NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

PH.D DEFENCE - PUBLIC SEMINAR

Title: **Semantics analysis for XML keyword search**

Speaker: Ms Le Thuy Ngoc

Date/Time: 14 November 2014, Friday, 02:00 PM to 03:30 PM

Venue: MR3, COM2-02-26

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

Abstract:

Since XML has become a standard for information exchange over the Internet, more and more data are represented as XML. XML keyword search has been attracted a lot of interests because it provides a simple and user-friendly interface to query XML documents. Existing approaches for XML keyword search can be classified into two types: tree-based approaches and graph-based approaches based on whether the considered XML document is modeled as a tree or a graph. Commonly, the tree-based approaches are for XML documents with no ID/IDREF and mainly follow the Lowest Common Ancestor (LCA) semantics (and thus they are also called LCA-based approaches), while the graph-based approaches are for XML documents with ID/IDREFs and usually apply the Steiner tree semantics. These tree-based and graph-based approaches work well for certain types of XML documents. However, since these approaches only rely on the structure of XML documents but do not consider the semantics of Objects, Relationships between/among objects, Attributes of objects, and Attribute of relationships (referred to as ORA-semantics), they may suffer from several problems, including meaningless answers, missing answers, duplicated answers, schema-dependent answers (i.e., different answers are returned for different schema designs of the same data content), and incomplete answers (when handling relationship attributes or n-ary ($n >= 3$) relationship types).

In this thesis, we propose to use the ORA-semantics for keyword search on a data-centric XML document to address the above problems. We classify nodes in a data-centric XML document into different types such as object class, object identifier (OID), object attribute, relationship attribute, etc. The ORA-semantics provides the type of each node in XML data. Based on the ORA-semantics, we can first distinguish an object node from an arbitrary node in XML data, e.g., attribute and value. Then we can detect whether the two object nodes refer to the same object based on object class and OID. These identifications enable us to have the following contributions.

First, we find that the LCA-based approaches (i.e., the tree-based approaches) only search up the XML tree from the matching nodes to find common ancestors but not search down the XML tree to find common information appearing as descendants (referred to ascommon descendants) due to many-to-many or many-to-one relationships among objects. Therefore, they can miss meaningful answers. We propose the new semantics, called Nearest Common Object Node (NCON), to take not only common ancestors but also common descendants into account. We then propose an approach using reversal mechanism to find NCONs for a keyword query over data-centric XML document with no ID/IDREF. Our approach is also able to avoid meaningless answers, duplicated answers and incomplete answer.

Second, we extend the NCON semantics for XML documents with ID/IDREFs, in which some or all objects are under ID/IDREF mechanism. This means objects with duplication and objects with ID/IDREFs can be co-existed in the considered XML documents. The challenge of finding NCONs from such XML documents is that they cannot be modeled as trees anymore. They are graph instead. However, searching over a graph has been known to be equivalent to the group Steiner tree problem, which is NP-Hard. To address this challenge, we discover that an XML graph still has hierarchical structure where a reference edge can be considered as a parent-child relationship, in which the parent is the referring node and the child is the referred node. The hierarchical structure of XML graph provides us an efficient algorithm to find NCONs for keyword queries over XML graph.

Third, not only common ancestors and common descendants provide meaningful answers for users, we discover thatcommon relatives of the matching nodes, which are common ancestors w.r.t. some other schemas, are also meaningful. Therefore, we propose the CR (Common Relative) semantics which includes all together common ancestors, common descendants and common relatives as answers. More interestingly, several XML documents can share the same content such as they are all transformed from the same relational database by picking up different entity as the root. The proposed CR semantics can return the same answers for different XML documents (in which objects with duplication and object with IDREFs can be co-existed) sharing the same data content. This is important because when users issue a keyword query, they often have some intention in mind about what they want to search for. Thus, for a query, they expect to have the same answers from different XML documents sharing the same content. However, for existing appros, for the same data content, different schema designs may provide different answers for the same query.

Finally, we study how to support group-by and aggregate functions in XML keyword search. It goes beyond the simple keyword query, and raises several challenges including: (1) how to address the keyword ambiguity problem when interpreting a keyword query; (2) how to identify duplicated objects and duplicated relationships in order to guarantee the correctness of the results of aggregate functions; (3) how to compute a keyword query with group-by and aggregate functions. We exploit the ORA-semantics to address the above challenges. We find that without the ORA-semantics, keyword search with group-by and aggregate functions cannot be processed correctly. We consider only XML documents with no ID/IDREF. We leave the case of XML documents with ID/IDREFs for future works.

After all, this thesis theoretically and experimentally demonstrates that using ORA-semantics to process XML keyword queries one can gain a lot of benefit in terms of both effectiveness and efficiency. This result is useful for future research and applications in XML keyword search.