Guest Lecture

HAEC

Text Mining on Large Document Collections

Dr. Rainer Gemulla

Max Planck Institute, Saarbrücken

Tuesday, 11 June 2013

1:00 pm - 2:00 pm

Room: NOE 3105

(building of the Faculty of Computer Science, Nöthnitzer Strasse 46)

Abstract: In this talk, I summarize some of our work on various text mining problems. I present ClausIE, a novel, clause-based approach to open information extraction, which extracts relations and their arguments from natural language text. ClausIE fundamentally differs from previous approaches in that it separates the detection of ''useful'' pieces of information expressed in a sentence from their representation in terms of extractions. The second part of my talk focuses MG-FSM, a scalable approach to pattern mining on textual data. While (some variants of) the problem have been extensively studied, few of the available techniques are sufficiently scalable and flexible to handle complex datasets with billions of sentences.

Bio: Rainer Gemulla received a Ph.D. in computer science from the Technische Universität Dresden. He subsequently joined the IBM Almaden Research Center in San Jose, USA, as a postdoc and now holds a senior researcher position at the Max Planck Institute in Computer Science in Saarbrücken. Dr. Gemulla's current research centers around algorithms and systems that can effectively and efficiently extract useful information from large, complex datasets. His interests include data mining, text mining and information extraction, data-intensive optimization, and approximation techniques.

TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

DFG