NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

PH.D DEFENCE - PUBLIC SEMINAR

Title: **ON REPAIRING STRUCTURAL ISSUES IN SEMI-STRUCTURED DOCUMENTS**

Speaker: Ms Ying Shanshan

Date/Time: 22 September 2014, Monday, 02:00 PM to 03:30 PM

Venue: Executive Classroom, COM2-04-02

Supervisor : Dr Anthony Tung, Associate Professor, School of Computing

Abstract:

Poor quality of data can have a substantial social and economic impact. Although data quality management is a well-established research area, the vast majority of prior works focus on relational data. Increasingly, semi-structured data, such as XML, JSON, etc., are becoming the de facto standard for a huge variety of data formats and applications. Their flexibility and easy-customization contribute to the soaring popularity of semi-structured data, but also serve as significant sources of major data quality errors. Well-formedness of structure, a prerequisite for many research works on semi-structured data, is an assumption often does not hold. Many XML documents suffer from erroneous structures, such as improper nesting where open- and close-tags are unmatched. Apart from this, elements are possibly organized in an incorrect hierarchy or sequence, leading to unexpected number of occurrence.

To enforce the balance of open- and close- tags, we propose in this thesis two algorithms targeting at different structural constraints. The first algorithm focuses on tags only while the second limits the occurrence of text in the document. Thorough proofs are presented on the completeness and approximation ratio of these algorithms. Besides we concentrate on detecting unexpected elements errors, when there are missing or spurious elements. We propose novel techniques to detect such errors and provide plausible reasoning for every reported error and a summarization technique based on variations of set cover for concise reporting. We demonstrate the effectiveness of these algorithms on real datasets through extensive experimental study.