Bachelor project for 2011/2012
Information Systems Group
Prof. Dr. Felix Naumann

HPI Information Systems Group
Hasso Plattner Institut | Universität Potsdam

CelebDB: Harvesting Celebrity Data

Finding well-known persons (celebrities) in text documents is a common task that can be used for many applications. There are thousands of celebrities around the world who are well-known in one or several domains, such as sports, science, film, politics, music, etc.

Many people are interested to gather information about celebrities, including their images, awards, age, family, background, activities, people's opinions about them, etc. In addition, companies that want to employ celebrities for their advertisements and other marketing activities are interested to find celebrities that are the best fit for their products and image. They also like to rank celebrities based on different aspects, such as trustworthiness, performance, opinion leadership, uniqueness, etc. Finally, celebrities themselves are eager to know what are the public opinions about them and how they rank among other celebrities in the same field.

This and more information about celebrities is readily available on the Web in the form of structured data, textual documents, blogs, tweets, web services, etc. We plan to build a system that can automatically harvest celebrity data from the Web and present them in a comprehensive portal for the general public, for marketing departments and for the celebrities themselves.

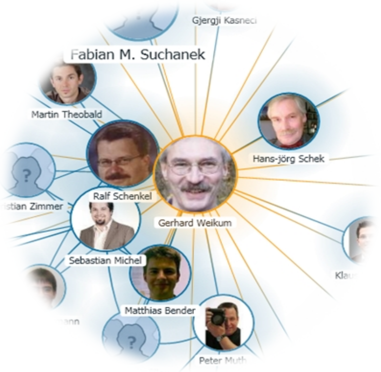
Project Description

The goal of the project is to populate the CelebDB database and present its data in a celebrity portal. The project partner provides a list of 5,000 celebrities of which only 500 celebrities are included in the database. The current version of the database is populated manually and it only contains factual information about celebrities. In the bachelor project, we want to expand the database automatically in order to cover more celebrities and include more information about them using the following techniques:

1. **Structured data collection:** Find factual information, such as name, date of birth, spouse, address, and awards, from various structured data sources including the general-purpose sources, such as DBpedia and Freebase and domain-specific sources, such as IMDB or sports.de.

2. **Unstructured data extraction:** Find textual content about celebrities, such as news snippets and peoples opinion. Opinionated text should be labeled by

its polarity (positive/negative), the strength of the polarity, and the aspect of the opinion (trustworthiness, performance, opinion leadership, uniqueness, etc.).



3. **Relationship extraction:** Find relations between each celebrity and other celebrities, products, companies, and topics.

4. **Relationship visualization:** Visualize the relations extracted in the previous step.

5. **Celebrity ranking:** Rank celebrities based on different weighted aspects (trustworthiness, performance, ...), their domain (sport, science, film, politics, music,...), or the number of followers on social networks (Facebook, Twitter,...).

6. **Celebrity classification:** Classify celebrities according to predefined classes, such as domains, age, products, etc.

7. **Celebrity clustering:** Cluster celebrities based on some similarities and building clouds/groups of relevant celebrities.

8. **Celebrity portal enrichment:** Enrich the celebrity portal by additional information from web services including images via Flickr, movies via YouTube, books, tweets, etc.

Project Cooperation and Supervision

CelebDB is a joint project between Celebrity Performance GmbH (CPI). CPI has developed a unique index to evaluate celebrities based on their media impact potential.

On the one side CPI supports marketing decision-makers in finding the perfect testimonial when marketing new products. On the other side, they offer celebrities and their management the opportunity to make all relevant information on their public perception and advertising effect potential available, in order to improve self-marketing strategies. For the perfect result CPI wants to develop a new research method that combines available information from the Web with accurate web-based-analysis on the perceived image of celebrities.

cpi
CELEBRITY
PERFORMANCE

The project starts October 1st, 2011 and will be advised by Prof. Dr. Felix Naumann and Dr. Saeedeh Momtazi. For further information about the project please contact saeedeh.momtazi@hpi.uni-potsdam.de.