Bachelor/Master Thesis
Best-effort Schemaless Reference Reconciliation

Information overload is a common symptom in the Internet age nowadays. Search engines assist users to seek a "needle in a haystack". However, the evolving demand of data intensive applications now asks for not only an isolated piece of information, but also a collection of interlinked data elements. Furthermore, in order to enable machine understandability, the information needs to be structured.

Our project is a first step towards the goal of finding useful structured information from unstructured data. We aim at consolidate a collection of triples extracted from the web, so that duplicates, either explicit or implicit, are identified and merged. The responsibility of the thesis candidate is to develop a scalable approach to reconcile natural-language triples using well-established algorithms in databases and information retrieval.

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

• Experienced in Java programming
• Good command of English
• Knowledge of Algorithm and Data Structures

http://dbis.rwth-aachen.de/cms/theses/ref-reconciliation

For further information please contact:

Xiang Li
Lehrstuhl Informatik 5
Phone: 0241/80-21514
lixiang@cs.rwth-aachen.de

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