Diploma/Master Thesis
Queries Crossing the Structure Chasm

Structured relational databases have gained great success for data management in the past three decades. However, with the advent of the internet, more and more information are presented in unstructured/semi-structured web texts. Emails and many other documents on personal computers are also plain texts. Such loosely structured data has great demands for data management facilities such as querying.

On one hand, the prevalent usage against texts nowadays is enabled by information retrieval (IR) techniques, e.g., via keyword search. Modern search engines like Google follow a keyword-in-document-out paradigm. Users have to follow links and navigate possibly large texts to locate interesting information they need. On the other hand, structured data stored in relational databases require expertise in database schemas to manipulate and is not friendly for casual users. Furthermore, there is no good technique that is able to explore relationship between structured databases and plain text documents or web texts.

Based on the above observation, our project is carried out in the context of dataspace, which is an abstraction of a collection of data sources including structured sources (e.g., RDBMS), unstructured text (e.g., emails), and possibly web data. The goal of the project is to investigate and develop query evaluation techniques crossing the structure chasm described above, namely, to enable structured query over texts.

A qualified candidate of the thesis project is expected to meet the following requirements:

- Good programming skills in Java (C# is also acceptable)
- Background in SQL and RDBMS
- Knowledge of query evaluation or optimization is a plus
- Knowledge of information retrieval and/or extraction is a plus

For further information please contact:

Xiang Li  
Lehrstuhl Informatik 5  
Phone: 0241/80-21505  
lixiang@cs.rwth-aachen.de

Dr. Christoph Quix  
Lehrstuhl Informatik 5  
Phone: 0241/80-21511  
quix@cs.rwth-aachen.de