Using a Mediated Query Approach for Matching Unstructured Query with Structured Resources

Gan Keng Hoon
Faculty of Computer Science and Information Technology
Universiti Malaya
50603 Kuala Lumpur, Malaysia
khgan@cs.usm.my

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – query formulation, *retrieval models.*

General Terms
Algorithms, Design, Experimentation.

Keywords
Query, Search, XML Retrieval.

Recent years, the availability of public accessible structured resources like XML on the web has led to active developments of structural retrieval systems. With these systems, users will be able to query for information from structured resources on the web efficiently. When querying, it is obvious that usage of structural information in query increases the precision of retrieval system. However, general web users are more familiar with unstructured query such as natural language or keywords, which contains no structural information. Therefore, a retrieval method that can handle matching between a simpler form of query (e.g. natural language, keywords, or less rigid syntax), and structured resources is required.

The needs to balance between effective querying models and exploitation of structural information of resources has spurs numerous solutions in the area structured retrieval. We classify the solutions based on the querying interfaces used.

i) Syntax-based querying, e.g. [4].
ii) Form-based querying, e.g. [5].
iii) Fragment-based querying, e.g. [1].
iv) Keyword-based querying, e.g. [2].

From the four querying interfaces mentioned above, we can see that there are basically two ways of specifying information needs in a query, either includes structural information in the query (i, ii, iii), or keywords only (iv). If we compared the information available in these two queries, it is quite obvious that usage of structural information in query can increase the precision of a retrieval system theoretically. But users need to explicitly add structural information, which means that they need to either know available structural information (i), or guess (iii), or select from a list (ii). A related work [3] presents that users have problem remembering how to write the syntax, do not have knowledge about or hard to guess the underlying structure, which result in no significant improvement of retrieval quality despite the usage of structural clues in query. On the other hand, keyword-based querying (iv) would be more user friendly, but since structural information is not explicitly specified in the query, retrieval system would not be able to utilize such information to further improve the precision of its result.

This motivates us to find a retrieval method that supports querying which is simpler and familiar to user, i.e. unstructured query, but at the same time, does not overlook the usage of structural information in query. For example, specifying query using keywords only (e.g. CO topics [2]) or loose structural clues resemble concepts such as "person: Tom Mitchell, course: machine learning". But, as structural information is not explicitly specified in the query, the system has to intelligently predict which structure to look up by optimizing query with sufficient structural information to improve the precision of retrieval model.

Hence, we propose a solution that automatically adds structural information to the unstructured query, and represents it as a *Mediated Query*. The mediated query is an intermediate query in structured form to bridge the gap of structural differences between unstructured query and structured resources. As the selection of correct structural information that reflects the query context is crucial for better retrieval performance, we develop a method to obtain this information by learning the semantics of a set of terms extracted from structured resources. The semantics of a term is defined by its concept and context. We represent the term and its semantics using the *Semantic Prediction Model*. The model will be used in reasoning the context of query and the process of creating mediated query. The mediated query is then matched against structured resources to obtain relevant results.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

References
[1] Carmel, D., Maarek, Y. S., Mandelbrod, M., Mass, Y. and Soffer, A. Searching XML documents via XML fragments. In *Proceedings of SIGIR'03* (Toronto, Canada, 28 Jul – 1 Aug, 2003). 151-158.

[2] Fuhr, N., Gövert, N., Kazai, G. and Lalmas, M. INEX: Initiative for the Evaluation of XML Retrieval, In *Proceedings of the ACM SIGIR Workshop on XML and Information Retrieval* (Tampere, Finland, August 15 2002).

[3] Trotman, A. and Lalmas, M. Why Structural Hints in Queries do not Help XMl-Retrieval. In *Proceedings of SIGIR'06* (Seattle, Washington, 6–11 Aug, 2006.

[4] Trotman, A. and Sigurbjörnsso, B. Narrowed Extended XPath I (NEXI). *Advances in XML Information Retrieval*. (May 2005). LNCS 3493, Springer-Verlag, 16–40.

[5] van Zwol, R., Bass, J., van Oostendorp, H. and Wiering, F. Bricks: The Building Blocks to Tackle Query Formulation in Structured Document Retrieval. In *Proceedings of ECIR '06*. LNCS 3936, Springer-Verlag Berlin Heidelberg, 314-325.