# User Needs and Strategies in Structured Information Retrieval

Georgina Ramírez

CWI
P.O. Box 94079
1090GB Amsterdam
The Netherlands
georgina@cwi.nl

**Abstract.** Structured information retrieval studies the combination of the content and the structure information of documents to perform different IR tasks. Different approaches such as *structure* or *context weighting* (e.g. [24], [10]) make use of the structural information of documents to improve information retrieval effectiveness. However, most of these studies do not take the user into account and they use the same strategy to perform all types of queries. This work aims to identify the relationship between user task and strategies on the usage of structural information when performing that task. The theoretical part of this research consists of three main phases designed to acquire a good understanding of (1) the nature of user tasks on structured documents, (2) the types of structural information and its role in retrieval strategies, and (3) the formalization of a model to correlate both, user tasks and strategies. Two different experimental studies are planned. The first one investigates the combination of evidence from different strategies for the defined user tasks, and the second one investigates how to use relevance feedback techniques to refine the structural information for a given user need.

## 1 Motivation

The growing amount of structured information available, e.g., web pages and XML documents, poses interesting new challenges to different information seeking research communities like digital libraries or information retrieval. On one hand, the structure of the documents provide a new source of information that retrieval systems may exploit to improve their search effectiveness. On the other hand, the appearance of new query languages that work on structure provides the users with a more powerful tool to express complex and specific needs.

This possibility of querying using structural constraints requires itself an understanding of two main issues that differ from traditional IR. On one hand, due to differences in the structure of the documents, some extra knowledge (e.g., semantic mapping) might be needed when querying on heterogeneous sources. On the other hand, the knowledge that the user has of the structure of the documents, might lead to different interpretations of the structural constraints of the query.

Surface features, i.e., anything other than content information, and, in particular, structural information, i.e, structural markup within a document or a collection, have been poorly used by information retrieval systems. Although some work has been done on using structural information to address some of the issues from above, how to choose and use the right surface features for a given user task is still an open research question.

Furthermore, although there is a growing interest in the information retrieval field in using user context information to improve retrieval effectiveness, still a gap exists between the advance done in information seeking research regarding user tasks and behaviour and the work done to apply these studies in the information retrieval systems [15]. This work investigates this relationship in the context of structured information retrieval.

On the other hand, studies done in other areas of information retrieval, like web search, have shown that the use of more specialized retrieval strategies that combine different sources of evidence for different categorizations of user tasks and intentions, improves system effectiveness significantly (e.g. [14]).

The main hypothesis of the present work is that the different types of information that can be extracted from the structural components of documents and from other surface features can be treated as multiple sources of information. The best combination of evidence from these sources will be determined by the different types of user tasks and intentions, as it happens already in different information retrieval tasks. Therefore, once we acquire a better understanding of surface features and a study on how they can help the system to perform several types of information needs, information retrieval systems will be able to use this extra source of information more effectively.

## 2 Research plan

Several aspects of the information seeking process on structured information need to be analyzed and understood to be able to develop the work presented. The five main steps of this research plan are:

1. **Definition of a taxonomy of user tasks.** Because the aim of any information system is to be able to answer effectively different user needs, it is important to understand the nature of these search tasks. The main hypothesis on this part of the research is that the type of search tasks users perform on structured collections are the same as those done on plain text collections. The difference between the two fields is that, when querying on structured data, users can provide the system with extra information about the type and location of the information they are looking for and that the knowledge the users have of the structure of the documents might differ in different degrees. Therefore, two main research questions are investigated:
   - Can we apply existing categorization of user tasks to structured information retrieval?

- How can we extend these categorizations to include the different levels of knowledge a user might have on the structure of the documents?

2. **Definition of a taxonomy of types and uses of structure.** Sructured 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 make an optimal use of this information. The main hypotheses here are that different types of structure have been created with different purposes. Therefore, the use that information retrieval systems can make of the structure can be also clustered into different views (dimensions), e.g., semantic or organizational. Different research questions are investigated in this part of the research:
   - Can we define a fixed set of independent dimensions from the structural features of the documents?
   - Can we classify the existing strategies into these set of dimensions?
   - Are there other strategies systems could use regarding structure?

3. **Formalization of a model to correlate user tasks and strategies.** Once the classifications of user tasks and possible strategies are defined, we need to connect both of them and formalize a prescriptive model that defines which strategies for the possible dimensions can be used by the retrieval system for each of the user tasks. The hypothesis here is that different user needs will require different structural features. For example, a query from a user searching for pictures might require to be processed strictly regarding the type of retrieval unit (pictures) and some semantic matching could be added (e.g., pictures or drawings or figures). However, this query most probably will not require any length normalization. The research question we investigate in this part is:
   - Can we correlate, according to their properties, different user needs with the different uses of structural features?

4. **Experimentation on the combination of evidence from different dimensions.** Different experiments need to be done in order to estimate the best way to combine the evidence from the different dimensions used in the different user tasks. A hypothesis in this part of the research is that the problem we address of combination of evidence is similar to the one of combining evidence from different classifiers [8]. We investigate the following research questions:
   - Can we cast the problem of combining the different dimensions to the one of combining classifiers?
   - Can we estimate the best parametrization for the combination of evidence dependent on the user task?

5. **Experimentation on the use of structural relevance feedback.** As the complexity of an information need increases, systems need to be able to process any information users might provide, e.g, by using special interfaces

or by relevance feedback strategies. The hypothesis in this experimental part of the research is that, in the same way as content is redefined on a relevance feedback process, relevant structural information can also be used to update the parameters of the combination of evidence during different search tasks. We investigate the following research questions:

- Which type of structural information can be extracted from a relevance feedback process?
- How can we use the relevant structural information to update the parametrization of the combination of evidence and improve the overall effectiveness of the retrieval system?

## 3 Experiments and evaluation

Once the model is defined, it will be implemented on top of an IR system that uses a probabilistic language modeling approach to information retrieval [20]. Different benchmarks will be used to test the effectiveness of the approach. The user tasks from several tracks in these benchmarks will be manually classified into the defined task categorization and processed with the system. The different tracks have been chosen for providing different types of information needs and different types of structured data. The main characterisitics of these tracks are highlighted in Table 1:

**Table 1.** Track's characteristics

| **Testbed and Track** | **Features** |
|---|---|
| INEX: Adhoc | Content oriented and content and structure queries. |
| INEX: Heterogeneous | Different types of user tasks on an heterogeneous collection. |
| INEX: Multimedia | Search task on a collection with different multimedia data. |
| INEX: Relevance Feedback | Adhoc task with relevance feedback. |
| TREC: Enterprise search | Different user tasks on different types of resources. |

## 4 Background and Related Work

A review on theoretical models and frameworks for information seeking and retrieval is presented in [16]. Although many studies have been done on understanding and modelling user needs and information seeking behaviour within the information science community, traditional information retrieval systems pretty much ignore the user. However, several efforts have been made towards defining information retrieval models and systems that take users and context into account (e.g. [11], [2]), and a growing interest within the IR community on these user aspects is leading to an increasing ammount of studies on user tasks and seeking behaviours. Some of these recent studies can be found in [12] and some of the specific work that inspired the research presented in this paper is describe in the following paragraphs.

Bhavnani et al. [4] define a framework for IR tasks and strategies with the goal of training users on t he use of effective strategies for information searches. They also analyse different studies on the categorization of tasks and IR strategies. Broder [7] introduces and analyses a taxonomy of web searches and shows how search engines evolve towards dealing better with these web-specific needs. Croft [8] analyses some work done in the area of combining evidence in information retrieval and shows how this problem can be modeled as the one of combining the outputs of multiple classifiers. Kang et. al [14] define another taxonomy for web searches and report improved effectiveness when using different strategies for each of the categorized web searches. In [5], the same taxonomy is extended to represent other information retrieval tasks. Although the authors do not show significant improvements yet, they believe that if different types of queries (user intentions) are defined and processed accordingly, the retrieval effectiveness may increase.

As mentioned earlier, different studies on the use of structure to improve retrieval effectiveness exist. In the few years of INEX existence [9], a lot of XML retrieval approaches have been presented. Studies have been done on the use of the structural relationship between elements [10, 26], mentioned previously as *context weighting*. Another group of researchers uses *structure weighting* in order to give importance to certain types of elements [24]. To address the problem of what is the best retrieval unit for each query, some works choose the approach of defining a subset of possible elements to be retrieved [27, 21] whereas some others, as in other information retrieval areas, use length normalisation [13]. Work done by the author on this area includes the study and implementation of some of these strategies [19] [20].

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 [6, 17, 3]. Kraaij et al. [18] demonstrate that using information obtained from URL-length can improve performance when querying for homepages. Also in the HARD track at TREC [1], surface features are studied, but there the features describe characteristics of the searcher rather than the documents. In other information retrieval areas the only surface feature that has been used widely is document length, which is typically used for normalisation.

Different 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 [25]. In [22] and [26], existing relevance feedback algorithms are applied to query on XML documents. All these approaches are using content-oriented feedback, whereas the feedback this research intends to study is based on structural features. Some work that the author has done in this area is explained in [23], where a first approach on using different structural features on a relevance feedback process is presented.

## 5 Conclusions

In this paper, I have summarized the main motivations, hypotheses and research questions of my research proposal. I conclude with the main points where the presented research might contribute to the information seeking community and in particular, to the IR field:

1. Investigation on the use of structural features for effective information retrieval and relevance feedback.
2. Definition of the set of features that an information system should support in order to use structural information effectively.
3. Study and development of an experimental approach to correlate user needs and strategies for structured information retrieval.

## References

1. J. Allan. HARD track overview in TREC 2003; high accuracy retrieval from documents. In *The Twelfth Text Retrieval Conference*, 2003.
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. M. Bomhoff, T. Huibers, and P. van der Vet. User intentions in information retrieval. In *Proceedings of the Fifth Dutch-Belgian Information Retrieval Workshop*, 2005.
6. S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.
7. A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.
8. W. B. Croft. *Combining approaches to information retrieval*, chapter 1, pages 1–36. Kluwer Academic Publishers, 2000.
9. N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas. INEX: INitiative for the Evaluation of XML retrieval. In *Proceedings of the SIGIR 2002 Workshop on XML and Information Retrieval*, 2002.
10. N. Fuhr and K. Großjohann. Xirql: A query language for information retrieval in xml documents. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 172–180, 2001.

11. P. Ingwersen. *Information Retrieval Interaction*. London: Taylor Graham, 1992.
12. P. Ingwersen, K. van Rijsbergen, N. Belkin, and B. Larsen, editors. *IRiX: ACM SIGIR 2004 Workshop on Information Retrieval in Context*, 2004. `http://ir.dcs.gla.ac.uk/context/IRinContext_WorkshopNotes_SIGIR2004.pdf%`.
13. J. Kamps, M. de Rijke, and B. 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.
14. 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.
15. K.Järvelin and P.Ingwersen. Information seeking research needs extension towards tasks and technology. In *Information Research, 10(1) paper 212*. 2004. `http://InformationR.net/ir/10-1/paper212.html`.
16. K.Järvelin and T.D. Wilson. On conceptual models for information seeking and retrieval research. In *Information Research, 9(1) paper 163*. 2003. `http://InformationR.net/ir/9-1/paper163.html`.
17. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
18. W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34. ACM Press, 2002.
19. J. List, V. Mihajlovic, A. P. de Vries, G. Ramirez, and D. Hiemstra. The tijah xml-ir system at inex 2003. In Norbert Fuhr, Saadia Malik, and Mounia Lalmas, editors, *INEX 2003 Workshop Proceedings*, pages 102–109, 2003. `http://inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf`.
20. J. List, V.Mihajlovic, G.Ramirez, A.P. de Vries, D. Hiemstra, and H.E. Blok. Tijah: Embracing ir methods in xml databases. *Information Retrieval Journal*.
21. Y. Mass and M. Mandekbrod. Retrieving the most relevant xml components. In Norbert Fuhr, Saadia Malik, and Mounia Lalmas, editors, *INEX 2003 Workshop Proceedings*, 2003. `http://inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf`.
22. Y. Mass and M. Mandelbrod. Relevance feedback for xml retrieval. In Norbert Fuhr, Saadia Malik, and Mounia Lalmas, editors, *INEX 2004 Workshop Proceedings*, 2004. notebook paper.
23. V. Mihajlovic, G. Ramirez, A.P. de Vries, , D. Hiemstra, and H.E. Blok. Tijah at inex 2004. modeling phrases and relevance feedback. In Norbert Fuhr, Saadia Malik, and Mounia Lalmas, editors, *INEX 2004 Workshop Proceedings*, 2004. notebook paper.
24. 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. `http://inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf`.
25. I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, 18(2):95–145, 2003.
26. 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.

27. A. Trotman and R. A. O'Keefe. Identifying and ranking relevant document elements. In Norbert Fuhr, Saadia Malik, and Mounia Lalmas, editors, *INEX 2003 Workshop Proceedings*, 2003. `http://inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf`.