Department of Computer Science
Research Focus
"Intelligent Formal Systems"

TECHNISCHE UNIVERSITÄT DRESDEN

Colloquium

Within the seminar of the Research Focus, the Ph.D. programme "Specification of discrete processes and systems of processes by operational models and logics" presents a talk by

Dmitrii Ignatov
Institut für Algebra
Technische Universität Dresden
Germany

about

"Finding near-duplicate documents in the web: an FCA approach"

Abstract: Near-duplicate documents make almost 30% of output of web search machines, so the general problem is to reduce the amount of output near-duplicates. To this end one should first define similarity on documents. This is ususally done in two ways: lexically (i.e., using word frequencies) and syntactically (treating texts as sequences of symbols). First concise representation of documents by subsets of lexical or syntactical features is obtained. Then, due to computation limitations, similarity of documents is usually defined as a relation on pairs of document images. This relation is non-transitive, which causes problems in definition of clusters of similar documents. We propose an approach where a cluster of documents is defined as an extent of a formal concept, such that the corresponding intent is the set of common features of the documents. For finding similar documents we intend to use efficient algorithms known in the data mining community.

Date:	31.05.2005
Place and time:	Department of Mathematics and Natural Sciences, Room: WIL 129, 1.15 pm

All interested persons are invited to attend!

Prof. Dr.-Ing. habil. Heiko Vogler