

Tutorial at the 14th International Conference on Advances in Databases, Knowledge, and Applications – DBKDA, Venice/Italy, May 22 – May 26, 2022

Title: From Ctrl-F to Information Retrieval – An Excursion into the World of Searching in Text

Organizer:

Prof. Dr. Andreas Schmidt

Department of Computer Science and Business Information Systems
University of Applied Sciences
Karlsruhe, Germany

Institute for Automation and Applied Informatics
Karlsruhe Institute of Technology
Karlsruhe, Germany
email: andreas.schmidt@kit.edu

Tutorial description:

This tutorial is about searching in text. We give an overview of the different possibilities to search within documents. Starting from the search in the original documents, different search variants are presented. One way to increase the speed and flexibility of the query is to pre-process the search request. But when the amount of data grows, the search in the original documents reaches its limits, which is why we present the Inverted Index, as it is used typically in Information Retrieval (IR). Different variants of the inverted index, for the different types of search queries, are presented. Next, the vector-space-model, together with the cosine similarity and the tf*idf measure is introduced. In this context, the ranking of result-document is explained. We finish the tutorial with a rough overview over the different preprocessing mechanisms, to adapt the search to the concrete demands and domains.

Materials to be distributed to the attendees:

- Slideset
- Refcard with the most useful shell commands in the context of searching.
- Practical exercise

Duration: (1 hour)

Short Bio:

Prof. Dr. Andreas Schmidt is a professor at the Department of Computer Science and Business Information Systems of the Karlsruhe University of Applied Sciences (Germany). He is lecturing in the fields of database information systems, data analytics and model-driven software development. Additionally, he is a senior research fellow in computer science at the Institute for Applied Computer Science of the Karlsruhe Institute of Technology (KIT). His research focuses on database technology, knowledge extraction from unstructured data/text, Big Data, and generative programming. Andreas

Schmidt was awarded his diploma in computer science by the University of Karlsruhe in 1995 and his PhD in mechanical engineering in 2000. Dr. Schmidt has numerous publications in the field of database technology and information extraction. He regularly gives tutorials on international conferences in the field of Big Data related topics and model driven software development. Prof. Schmidt followed sabbatical invitations from renowned institutions like the Systems-Group at ETH-Zurich in Switzerland, the Database Group at the Max-Planck-Institute for Informatics in Saarbrücken/Germany and the Data-Management-Lab at the University of Darmstadt.