NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

PH.D DEFENCE - PUBLIC SEMINAR

Title: **Towards Effective Relational Keyword Search Using Semantics**

Speaker: Mr Zeng Zhong

Date/Time: 7 April 2016, Thursday, 10:00 AM to 11:30 AM

Venue: Executive Classroom, COM2-04-02

Supervisor : Dr Lee Mong Li, Janice, Professor, School of Computing
Dr Ling Tok Wang, Professor, School of Computing

Abstract:

Keyword search over relational databases has been widely studied in recent years. In contrast to structured queries, it enables users to pose queries without learning query languages or database schemas, and has become a flexible and popular approach to access database information. However, the interpretation of a keyword query is ambiguous. Existing research on relational keyword search has been focused on the efficient computation of search results and ranking strategies to improve the quality of results. But they do not consider the Object-Relationship- Attribute (ORA) semantics in the database, and thus suffer from the problems of returning incomplete answers, overwhelming number of answers, and even incorrect answers. In addition, they do not consider the normal forms of the database relations, and return different answers for different schemas of the same data content. Hence, users may fail to find the answers that satisfy their search intention. In this thesis, we address these problems by exploiting the ORA semantics in relational databases for keyword query processing in order to improve the completeness and correctness of keyword search.

First, we analyze the semantics of the relations in a database. We classify the relations into object relations, relationship relations, mixed relations, and component relations. An object (relationship resp.) relation captures the information of objects (relationships resp.), while a mixed relation captures the information of both objects and their associated many-to-one relationships. The information of multivalued attributes of objects and relationships are captured by component relations. We refer to these semantics as the Object-Relationship-Attribute (ORA) semantics. We construct an Object-Relationship-Mixed (ORM) data graph where each node represents either an object, or a relationship, or an object together with its many-to-one relationship in the database. Keyword queries are processed via the ORM data graph because the information of objects and relationships in the ORM data graph enable us to retrieve more complete and informative answers compared to existing methods.

Second, we investigate how objects in a relational database are related via relationships. We identify four types of paths, namely, simple path, recursive path, palindrome path and complex path, whereby a pair of nodes in the ORM data graph can be connected. These paths capture the semantic meanings between objects together with relationships in the database, and reflect different query interpretations. We compute and rank query answers by considering the semantic paths between each pair of keyword match nodes in the answers. Compared to existing ranking schemes which typically rank answers based on the number of tuples, our approach ranks answers based on the semantic paths and is more meaningful. Even if two answers contain the same number of tuples, they can still be distinguished by their different semantic paths. Further, the semantic paths are used to annotate answers to facilitate users? understanding.

Third, we extend the keyword query language to include keywords that match meta-data, that is, the names of relations and attributes. These keywords provide the context of subsequent keywords and explicitly indicate the search targets of the query. Thus, the ambiguity of keyword queries are significantly reduced and we can infer users? search intention more precisely than existing methods. We use the ORA semantics to construct an ORM schema graph and determine the objects and relationships referred to by the keywords in a query. We obtain a set of minimal connected graphs called query patterns to represent the possible search intentions of the user. We rank the query patterns based on the search targets of the query and the number of objects captured in the patterns. The top-k ranked query patterns are translated into SQL statements to retrieve the information that the user is interested in.

Finally, we further extend keyword queries to incorporate aggregate functions and GROUPBY, e.g., {John COUNT Course}. The work SQAK supports aggregate functions in keyword queries. However, it does not have the concept of objects and cannot distinguish objects with same attribute value, e.g., two students called John. As a result, it may return incorrect answers. In contrast, we utilize the ORA semantics to distinguish objects with the same attribute value and detect duplications of objects in relationships in order to compute aggregates correctly. Based on the ORM schema graph, we generate a set of query patterns and annotate these patterns to determine various query interpretations. Furthermore, we detect duplications of objects/relationships arising from unnormalized relations, and extend our approach to handle aggregate queries on unnormalized databases. We show that without the ORA semantics, it is impossible to process aggregate functions correctly.