Tracking Entities in Web Archives: The LAWA Project*

Marc Spaniol
Max Planck Institute for Informatics, Germany
mspaniol@mpi-inf.mpg.de

Gerhard Weikum
Max Planck Institute for Informatics, Germany
weikum@mpi-inf.mpg.de

ABSTRACT

Web-preservation organization like the Internet Archive not only capture the history of born-digital content but also reflect the zeitgeist of different time periods over more than a decade. This longitudinal data is a potential gold mine for researchers like sociologists, politologists, media and market analysts, or experts on intellectual property. The LAWA project (Longitudinal Analytics of Web Archive data) is developing an Internet-based experimental testbed for large-scale data analytics on Web archive collections. Its emphasis is on scalable methods for this specific kind of big-data analytics, and software tools for aggregating, querying, mining, and analyzing Web contents over long epochs. In this paper, we highlight our research on *entity-level analytics* in Web archive data, which lifts Web analytics from plain text to the entity-level by detecting named entities, resolving ambiguous names, extracting temporal facts and visualizing entities over time periods. Our results provide key assets for tracking named entities in the evolving Web, news, and social media.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Experimentation, Measurement

Keywords

Temporal Web Analytics, Entity Analytics, FIRE

1. INTRODUCTION

National libraries and organizations like the Internet Archive (archive.org) and its European sibling (internetmemory.org) have been capturing Web contents over more than a decade and have protected Web contents from vanishing [6]. The emergence of large Web-contents repositories and digitization projects open up anew range of analytical opportunities and challenges along the temporal dimension [1]. Studies reveal "culturomics" phenomena [8], track the trustworthyness of memes over time (truthy[1]) or even investigate the Web's predictive power (such as recorded future[2] or time explorer[3]). The Temporal Web Analytics workshop series (TempWeb[4]) has been launched as a forum for such topics.

Archives of timestamped Web pages host a wealth of information, providing a gold mine for sociological, political, business, and media analysts. For example, one could track and analyze public statements made by representatives of companies such as SAP or Oracle, characterizing the evolution of their attitude towards green IT. Another example could be tracking, over a long time horizon, a politician's public appearances: which cities has she/he visited, which other politicians or business leaders has she/he met, etc.

Tracking entities on the Web or in Web archives involves finding names of people, companies, products, songs, etc. in Web pages and social media. However, names are often ambiguous. The disambiguation of named entities in natural language text needs to map mentions onto canonical entities allows a semantically exploit Web data. For that purpose, mentions of people, places, or organizations need to be raised to the entity level. This entity information is a great asset for making sense of the raw and often noisy data.

LAWA is a focused research project (STREP) funded by the European Union since September 2010, which addresses these challenges on temporal analytics of Web contents as part of the Future Internet Research and Experimentation (FIRE[5]) initiative. Its consortium consists of a total of six partners: Max Planck Institute for Informatics (Germany), Hebrew University (Israel), Hungarian Academy of Sciences (Hungary), University of Patras (Greece), Internet Memory Foundation (France) formerly called European Archive, and Hanzo Archives Ltd. (UK). The latter two are professional archival organizations, one being a non-profit foundation and the other one being commercial. The project investigates temporal Web analytics with respect to semantic and scalability issues [12]. To appreciate the latter, let us merely point out that the Internet Archive currently holds more than 150 Billion versions of Web pages, captured during the timeframe from 1996 until now. Its coverage is getting sparser as Web contents has become so diverse, dynamic, and humongous. A high-coverage archive would have to be an order of magnitude larger. Therefore, research questions of LAWA include:

- *Web Scale Data Provisioning*: Development of methods

*http://www.lawa-project.eu
[1]http://truthy.indiana.edu/

WWW 2012 Companion, April16-18,2012, Lyon, France.S
ACM 978-1-4503-1229-5/12/04.

[2]https://www.recordedfuture.com/
[3]http://fbmya01.barcelonamedia.org:8080/future/
[4]http://temporalweb.net/
[5]http://cordis.europa.eu/fp7/ict/fire/

for large scale, on demand crawling services that can be used for research purposes, and optimized Web content storage for processing.

- *Distributed Storage and Computations*: Support of computations required on summaries of data gathered, their distribution across the network (as opposed to executed inside the data centre) and development of distributed indexes for heterogeneous data stores.
- *Web Analytics*: Development of algorithms and software for systematically aggregating, querying, mining, and analyzing statistical patterns, cross-data dependencies, and temporal variabilities, in order to reveal latent knowledge in Web sources.

In this paper, we spotlight ongoing research in the area of entity analytics and will present initial results.

2. ENTITY DISAMBIGUATION

Disambiguating named entities in natural language text maps mentions of ambiguous names onto canonical entities like people or places, registered in a knowledge base such as DBpedia [2] or T-YAGO [11, 4]. Within AIDA (Accurate Online Disambiguation of Named Entities), we developed a method that unifies prior approaches into a comprehensive framework that combines three measures: the prior probability of an entity being mentioned, the similarity between the contexts of a mention and a candidate entity, as well as the coherence among candidate entities for all mentions together. Key contributions compared to the previously published our work are spatial and temporal extensions for improved named entity extraction and disambiguation.

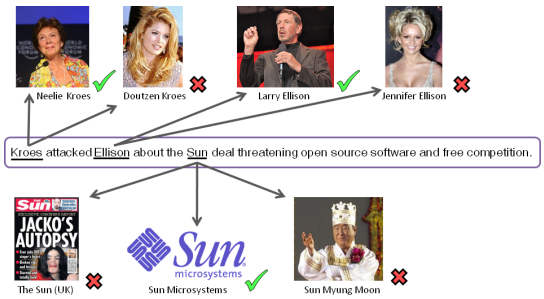

Figure 1: Example of entity disambiguation

We aim to identify surface strings representing named entities and map them to their proper entries in a knowledge base, thus giving a disambiguated meaning (cf. figure 1 for a graphical example of entity disambiguation). Given the previous example and surface strings "Kroes", "Ellison" and "Sun" extracted with the help of Stanford NER Tagger [3], we now aim to disambiguate the text mentions using coherence by combining three different disambiguation measures:

1. *Prior*: How often did other entities link to this entity in Wikipedia?
2. *Similarity*: How good do entity keyphrases and the text context overlap?
3. *Coherence*: Are the disambiguated entities related?

Since, each entity has a context in the underlying knowledge base(s): other entities that are connected via semantic relationships (e.g., memberOf) or have the same semantic type (e.g., politician). An asset that knowledge bases like DBpedia and T-YAGO provide us with is the same-as cross-referencing to Wikipedia. This way, we can quantify the coherence between two entities by, e.g., the overlap among their related entities or some form of type distance. In addition, we have exploited the spatial and temporal information to better disambiguate named entities. The spatial distance between two named entities with geo-coordinates is defined as the normalized great circle distance, while the temporal coherence of two named entities is defined as the difference of the center points of the entity's existence time interval, normalized by the maximum distance of any two entities in the current set of entity candidates.

3. TEMPORAL FACT EXTRACTION

The world is highly dynamic and nothing lasts forever! Knowledge about entities evolves over time, and many facts are fairly ephemeral, e.g., winners of sports competitions, and occasionally even CEOs and spouses. In addition, many information needs by advanced users require *temporal knowledge* [9, 5, 7]. For example, consider the following example question: "When did Dietmar Hopp found SAP and when did he leave the company?" Such a question is not being supported by existing knowledge bases. The problem we tackle is to automatically distil, from news articles and biography-style texts such as Wikipedia, *temporal facts* about entities for a given set of relations. By this we mean instances of the relations with additional time annotations that denote the validity point or span of a relational fact. For example, for the *wasCreatedOnDate* relation between people and companies, we want to augment facts with the time points of the respective events; and for the *worksForCompany* relation between business people and companies, we would add the timespan during which the fact holds. This can be seen as a specific task of extracting ternary relations, which is much harder than the usual information extraction issues considered in prior work.

Our system called PRAVDA (label PRopagated fAct extraction on Very large DAta) gathers fact candidates and distills facts with their temporal extent based on a new form of *label propagation (LP)* [10]. This is a family of graph-based semi-supervised learning methods, applied to (in our setting) a similarity graph of fact candidates and textual patterns. The system consists of four components: *candidate gathering*, *pattern analysis*, *graph construction*, and *label propagation*. The components are invoked in four phases.

1. *Candidate Gathering*: This phase serves to collect potentially relevant sentences. A sentence is interesting if it contains entities in relevant types for a target relation of interest. We employ a test for the potential presence of a base fact in a sentence, by checking for the existence of two noun phrases (denoting two entities) in the same sentence. In addition, we test for the existence of a temporal expression (currently only explicit date) in the sentence, thus producing raw input also for temporal fact candidates.

2. *Pattern Analysis*: Based on a (small, manually crafted) set of seed facts for a particular relation (either base or temporal), seed patterns in the form of sets of word-level n-grams are extracted from the interesting sentences. For each target relation that we aim to extract, a statis-

tical prior is then computed based on how strongly fact candidates have evidence in the form of seed patterns.

3. *Graph Construction*: To facilitate the knowledge harvesting procedure, the fact candidates and pattern candidates are represented as vertices of a graph. Edges represent the evidence connection between a pattern and a fact candidate, as well as the similarity between two patterns.

4. *Label Propagation*: Finally, a graph-based semi-supervised learning method is employed to harvest both base facts and temporal facts. To this end, we have developed an extended label propagation algorithm.

4. DEMO PRESENTATION

The actual demo consists of AIDA (**entity disambiguation**) and T-YAGO (**querying and visualization of temporal facts**). Users may freely interact with our systems.

Entity Disambiguation. The AIDA demo allows users to type in their own texts with ambiguous names. This may be a short and difficult text, like the one we have as example earlier; or it can be a news article, blog posting, or content from a discussion forum that is copy-and-pasted into AIDA [13]. Users can choose different methods, vary their configurations, explore the effects in terms of candidate entities, the weighted graph of mentions and entities, and the output quality. Figure 2 shows such an example using our graph-based disambiguation method. In our example: "Kroes attacked Ellison about the Sun deal threatening open source software and free competition." The prior-only and the prior-and-similarity methods may make mistakes by mapping "Sun"to the English newspaper instead of the software company, and "Kroes" to the Dutch model and actress Doutzen Kroes instead of the EC Vice-President.

Querying and Visualization of Temporal Facts. To facilitate querying of YAGO [4], each **S**ubject, **P**redicate, and **O**bject triple has been enhanced along by the dimensions **T**ime, **L**ocation, and conte**X**t. This yields a 6-tuple representation, which we call *SPOTLX*. SPOTLX is obtained by joining in the occurrence spans of facts, the locations of facts, and the context of the involved entities. Our interface shows a SPARQL-like query form for SPOTLX tuples. If results contain events or entities with an associated timespan, they are visualized on an interactive map with a time line. When the query changes, the information is updated. Users may also zoom and scroll in the map or timeline, which updates the query. For the visualization, we use TimeMap (`http://timemap.googlecode.com`). Combining the new dimensions, we can solve different kinds of queries. Adding the time dimension, we can query for companies founded in Germany since 1970:

```
?c type company .
?f created ?c after 1970
```

We can refine this query by adding a restriction on the location where the company was founded, e. g. `nearby Stuttgart` and dealing with `Software`. Our interface visualizes the retrieved entities on a timeline accompanied by a map (Figure 3). In the screenshot, we have zoomed in on Germany's state of Baden-Württemberg, which is also called Musterländle (coll. for little model state). The markers show locations where companies have been founded since 1970.

Acknowledgements
This work is supported by the 7^{th} Framework IST programme of the European Union through the focused research project (STREP) on Longitudinal Analytics of Web Archive data (LAWA) under contract no. 258105. We are grateful to our colleagues and project co-investigators András Benczúr, Scott Kirkpatrick, Philippe Rigaux, Peter Triantafillou, and Mark Williamson.

5. REFERENCES

[1] Omar Alonso, Michael Gertz, and Ricardo A. Baeza-Yates. On the value of temporal information in information retrieval. *SIGIR Forum*, 41(2):35–41, 2007.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In *ISWC/ASWC*, pages 722–735, 2007.

[3] Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, 2005.

[4] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard de Melo, and Gerhard Weikum. Yago2: exploring and querying world knowledge in time, space, context, and many languages. In *WWW (Companion Volume)*, pages 229–232, 2011.

[5] Xiao Ling and Daniel S. Weld. Temporal information extraction. In *AAAI*, 2010.

[6] Julien Masanès. *Web archiving*. Springer, 2006.

[7] Pawel P. Mazur and Robert Dale. Wikiwars: A new corpus for research on temporal expressions. In *EMNLP*, pages 913–922, 2010.

[8] Jean-Baptiste B. Michel, Yuan Kui K. Shen, and Aviva Presser P. Aiden et al. Quantitative analysis of culture using millions of digitized books. *Science (New York, N.Y.)*, 331(6014):176–182, 2011.

[9] Jannik Strötgen and Michael Gertz. Timetrails: A system for exploring spatio-temporal information in documents. *PVLDB*, 3(2):1569–1572, 2010.

[10] Yafang Wang, Bin Yang, Lizhen Qu, Marc Spaniol, and Gerhard Weikum. Harvesting Facts from Textual Web Sources by Constrained Label Propagation. In *Proceedings of the* 20^{th} *ACM Conference on Information and Knowledge Management (CIKM), Glasgow, Scotland, UK, October 24-28, 2011*, pages 837–846, 2011.

[11] Yafang Wang, Mingjie Zhu, Lizhen Qu, Marc Spaniol, and Gerhard Weikum. Timely yago: harvesting, querying, and visualizing temporal knowledge from wikipedia. In *EDBT*, pages 697–700, 2010.

[12] Gerhard Weikum, Nikos Ntarmos, Marc Spaniol, Peter Triantafillou, Andras Benczúr, Scott Kirkpatrick, Philippe Rigaux, and Mark Williamson. Longitudinal Analytics on Web Archive Data: It's About Time! In *Proceedings of the* 5^{th} *biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, CA, USA, January 9-12*, pages 199–202, 2011.

[13] Mohamed Amir Yosef, Johannes Hoffart, Ilaria Bordino, Marc Spaniol, and Gerhard Weikum. Aida: An online tool for accurate disambiguation of named entities in text and tables. In *Proc. of the* 37^{th} *Intl. Conference on Very Large Databases (VLDB 2011), August 29 - September 3, Seattle, WA, USA*, pages 1450–1453, 2011.

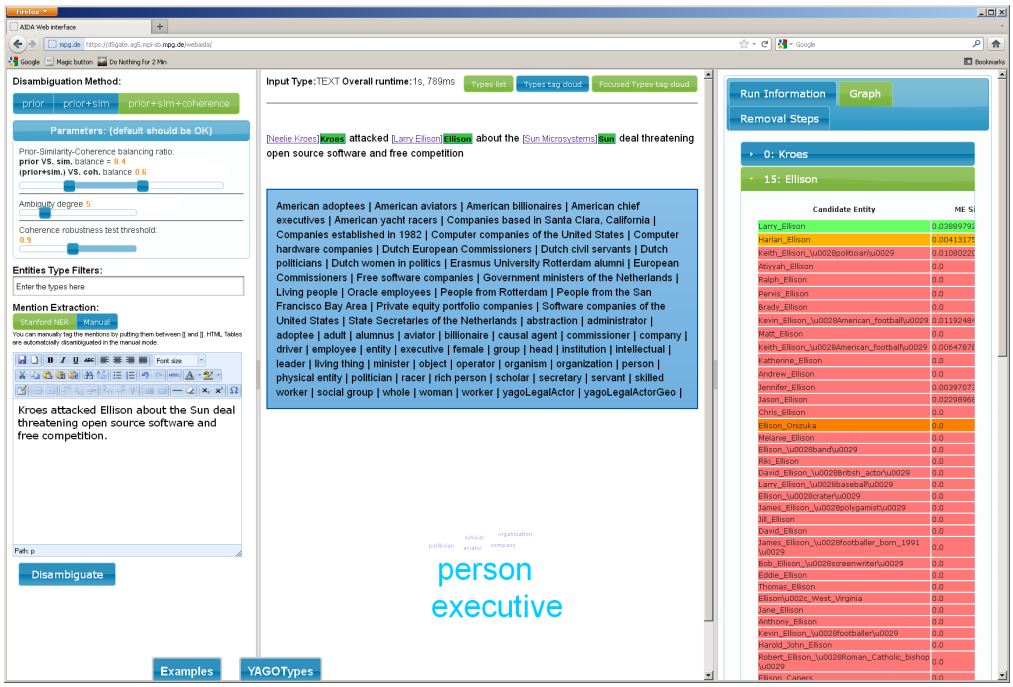

Figure 2: AIDA user interface for entity disambiguation

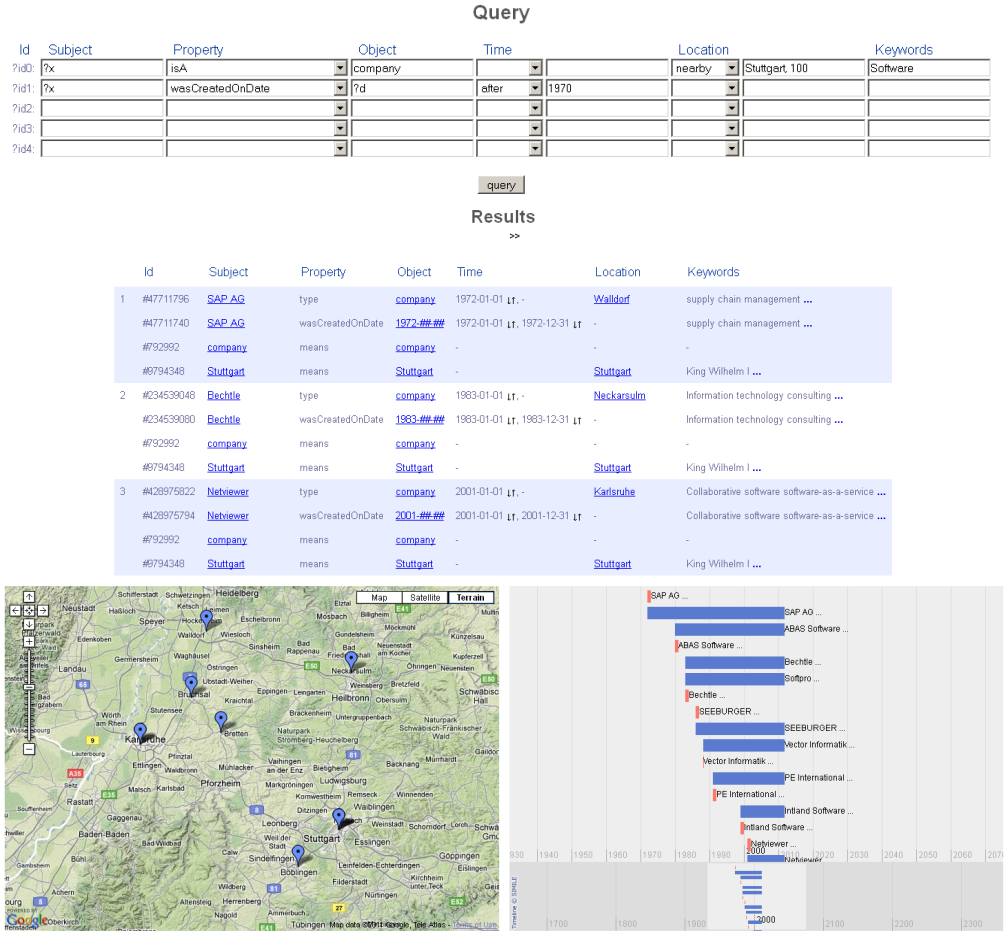

Figure 3: YAGO visualization of software companies founded since 1970 in Baden-Württemberg, Germany