

VIP 2.0: Celebrity Exploration



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.

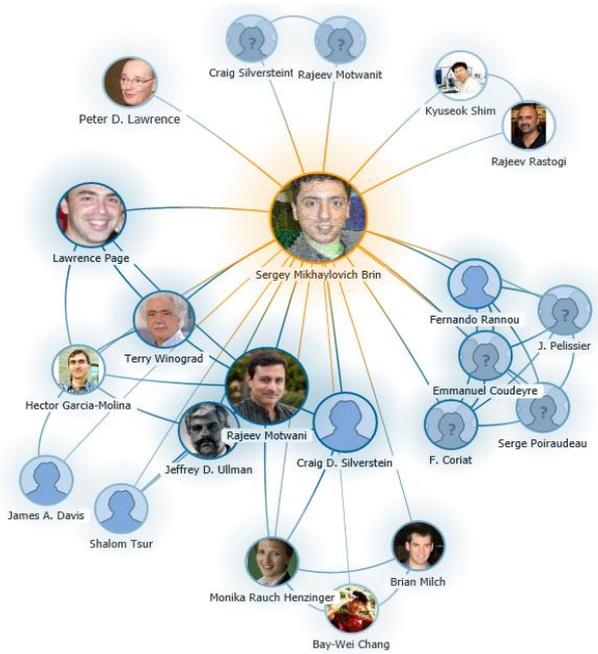
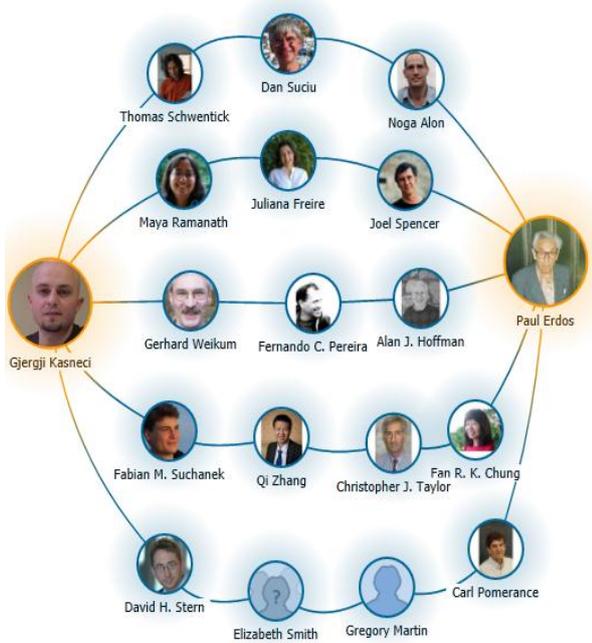
Companies that want to employ celebrities for their advertisements and other marketing activities are interested in finding celebrities that are the best fit for their **products** and **brand image**. Also, celebrities themselves are eager to learn about their public image, opinions about them, and how they rank among other celebrities in the same field. Finally, celebrities stand in many **relationships** to other celebrities (spouses, friends, colleagues), to topics (sports, health, etc.) and to brands and products (DiBa, Gummibärchen, etc.).

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 extend a system that can automatically gather celebrity data from the Web in a focused manner, analyze them, 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 celebrity database of our project partner **cpi Celebrity Performance GmbH** and present its data on its portal. The project partner provides a list of 1,500 celebrities: Based on the result of the previous bachelor project, the current version of the database is partially populated automatically with some factual information and some popularity factors about celebrities. In this bachelor project, we will develop methods for automatically enriching the database with new relationships between celebrities, celebrities and products, celebrities and companies, etc. We intend to use the following techniques:

1. **Supervised relationship extraction:** In order to enrich the database with new relations between celebrities and other celebrities, products, companies, and topics, we plan to start with known relationship instances and search on various web corpora (e.g., Wikipedia, news articles, etc.) for new instances that adhere to certain relationship patterns that will be automatically learned.
2. **Relationship visualization:** Visualize the relations extracted in the previous step. A hyperbolic graph browsing interface would allow the access and discovery of interesting connections between celebrities, celebrities and products, celebrities and companies, etc. As an example, consider the following two graphs from Microsoft's Academic Search:



On the left, we see various paths that interconnect two persons; on the right, the local graph surrounding the Google founder Sergey Brin is shown.

3. **Focused Crawling:** Besides gathering information from different sources that provide APIs for accessing their data (Twitter, Google, ...), cpi wants to extract relevant text also from other sources, such as news articles or blogs. Hence, we plan to implement a topic-based crawler that can be used to find relevant text snippets about celebrities.
4. **Event Detection:** The awareness of celebrities in the media is often also based on events. For instance, during June 2012 the German soccer players were more present in media, current film releases boost the media presence of the actors, etc. We plan to create a system that is able to detect relevant events automatically and relate them to statistical peaks in the celebrity data.

Project Cooperation and Supervision

VIP 2.0 is a joint project with cpi Celebrity Performance GmbH. 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 is developing methods that combine available information from the Web with accurate web-based-analysis on the perceived image of celebrities.



The project starts in October 2012 and will be advised by Prof. Dr. Felix Naumann and Dr. Gjergji Kasneci. For further information about the project please contact gjergji.kasneci@hpi.uni-potsdam.de.