspect in status updates and information sharing in these networks.

Information Retrieval is the Informatics field primarily focused on all problems and challenges related to information storage and access. The large majority of works in this area are based on static collections of documents. However, many of these collections are dynamic, and have evolved over time with documents being added, edited or simply removed at different times. Even in highly dynamic environments such as the World Wide Web, research tends to be centered on the most recent version of the documents and all past information is normally discarded.

It is recognized that retrieval over dynamic collections introduces new and relevant research challenges. Incorporating time in search and retrieval has been recently identified as one of the challenges and opportunities within the field [2]. This proposal addresses this opportunity to investigate a problem that gains relevance in this context – the summarization of changes.

2 Research Goals

This is a proposal in the area of Information Retrieval, specifically related to information access in a scenario of dynamic text collections. We think that a setting where the dynamic facet of document collections is taken into account introduces several new challenges. The primary goal of this work is to address the problem of summarizing changes in dynamic text collection. In standard text summarization, retrieval techniques are used to produce a summary of the current version of an entire document (or

collection of documents). When addressing the summarization of changes, we are interested in producing a summary that describes the alterations that were made to a document or set of documents. In other words, the focus is on the development of tools that are able to produce an automatic summary of a given period in the lifetime of a text collection. This would allow users to answer questions such as:

- What were the significant changes that have occurred in a collection of documents between two dates?
- What is the summary of the changes made to a document between two specific revisions to that document?
- Which were the important events that took place during a given period?

The summarization of changes is a relevant task when dealing with dynamic collections, such as wikis, collaborative documents, or even the collection of messages shared in a social network. Wikipedia will be one of the core resources for this work due to two major reasons. First, it is a collection where the full revision history for each document is fully available through an API. Second, it is a public resource, a critical requirement for the reproducibility of the experiments done – i.e. other researchers can access exactly the same dataset and validate or refute our findings.

Previous work on this particular problem is relatively scarce. The only explicit reference that we are aware of is by Jatowt et al. [4], introducing the idea of change summarization for document collections. However, and contrary to our idea, the information need being addressed is restricted to "recent, important changes". In this proposal we plan to adopt a broader view and define a more general information need. We consider that changes summarization is not limited to recent changes and should address any user defined period.

Our group has been actively focused on the topic of information retrieval in time-dependent collections [7]. Within this line of research, we have published work on different topics, specifically: web information retrieval [6, 8, 9], blog search [12, 11], and term weighting [14]. Two specific contributions are directly related to this proposal. The first was published at WikiSym 2008 and studies the revision patterns found in collaborative documents, specifically on Wikipedia articles [10]. The second contribution was published at DocEng 2010 and is focused on the term frequency dynamics within evolving documents [13].

References

[1] E. Adar, J. Teevan, S. T. Dumais, and J. L. Elsas. The web changes everything: understanding the dynamics of web content. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 282–291, New York, NY, USA, 2009. ACM.

[2] J. Allan, B. Croft, A. Moffat, and M. Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne. *SIGIR Forum*, 46(1):2–32, May 2012.

[3] D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener. A large-scale study of the evolution of web pages. *Softw. Pract. Exper.*, 34(2):213–237, February 2004.

[4] A. Jatowt, K. K. Bun, and M. Ishizuka. Change summarization in web collections. In *Innovations in Applied Artificial Intelligence*, Lecture Notes in Computer Science, pages 653–662. Springer Berlin / Heidelberg, 2004.

[5] A. Ntoulas, J. Cho, and C. Olston. What's new on the web?: the evolution of the web from a search engine perspective. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 1–12, New York, NY, USA, 2004. ACM Press.

[6] S. Nunes. Exploring temporal evidence in web information retrieval. In *BCS IRSG Symposium Future Directions in Information Access (FDIA 2007)*, pages 44–50, Cambridge, England, 2007. BCS IRSG, BCS IRSG.

[7] S. Nunes. *Information Retrieval on Time-Dependent Collections*. PhD thesis, Faculdade de Engenharia, Universidade do Porto, December 2010.

[8] S. Nunes, C. Ribeiro, and G. David. Using neighbors to date web documents. In *Proceedings of the 9th ACM International Workshop on Web Information and Data Management (WIDM)*, November 2007.

[9] S. Nunes, C. Ribeiro, and G. David. Use of temporal expressions in web queries. In *Proceedings of the 30th European Conference on Information Retrieval (ECIR 2008)*, pages 580–584, March 2008.

[10] S. Nunes, C. Ribeiro, and G. David. Wikichanges - exposing wikipedia revision activity. In *Proceedings of the 4th International Symposium on Wikis (WikiSym 2008)*, WikiSym. ACM, September 2008.

[11] S. Nunes, C. Ribeiro, and G. David. Feup at trec 2009 blog track: Temporal evidence in the faceted blog destillation task. In *Proceedings of the 18th Text REtrieval Conference (TREC 2009)*, TREC. NIST, November 2009.

[12] S. Nunes, C. Ribeiro, and G. David. Using temporal evidence in blog search. In *Proceedings of the ECIR'09 Workshop on Information Retrieval over Social Networks (IRSN 2009)*, April 2009.

[13] S. Nunes, C. Ribeiro, and G. David. Term frequency dynamics in collaborative articles. In *Proceedings of the 10th ACM Symposium on Document Engineering (DocEng'10)*, DocEng, pages 267–270. ACM, September 2010.

[14] S. Nunes, C. Ribeiro, and G. David. Term weighting based on document revision history. *Journal of the American Society of Information Science and Technology (JASIST)*, 62(12):2471–2478, December 2011.