An Algebraic Language for Automatic Semantic Data Integration on the Hidden Web

Shazzad Hosain
Department of Computer Science
Wayne State University

Tuesday, September 22, 2009
3:00pm Rm. 110 Purdy-Kresge Library

Abstract:
Semantic integration of the hidden Web is an emerging area of research where traditional assumptions about schema do not always hold and semantic heterogeneity poses serious challenge. Constant changes, conflicts and sheer size in the world of hidden Web demand integration techniques that rely on autonomous detection and resolution heterogeneity, correspondence establishment and information extraction strategies. First it needs to automate those techniques and then to integrate those techniques or sub-systems automatically into a single system. Though many such sub-systems have been automated, to our knowledge, there is no integrated framework for combining those technologies automatically. Our idea is to exploit the flexibility and strengths of a declarative language and the first step of such a language is to give an algebraic foundation that takes various integration techniques into consideration. In this thesis, we present an algebraic language, called \texttt{Integra}, as a foundation for an SQL like query language such as BioFlow for the integration of Life Sciences data on the hidden Web. The algebra presented here assumes that all web pages can be thought of as traditional relations and the integration techniques can be considered as user defined functions. These assumptions make it possible for us to extend the traditional relational algebra to include integration primitives such that a database with traditional relations reduces to a special case in our model. The algebra relies on a schema matching function, a key discovery function, a wrapper or extraction function and two new operators \texttt{link} and \texttt{combine} that embody the well known concepts of horizontal and vertical integration.

Biography:
Md. Shazzad Hosain did his Masters degree on August 2005, from Computer Science and Engineering department of Bangladesh University of Engineering and Technology. His M.Sc. thesis focused on measuring software reliability using Markov chain usage model. He then joined at the Computer Science department of Wayne State University during Fall 2005 and waiting to defend his dissertation soon. His PhD thesis focuses on developing an algebraic language for autonomous semantic data integration on the hidden Web. His other research interests involve in ontology language, knowledge representation and schema reconciliation.