Bachelor project for 2009/2010
Information Systems Group
Prof. Dr. Felix Naumann

Midas: Extreme Web Data Integration for Government Data

Project Background

The wealth of freely available, structured information on the Web is constantly growing. This is especially true for public data from and about governments and administrations. Data-providing projects, such as DBPedia and Freebase from the linked open data community, as well as structured data from domain-specific sites, such as senate.gov, USASpending.gov, or epp.eurostat.ec.europa.eu, make it possible to integrate data from multiple sources and thus create new data sets with added value. The recent appointment of Tim Berners-Lee to lead a review on how the UK government can open up access to official information reinforces this trend. However, the integration of such data sources is far from trivial: Apart from technical difficulties of accessing the data, structural and semantic differences in the data must be overcome. In particular, the various data sets must be standardized, transformed to a common structure, cleaned and finally consolidated into a single, consistent and complete data set.

Government data is a rich domain, but notoriously difficult to access and comprehend. The combination of budget data, spending recipients, data about representatives, data from various government agencies and geographic data promises several rich applications and use cases. For instance, in the 2009 American Recovery and Reinvestment Act $ 787 billion of the Stimulus Plan has been committed in various investment funds. The Recovery Accountability and Transparency Board is charged with overseeing all funds dispersed under the Act. It is charged with ensuring that funds are used in a manner consistent with the goals of economic recovery and job creation, as well as provide the American public with a transparent view of how that process is unfolding. Earlier this year, the Recovery Board decided to build a broader problem-solving community. The result was a week-long, online Recovery Dialogue on IT Solutions (http://www.thenationaldialogue.org/) and a report at http://www.recovery.gov/sites/default/files/NAPA_Recovery+Dialogue_Final+Report_5-20-09_0.pdf.

In the report, “data analysis and visualization” and “data collection” are the top-two most tagged idea topics. Among all companies, IBM has committed the highest number of participants. Indeed, IBM has tremendous interests to provide advanced research, developments, and solutions to solve the IT challenges described in the report

Cooperation Partner

Midas is a joint project between IBM’s Almaden Research Lab and IBM’s Silicon Valley Lab. The project’s goal is to provide an end-to-end framework for the process of integrating heterogeneous Web data into a common, clean and consistent data set. Individual components extract data, scrub it in a source-specific manner, identify common entities across multiple sources, transform data to a common structure and finally fuse possibly conflicting data into a value-added, rich and clean data set. Underlying technologies of this Java-based

tool include the relational and json data models, database operations based on SQL and Jaql, text extraction rules, similarity-based duplicate detection, mapping-based data transformation, and data fusion.

The project will commence in close cooperation with the partners, in particular through regular telephone-conferences. We will be working jointly on the source code, giving students the chance of strong impact on an ongoing, high-profile research and development project.

Project Description

The goal of the bachelor project is to establish the *government* domain for Midas. This goal includes the discovery of relevant data sources both from the US and the EU, to explore and extract data from those sources, to develop methods for scrubbing such data, develop techniques to discover duplicate entries and links among data, and finally to fuse duplicate data entries. Thus, the project team will add a new set of sources to Midas, it will adapt existing techniques for the individual steps, and it might develop new techniques that are more suitable for the domain and use case.

Participation is limited to 8 students who should have a strong database background. The project starts on October 5, 2009. Among the preparatory steps are tutorials on Midas, on the domain and on the use case.

Project advisors are Felix Naumann, Jana Bauckmann, and Christoph Böhm. For further information please contact any of the advisors directly.