# EvenPers: Event-based Person Exploration and Correlation

Christian Kapp, Jannik Strötgen, Michael Gertz

Institute of Computer Science, Heidelberg University
69120 Heidelberg, Germany
c.kapp@stud.uni-heidelberg.de
{stroetgen,gertz}@informatik.uni-heidelberg.de

**Abstract:** Searching for people on the Internet is one of the most frequent search activities. In this paper, we present *EvenPers*, a system for the event-based exploration of persons and person similarities. We address challenges such as cross-document person name normalization and present a novel approach to calculate person similarities based on their event information. In our demonstration, we show several exploration scenarios illustrating the usefulness of *EvenPers* and its exciting functionality.

## 1 Motivation and Objectives

A very common activity on the Web is that users search for information about (prominent) people. A major problem in identifying documents that are relevant to a user's search query is that person names need to be disambiguated. That is, a person can be referred to by different expressions (including personal pronouns) in documents, which need to be identified as such and mapped to some *normalized* person name. A typical example is the different ways the president of the US is referred to in documents. All expressions relating to the president first need to be identified as such in documents and mapped to a single expression, ideally the full name of the person, before a ranking of documents relevant to the search query can be determined. Several approaches have been proposed for person name disambiguation, see, e.g., [Chr06, YIO$^+$10].

In our approach, we go a step further and consider an *event-based context* in which person expressions occur in documents. The motivation is that a person can be well characterized by the events he or she was or will be involved in. For this, we use a simple yet powerful notion of event as a combination of a time and geographic expression, typically at the sentence level [SGJ11, SG12a]. The idea then is to combine such event information with person expressions, leading to the construction of a *person's event profile*. A search result then is not a list of relevant documents but a ranked list of events related to the person. Such a view leads to new functionality to explore information about a person, e.g., a search based on a certain time interval or geographic region of interest. A further novelty of our approach is that event-profiles for persons are used to determine persons that are similar based on the events they were involved in. Thus, our approach and system outlined in the following sections provide new functionality to search for and explore person information.

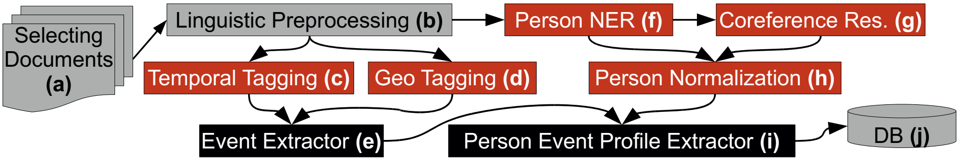

Figure 1: *EvenPers* document processing pipeline with extraction and normalization of temporal, geographic, and person information (red) and the event and person event profile extractors (black).

## 2 System Description

**Processing Pipeline:** Our document processing pipeline is depicted in Figure 1. After selecting documents from the corpus (a), linguistic preprocessing, such as sentence splitting, tokenization, and part-of-speech tagging, is performed (b). The results of this step are then available for further annotation tools. All temporal expressions (c) and geographic expressions (d) are extracted and normalized using HeidelTime [SG12b] and Yahoo! Placemaker[1]. These are then combined into events of the form $e = \langle t, dp_t, c_t, g, dp_g, c_g \rangle$ with normalized temporal and geographic values $t$ and $g$, document/position information $dp_t$ and $dp_g$, and confidence values for correct normalization $c_t$ and $c_g$, respectively (e).

In parallel, named entity recognition for detecting person names is performed (f) by StanfordNER [FGM05] and the OpenNLP NER tool[2]. In the next step, several tools are applied to resolve coreferences (g), namely Arkref [HK09], Cherrypicker [RN09], and the Illinois Coreference Package [BR08]. Since there are hardly any sophisticated person name normalization tools, we developed our own tool for this task (h): Based on the extracted NER and coreference information of the previous applied tools, the person information is merged into person chains before Wikipedia and JRCNames[3] are checked for different name variations of each item of the person chain to associate a Wikipedia and/or JRCNames ID, if available. Depending on detection and normalization details, a confidence value $c_p$ can be added to every reference to a person in addition to document/position ($dp$) and normalization information ($id_p$), resulting in a reference to a person as $p = \langle id, dp_p, c_p \rangle$.

Finally, the Person Event Profile Extractor (i) combines events with person information to create a person event profile $pep(p)$ for every person detected in the corpus. For this, co-occurrences of events and references to persons within a specified window size are determined in every document of the corpus, resulting in a person event profile of the form $pep(p) = \{\langle e_1, p_1 \rangle, ..., \langle e_n, p_n \rangle\}$. Note that for every item, a confidence value can be calculated depending on $c_t$, $c_g$, and $c_p$. All $pep$ are stored in a PostGIS database (j).

**Person Similarity Calculation:** For determining the similarity between persons, we rely on a model for event-centric document similarities described in [SGJ11]. For every two

[1] Yahoo! PlaceMaker: http://developer.yahoo.com/geo/placemaker/
[2] OpenNLP: http://opennlp.apache.org/index.html
[3] JRCNames: http://langtech.jrc.it/JRC-Names.html

persons $p_1$ and $p_2$, event similarities $esim$ for the cross-product of the events in $pep(p_1)$ and $pep(p_2)$ are calculated based on the events' temporal and geographic granularities as detailed in [SGJ11]. Then, we build the sum over all $esim$ weighted by their confidence weighting factor ($wf$), and normalize the value by the size of $pep(p_1)$ and $pep(p_2)$ and the average confidence $avgc$, resulting in a person similarity function of the form:

$$personSim(p_1, p_2) := \frac{\sum_{e_i \in pep(p_1)} \sum_{e_j \in pep(p_2)} wf(e_i, e_j) \times esim(e_i, e_j)}{|pep(p_1)| \times |pep(p_2)| \times avgc}$$

## 3 Demonstration

**Corpus:** The prerequisites of the underlying data set for our demonstration are that the corpus contains (i) information of many different persons, (ii) about different times, (iii) with the information about the persons being somehow related to each other. Then, it should be possible to identify reasonable person similarities based on our approach described in Section 2. For this, we selected all documents with the same category tag (politics) of the New York Times corpus[4], resulting in 209,795 news documents with publication times between 1987 and 2007. These documents contain over 5 million personalized events about 82,616 different persons.

**Demonstration Scenarios:** We present the following demonstration scenarios, which we briefly explain based on the screenshot of *EvenPers* depicted in Figure 2:

(a) Searching for persons ("Helm") results in a hit list. After the user makes a selection ("Helmut Kohl"), the person's details are presented: a picture, the Wikipedia link (if available), the person's most important events and most similar persons (left side of Figure 2). On the map, the person's events are anchored at the places where the events occurred.

(b) In the *person-centric exploration scenario*, a second person is selected from the "most similar persons" list ("Bill Clinton"). His/her events are added to the map and shared events are highlighted in a special color. The events of the two persons can be directly compared with each other and explored using event snippets described in (d).

(c) In the *event-centric scenario*, the user selects an event on the map or in the "most important events" list. In our example, the user selected the event "1998-05-14 – Potsdam, Brandenburg, DE". The event information is presented in a snippet anchored at the event's location as shown in the figure and described next.

(d) Event snippets contain: (i) normalized time and place information, (ii) a list of persons sharing the event – the user can select a person from the list to compare it with the first person as in the person-centric scenario, (iii) a list of contexts, containing event occurrences of the persons under investigation, with highlighted time, place, and person information.

(e) Finally, the user can filter events based on their importance, confidence, and year.

As the scenarios demonstrate, *EvenPers* provides exciting functionality on exploring single persons and person similarities based on their events extracted from different documents.

[4] New York Times Corpus is available from LDC (http://www.ldc.upenn.edu/), catalog number LDC2008T19.

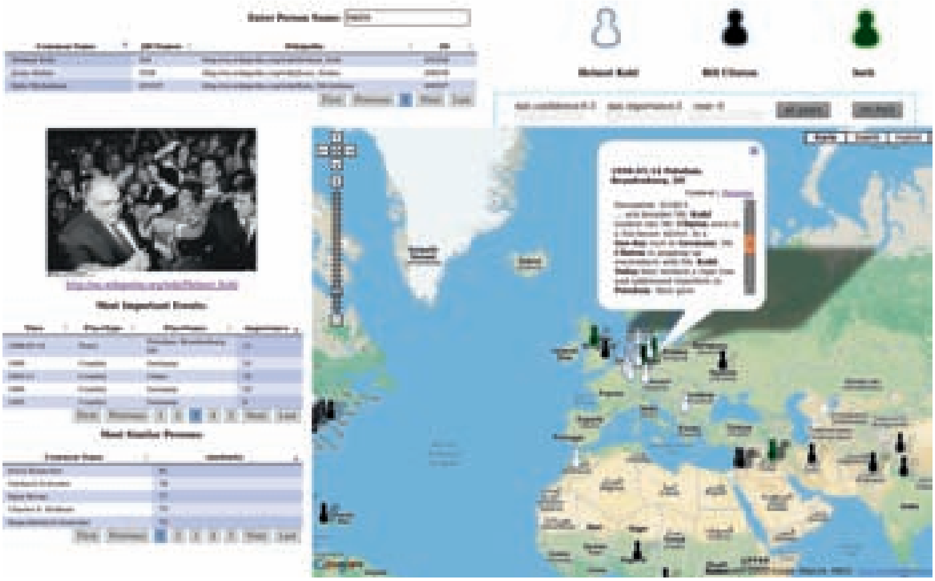

Figure 2: The *EvenPers* system: Exploring similarities between Helmut Kohl and Bill Clinton.

# References

[BR08] Eric Bengtson and Dan Roth. Understanding the Value of Features for Coreference Resolution. In *EMNLP'08*, pages 294–303, 2008.

[Chr06] Peter Christen. A Comparison of Personal Name Matching: Techniques and Practical Issues. In *ICDM'06 Workshops*, pages 290–294, 2006.

[FGM05] Jenny R. Finkel, Trond Grenager, and Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *ACL'05*, pages 363–370, 2005.

[HK09] Aria Haghighi and Dan Klein. Simple Coreference Resolution with Rich Syntactic and Semantic Features. In *EMNLP'09*, pages 1152–1161, 2009.

[RN09] Altaf Rahman and Vincent Ng. Supervised Models for Coreference Resolution. In *EMNLP'09*, pages 968–977, 2009.

[SG12a] Jannik Strötgen and Michael Gertz. Event-centric Search and Exploration in Document Collections. In *JCDL'12*, pages 223–232, 2012.

[SG12b] Jannik Strötgen and Michael Gertz. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, pages 1–30, 2012. 10.1007/s10579-012-9179-y.

[SGJ11] Jannik Strötgen, Michael Gertz, and Conny Junghans. An Event-centric Model for Multilingual Document Similarity. In *SIGIR'11*, pages 953–962, 2011.

[YIO$^+$10] Minoru Yoshida, Masaki Ikeda, Shingo Ono, Issei Sato, and Hiroshi Nakagawa. Person Name Disambiguation by Bootstrapping. In *SIGIR'10*, pages 10–17, 2010.