

NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

PH.D DEFENCE - PUBLIC SEMINAR

Title: **Enhancing the Usability of XML Keyword Search**

Speaker: Mr Zeng Yong

Date/Time: 19 September 2014, Friday, 02:00 PM to 03:30 PM

Venue: MR1, COM1-03-19

Supervisor : Dr Ling Tok Wang, Professor, School of Computing

Abstract:

XML has become a de facto standard of information representation and ex-change over the Internet. It has been used extensively in many applications. Such semi-structured data is normally queried by rigorous structured query languages, e.g., XPath, XQuery, etc. In recent years, keyword search on XML has become more and more popular due to its easy-to-use query interface. It provides an opportunity to explore the semi-structured data without knowing the data schema or learning the sophisticated structured query languages. It is becoming an equally important counterpart of structured query and an important way for novice to explore XML database.

XML keyword search has been abundantly studied in the last ten years. The research efforts mainly focus on defining what should be returned as results (matching semantics) and designing efficient algorithms for a certain matching semantics.

However, in XML keyword search, how to reduce the gap between users' search intention and the query results remains a challenge. Even for the mature web search, users have to reformulate and resubmit their queries 40% to 52% of the time in order to get what they want. Therefore, enhancing the usability by handling the mismatch between users' search intention and the query results is an important issue, no matter for web search, XML keyword search, or any other kind of search. In this dissertation, we will study how to enhance the usability of XML keyword search by addressing the following challenges.

First, we study the mismatch results in XML keyword search without considering ID references. In this case, the XML data can be modeled as a tree. We develop a low-cost post-processing algorithm on the results of query evaluation to detect the mismatch and generate helpful suggestions to users. The solution is based on two novel concepts that we introduce: Target Node Type and Distinguishability. Target Node Type represents the type of node a query result intends to match, and distinguishability is used to measure the importance of the query keywords. Our solution can work with any LCA-based matching semantics and is

orthogonal to the choice of result retrieval method adopted. We have also built an interactive XML keyword search engine, called XClear, with our mismatch solution incorporated.

Second, we try to extend our mismatch solution to XML data with ID references considered. Then the XML data is usually modeled as a digraph, where keyword query results are usually computed by graph traversal. We call such a digraph as XML IDREF digraph in this dissertation. We observe that an XML IDREF digraph is mainly a tree structure with a portion of reference edges. It motivates us to propose a novel method to transform an XML IDREF digraph with ID references to a tree model, such that we can exploit abundant efficient XML tree search methods. Subsequently our mismatch solution designed for an XML tree can still apply.

Third, after the results are retrieved from the search engine, they need to be presented to users. To further bridge the mismatch gap between users' search intention and the query results, we improve the result presentation method for XML keyword search, which plays an important role in users' digesting and exploring of the query results. The traditional way of returning a list of subtrees as query results is insufficient to meet the information needs of users. We find that such a presentation is imprecise and could be misleading. Users could misunderstand the query results. Therefore We propose an interactive and novel result presentation model, call XMAP, to visualize and work as a complementary component of the XML keyword search engine, in order to enhance the usability of XML keyword search. It allows users to view the inter-relationship among the query results and also further explore the query results according to their information needs. A demo system of XMAP has also been built.