XML based Keyword Search

Abstract

The success of information retrieval style keyword search on the web leads to the emergence of XML based keyword search. The text database and XML database differences leads to three new challenges: 1) The users search intention is to be identified, i.e., the XML node types that user wants to search for and search via is identified. 2) The similarities in tag name, tag value and the structure of tags are identified. 3) New scoring function is needed to estimate the output of the search results (XML document) relevance to the given query. However, these challenges cannot be addressed by the existing system, which results in low quality results in terms of query relevance. In this paper, an IR-style approach is proposed which basically utilizes the statistics of underlying XML data to address these challenges. First, specific guidelines that a search engine should meet in both search intention identification and relevance oriented ranking for search results is proposed. Then, based on these guidelines, a novel XML TF*IDF ranking strategy to rank the individual matches of all possible search intentions is proposed.

References

- Yunyao Li, Cong Yu, H. V. Jagadish, "Schema-Free XQuery"; Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004
- Norbert Fuhr, Kai Großjohann, "XIRQL: A Query Language for Information Retrieval in XML Documents SIGIR;01, September 9–12, New Orleans, Louisiana, USA

**Index Terms**

Computer Science

Web Technology

**Keywords**

XML Search ranking.