Tobias Berka, Ph.D. – Research Statement

Born: 28th of October 1978
Nationality: Austrian
Occupation: Post-Doc
e-mail: tberka@cosy.sbg.ac.at
Web Site: www.cosy.sbg.ac.at/~tberka

Research Statement

My overall research focuses on text search using the most flexible and reliable mathematical family of representations: vector space semantics. But for some time now, these models have been quite costly in computational resources - in both computation and storage. I work on improved representations and algorithms to further improve the precision, reduce the cost, and increase our ability to pay it.

The goal is to provide the gold standard in retrieval accuracy and search flexibility, by providing interactive query refinement, query expansion and query-by-example functionality. Our scenario is to support search for corporate intranets, knowledge bases, case files, design documents, e-mail archives and other document collections containing millions or up to hundreds of millions of documents. The objective is thus not web search, which has to meet very different information needs, and where very few users are interested in more than one hundred results. Our users are professionals and experts, so both completeness and accuracy of search results play a much bigger role. Our scenario is also limited to mere hundreds of millions of documents, making a more comprehensive result analysis possible. At the same time, practical applications in criminal, legal, medical, or engineering information retrieval can impose stringent deadlines on the document analysis process. With deadlines of mere hours or days, the initial corpus analysis has to be completed as quickly, to allow the human experts to set to work.

The prototype code runs well on a range of different platforms, from desktops and servers to off-the-shelf clusters, server racks and, potentially, supercomputers. It exploits the parallelism available in today’s multi-core processors but scales up to large parallel server installations found in data centers. In the long run, it will be able to install and run the search engine locally on your personal computer or to deploy it in a data center as a search cloud for groups of users.