The Role of Non-Content Features in XML Retrieval

Georgina Ramírez

Center for Mathematics and Computer Science (CWI)
Kruislaan, 413,1098 SJ Amsterdam, The Netherlands
phone:+31 205924319 fax: +31 205924312
Supervisor: Arjen P. de Vries
Project duration: 4 years, ending in March 2007.
georgina@cwi.nl

Abstract. The research presented investigates the use of *non-content* features for effective information retrieval. We use the expression *non-content features* to refer to the structural markup within a document or a collection and the document's surface features, i.e. document's (derived) metadata (e.g. size). Our main hypothesis is that the best use information retrieval systems can make of this type of information will be determined by the different types of search tasks and contextual factors. We focus our investigation on three main aspects: (1) The analysis of existing and the creation of new retrieval strategies on the use of non-content features, (2) the use of relevance feedback techniques to refine the non-content information given a user need, and (3) the study of the relationships between user search tasks and contextual factors and the structural characteristics of the relevant information.

1 Motivation

The growing amount of structured information available, e.g., web pages and XML documents, poses interesting new challenges to the information retrieval community. The structure of a document provides a new source of information that retrieval systems may exploit to improve their search effectiveness. New query languages have appeared that exploit structure and provide users with a more powerful tool to express complex and specific needs. Although some work has been done on the use of non-content features for retrieval, it is still an open research question how IR systems can make the best use of this type of information.

The main hypothesis of the present work is that the information that can be extracted from the structural components of documents and from other surface features can be further exploited to improve retrieval effectiveness. Furthermore, the best use IR systems can make of it will be determined by the different types of user search tasks and contextual factors.

By acquiring a better understanding of the role of non-content information and how it can help systems to perform several types of information needs, IR

systems will be able to use this extra source of information more effectively when adapting their retrieval strategies to different users, contexts and information needs.

2 Research questions

To be able to define the best use of structural information and surface features regarding search tasks, several aspects of the information seeking process on structured information need to be analyzed and understood. This research focuses on the three aspects described in the following subsections.

2.1 Non-content information: types and uses

Structured documents provide information systems with an extra source of information. A good understanding of the nature of this structure and its possible uses is needed in order to be able to choose a proper strategy when processing an information need.

Different studies on the use of structure to improve retrieval effectiveness exist. In the few years of INEX existence [8], many XML retrieval approaches have been presented (see [10, 9]). Although many of these approaches simply use standard IR techniques to rank (independently) the document's elements, several efforts have been made towards defining new retrieval models and techniques that take non-content features into account. For instance, by using structural relationships between elements to propagate or weight scores [7, 31, 1], or by weighting the content information contained in certain structural components [22, 19]. Outside the area of XML retrieval, surface features have been studied mainly in the context of web retrieval. A host of work exists that studies ways to exploit the hyperlink structure between documents. See for example [5, 17, 3]. Kraaij et al. [18] demonstrate that using information obtained from another surface feature (URL-length) can improve performance when querying for homepages. In other information retrieval areas (including XML retrieval) the most used surface feature has been document/component length, which is typically used for length normalization (e.g. [15]).

A good understanding of these techniques and an analysis of their properties is needed to determine their potential use for processing different search tasks. Furthermore, we argue that the non-content part of documents can be further exploited and that this type of information can be used more effectively. In particular, we investigate the following research questions:

- What are the main characteristics of the existing retrieval strategies that use non-content features and what are they good for?
- Can we define new retrieval strategies that exploit this type of information more effectively?

Some work towards answering these questions has been presented in [28] and [25], where we use structural relationships between XML elements to improve

retrieval effectiveness. This approach has been shown to be effective in different search tasks and in combination with other XML retrieval techniques.

2.2 Structural relevance feedback

As the complexity of information needs increases, systems need to be able to process any information users might provide, e.g, by using special interfaces or by relevance feedback strategies.

For many years now, relevance feedback techniques have been used in IR systems. Although extensively used, these techniques focus uniquely on the content part of a document. A survey of these techniques applied to different information retrieval models is presented in [29]. In the area of XML retrieval, existing relevance feedback algorithms have been applied to query on XML documents (e.g. [20], [31]). However, its use has also concentrated on the content part of the documents.

Our hypothesis is that, in the same way as content is refined on a relevance feedback process, relevant structural information and surface features can also be used to update search parameters and to refine the structural characteristics of the desired information. In particular, we investigate the following research questions:

- Which type of structural information and surface features can be extracted from a relevance feedback process?
- How can we use relevant non-content features to update search parameters and improve the overall effectiveness of the retrieval system?

Working in this direction, we introduced in [21] the concept of structural feedback and used three different surface features to model *structural relevance*. Although this first attempt was not very successful, further analysis showed that the use of structural information on a feedback process can improve retrieval effectiveness significantly ([27, 24]). Later, Schenkel and Theobald [30] showed that when the structural information is used in combination with the content one, larger improvements can be achieved.

2.3 Search tasks and context in XML retrieval

Because the aim of any information system is to be able to answer effectively different types of search tasks and information needs, it is important to understand the nature of these tasks and the different contextual factors that might influence the search.

Although many studies have been done on understanding and modeling user needs and information seeking behavior within the information science community, traditional information retrieval systems, with few exceptions (e.g.[11], [2]), pretty much ignored the user. However, there is a growing interest within the IR community to incorporate the use of contextual information to improve effectiveness. Understanding and modeling contextual information is becoming an

important issue (see for instance [13, 12]). In [14] Ingwersen and Järvelin analyze the work done on these areas and propose new directions towards the integration of information seeking and retrieval in context research. Besides that, some work has been done in the IR community to classify types of user needs and intentions [4, 6] and to improve retrieval effectiveness by using different retrieval strategies for each of these categorizations [16].

Our research concentrates on a very specific aspect of contextual information retrieval: The influence of search task type and several contextual factors on the structural characteristics of relevant information. We argue that if there are differences in the distribution of structural features on the relevant information regarding different search tasks or context situations, retrieval systems should be able to use this information more effectively.

In particular, we study four different contextual features that we considered to be possibly relevant in the XML information retrieval setting. Precisely, the knowledge users have on the topic of the request, the type and specificity of the request, user's intention or work-task behind the request and user's familiarity with the document's structure. The latter, specific for this retrieval setting, is introduced with the hypothesis that the knowledge users have about the structure of the documents might lead to the use of different retrieval strategies. Thus, it might be an important contextual factor to consider.

In general, two main research questions are investigated:

- Can we identify a measurable dependency between a topic's task type and some of its contextual factors and the structural aspects of the topic's relevant components?
- Can we correlate, according to their properties, different search tasks and contextual factors with the different uses of non-content features?

So far, we have concentrated in the first research question and showed that the distributions of some structural features differ for different search tasks and contextual information [23, 26]. Although some contextual factors such as the user's familiarity with the topic or the specificity of the request might influence the relevance judgments, the differences are rather small and it is not clear yet whether IR systems will be able to make use of this type of information.

3 Conclusions

This paper summarizes the main motivations, hypotheses and research questions of my research proposal. The main contribution of this work is the investigation of the use of non-content features (structural information and surface features) for effective information retrieval and relevance feedback on XML documents collections given different search tasks and contexts.

References

1. Paavo Arvola, Marko Junkkari, and Jaana Kekäläinen. Generalized Contextualization Method for XML Information Retrieval. In *CIKM '05: Proceedings of the*

14th ACM International Conference on Information and Knowledge Management, pages 20–27, New York, NY, USA, 2005. ACM Press.
2. N.J. Belkin, R.N. Oddy, and H.M. Brooks. Ask for information retrieval: Part I and II. *Journal of Documentation*, 38(2 and 3), 1982.
3. K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 104–111, 1998.
4. S. K. Bhavnani, K. Drabenstott, and D. Radev. Towards a unified framework of ir tasks and strategies. In *Proceedings of ASIST'2001*, pages 340–354, 2001.
5. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.
6. A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.
7. N. Fuhr and K. Großjohann. XIRQL: A Query Language for Information Retrieval in XML Documents. In W. B. Croft, D. Harper, D. H. Kraft, and J. Zobel , editors, *Proceedings of the 24th Annual International Conference on Research and development in Information Retrieval*, pages 172–180. ACM, 2001.
8. Norbert Fuhr, Norbert Gövert, Gabriella Kazai, and Mounia Lalmas. INEX: INitiative for the Evaluation of XML Retrieval. In *Proceedings of the SIGIR 2002 Workshop on XML and Information Retrieval*, 2002.
9. Norbert Fuhr, Mounia Lalmas, Saadia Malik, and Gabriella Kazai, editors. *Advances in XML Information Retrieval. Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005)*, volume 3977 of *Lecture Notes in Computer Science*. Springer, 2006.
10. Norbert Fuhr, Mounia Lalmas, Saadia Malik, and Zoltán Szlávik, editors. *Advances in XML Information Retrieval. Third International Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2004)*, volume 3493 of *Lecture Notes in Computer Science*. Springer, 2005.
11. P. Ingwersen. *Information Retrieval Interaction*. London: Taylor Graham, 1992.
12. P. Ingwersen, K. Järvelin, N. Belkin, and B. Larsen, editors. *ACM SIGIR 2005 Workshop on Information Retrieval in Context (IRiX)*, 2005. `http://irix.umiacs.umd.edu/ACM-SIGIR2005-IRiX-proceedings.pdf`.
13. P. Ingwersen, K. van Rijsbergen, N. Belkin, and B. Larsen, editors. *ACM SIGIR 2004 Workshop on Information Retrieval in Context (IRiX)*, 2004. `http://ir.dcs.gla.ac.uk/context/IRinContext_WorkshopNotes_SIGIR2004.pdf%`.
14. Peter Ingwersen and Kalervo Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*, volume 18 of *The Information Retrieval Series*. Springer, 2005.
15. Jaap Kamps, Maarten de Rijke, and Börkur Sigurbjörnsson. Length Normalization in XML Retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 80–87. ACM Press, 2004.
16. I. Kang and G. Kim. Query type classification for web document retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64–71. ACM Press, 2003.
17. Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, 1999.
18. Wessel Kraaij, Thijs Westerveld, and Djoerd Hiemstra. The Importance of Prior Probabilities for Entry Page Search. In *Proceedings of the 25th Annual International ACM SIGIR C onference on Research and Development in Information Retrieval*, pages 27–34. ACM Press, 2002.

19. Wei Lu, Stephen Robertson, and Andrew Mcfarlane. Field-Weighted XML Retrieval Based on BM25 . In N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, editors, *INEX 2005 Workshop Proceedings*, Dagstuhl, Germany, 2005.
20. Yosi Mass and Matan Mandelbrod. Relevance Feedback for XML Retrieval. In Norbert Fuhr, Mounia Lalmas, Saadia Malik, and Zoltán Szlávik, editors, *Advances in XML Information Retrieval. Third International Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2004)*, volume 3493, pages 303–310. Springer, 2005.
21. Vojkan Mihajlović, Georgina Ramírez, Arjen P. de Vries, , Djoerd Hiemstra, and Henk Ernst Blok. TIJAH at INEX 2004. Modeling Phrases and Relevance Feedback. In Norbert Fuhr, Mounia Lalmas, Saadia Malik, and Zoltán Szlávik, editors, *Advances in XML Information Retrieval. Third International Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2004)*, volume 3493, pages 276–291. Springer, 2005.
22. P. Ogilvie and J. Callan. Using language models for flat text queries in XML retrieval. In Norbert Fuhr, Saadia Malik, and Mounia Lalmas, editors, *INEX 2003 Workshop Proceedings*, 2003.
23. G. Ramírez and A. P. de Vries. XML and Context: Structural Features Relevant to Search Tasks. In *Proceedings of the ACM SIGIR 2005 Workshop on Information Retrieval in Context, IRiX*, pages 24–26, Salvador, Brazil, 2005.
24. G. Ramírez, T. Westerveld, and A.P. de Vries. Structural Features in Content Oriented XML Retrieval. Technical Report INS-E0508, CWI,Centre for Mathematics and Computer Science, 2005.
25. G. Ramírez, T. Westerveld, and A.P. de Vries. Using Structural Relationships for Focused XML Retrieval. In *Proceedings of the Seventh International Conference on Flexible Query Answering Systems (FQAS 2006)*. Springer, 2006.
26. Georgina Ramírez and Arjen P. de Vries. Relevant Contextual Features in XML Retrieval. In *1st Symposium on Information Interaction in Context (IIiX)*, Copenhagen, Denmark, 2006.
27. Georgina Ramírez, Thijs Westerveld, and Arjen P. de Vries. Structural Features in Content Oriented XML Retrieval. In *CIKM '05: Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 291–292, New York, NY, USA, 2005. ACM Press.
28. Georgina Ramírez, Thijs Westerveld, and Arjen P. de Vries. Using Small XML Elements to Support Relevance. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 2006.
29. Ian Ruthven and Mounia Lalmas. A Survey on the Use of Relevance Feedback for Information Access Systems. *Knowl. Eng. Rev.*, 18(2):95–145, 2003.
30. Ralf Schenkel and Martin Theobald. Structural Feedback for Keyword-Based XML Retrieval. In *Advances in Information Retrieval. 28th European Conference on IR Research (ECIR 2006)*, pages 326–337, 2006.
31. B. Sigurbjörnsson, J. Kamps, and M. de Rijke. An element-based approach to XML retrieval. In N. Fuhr, M. Lalmas, and S. Malik, editors, *Proceedings of the Second Workshop of the INitiative for the Evaluation of XML retrieval (INEX)*, ERCIM Publications, 2004.