Talk Title: EROCS: Automatically linking documents with relevant structured information

Speaker: Dr. Mukesh Mohania

Affiliation: IBM Research, India

Venue: G31 Electrical Engin.

Time: May 1st Friday, 12:30 – 1:30 pm

Abstract:

Consolidated analysis of critical business information distributed across structured and unstructured data is a key enabler for next generation business intelligence and search. In this work, we address the problem of linking a given text document with relevant structured data, retrieved automatically from a RDBMS. We have developed a prototype system, called EROCS, that views the structured data as a predefined set of ``entities'' and identifies the entities that best match the given document. EROCS also embeds the identified entities in the document, effectively creating links between the structured data and segments within the document. Unlike prior approaches, EROCS identifies such links even when the relevant entity is not explicitly mentioned in the document. EROCS exploits sophisticated optimizations in order to perform this task keeping the amount of information retrieved from the database at a minimum. We are also working on extensions of EROCS that enable such linkage even in the presence of noise and errors in the unstructured documents, and, in parallel, perform cleansing on the noisy documents with respect to the relevant structured data.

Bio

Mukesh Mohania received his Ph.D. in Computer Science & Engineering from Indian Institute of Technology, Bombay, India in 1995. He was a faculty member in University of South Australia and Western Michigan University from 1995-2001. He was also associated with Kyoto University and Purdue University as Senior Researcher from 1996-2001. Currently, he is a senior manager in IBM India Research Lab, and leading Information Management research group. He has worked extensively in the areas of distributed databases, data warehousing, data integration, and autonomic computing. He has published more than 100 papers and also filed more than 20 patents in these or related areas. He has also organized several conferences and workshops as PC Chair, General Chair, and Industrial Chair, and has served the program committee of several conferences. He is also associated with the Journal of Database Management and Transaction on Data Privacy as editorial board member.

He was awarded Technical Achievement Award in the area of Web Database Management and Data Warehousing by Association of Database and Expert Systems Applications in Greenwich, U.K., 2000. He received the He received the best paper award for his XML and data integration work in CIKM 2004 and CIKM 2005, respectively. He received an award from IBM Tivoli Software in 2004 for his research contribution to Policy Management for Autonomic Computing product. He was also a recipient of the "Excellence in People Management" award in IBM India in 2007. He received the “Outstanding Innovation Award” from IBM Corporation in 2008 for his Context-Oriented Information Integration work. He is an IEEE Distinguished Speaker.