Title: Profiling Entities over Time with Unreliable Sources

Speaker: Ms Li Furong

Date/Time: 22 March 2017, Wednesday, 02:00 PM to 03:30 PM

Venue: Executive Classroom, COM2-04-02

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

Abstract:

In the age of Big Data, an entity’s information is, more often than not, published by more than one data source. Each source may describe the same entity in different ways and from different aspects; their provided information may be incomplete, imprecise, and valid for different time periods. In order to obtain a complete picture of a real-world entity, one has to integrate data records that refer to the same entity.

This thesis studies how to construct complete and accurate profiles for real-world entities by integrating data records from different sources. We first present a framework called Comet to profile entities in the presence of erroneous values. It interleaves record matching with error correction, taking into consideration the varying source reliabilities on different attributes.

Next we consider how to decide the true attribute values of an entity when there may exist multiple truths. We propose a truth discovery model called Hybrid that jointly makes two decisions: how many truths there are, and what they are. It considers the conflicts between values as important evidence for ruling out wrong values, while keeps the flexibility of allowing multiple truths. In this way, Hybrid is able to achieve both high precision and high recall.

Subsequently, we examine how to construct a historical profile real-world entities when the attribute values may change over time. In particular, we develop a transition model to capture the probability that an entity changes to a particular attribute value after some time period. The transition model provides a fine-grained understanding of how entities may evolve over time, and thus enables us to identify the records describing the same entity but at different times.

We build upon the above techniques and present a Maroon+ framework for profiling entities over time with unreliable sources. It considers various source characteristics (correctness,
completeness, and timeliness), as well as the probability of value transitions when integrating information from different sources. We also developed a prototype of Maroon+, which allows interactive entity profiling, and provides explanations on the obtained results.