Entity Timelines: Visual Analytics and Named Entity Evolution

Arturas Mazeika Tomasz Tylenda Gerhard Weikum
Max-Planck-Institute for Informatics
Building 46.1, Campus E1 4, 66123 Saarbrücken, Germany
{amazeika, ttylenda, weikum} @mpi-inf.mpg.de

ABSTRACT

The constantly evolving Web reflects the evolution of society. Knowledge about entities (people, companies, political parties, etc.) evolves over time. Facts add up (e.g., awards, lawsuits, divorces), change (e.g., spouses, CEOs, political positions), and even cease to exist (e.g., countries split into smaller or join into bigger ones). Analytics of the evolution of the entities poses many challenges including extraction, disambiguation, and canonization of entities from large text collections as well as introduction of specific analysis and interactivity methods for the evolving entity data.

In this demonstration proposal[1], we consider a novel problem of the evolution of named entities. To this end, we have extracted, disambiguated, canonicalized, and connected named entities with the YAGO ontology. To analyze the evolution we have developed a visual analytics system. Careful preprocessing and ranking of the ontological data allowed us to propose wide range of effective interactions and data analysis techniques including advanced filtering, contrasting timeliness of entities and drill down/roll up evolving data.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Measurement, Experimentation

1. INTRODUCTION

Which companies have been associated with mobile music devices in the last decade? What are the similarities and the differences between Mikhail Gorbachev and Ronald Reagan? What are the key people associated with the European countries? What are the similarities between the Baltic States and Germany in 1998, and how do the similarities evolve over time? How does the news coverage of a particular product, event, or entity compare to blog and tweet coverage? Knowledge extraction from large text collections spanning long time intervals, organization of knowledge, and answering the above and similar questions is the key topic of this demo.

Knowledge extraction, summarization, and development of a (visual) analytic system impose a number of challenges. First, the information is spread over many articles (news, Web, blogs, tweets) and thus the information must be extracted and aggregated. Second, entities need to be identified in order to move them from lexical to semantic level. Tracing entities along the temporal dimension is possible only if their semantics is taken into account. Simple keyword tracking is insufficient to connect semantically similar entities from different time periods, e.g., Walkman and iPod. Third, named entities must be properly disambiguated (George Bush may mean the 41st or the 43rd president of the U.S., Waterloo can be the location, the battle, or the song). Fourth, due to accumulation and contrasting functionality a single, unique identifier should be used to denote the same named entity (Michail Sergejevic Gorbachov and Mikhail Gorbachev should be replaced with, e.g., `Mikhail_Gorbachev`). Finally, an efficient visual analytics system with rich interactive capabilities must be developed to allow the analyst to interact, filter, view, project, compare (contrast), and group over the evolving named entity data.

Information presentation and analytics tools have been proposed in the information retrieval and ontology visualization (e.g., YAGO browser [5]) and visual analytics communities (e.g., Multiple Facets [3]); there are commercial system that link news articles based on keywords, tags, and topics (e.g., Google News). Ontology visualization systems aim to present the ontological knowledge to the user. Since ontologies form graphs, the data is usually presented as a graph with a rich interactive capabilities to show/hide selected parts of the subgraph or identify a handful number of the most relevant nodes. Evolution and contrasting of named entities is usually not considered. Conversely, text visualization community proposed tools to visualize evolving statistics of (key-)words in the text collections. Since only words rather than named entities are considered in such visualizations, the visualizations cannot benefit from the comprehensive type system offered by ontologies (e.g., `Albert_Einstein` is a `Theoretical physicists`→`physicist` → `scientist`→ `person`→ `entity`; `Lithuania` → `Baltic_State` → `European_State`). Such systems therefore, are limited to basic interactions with the data. Article linkage systems aim to group, recommend, and cross link articles with other data sources (e.g., Wikipedia). The extraction (and disambiguation) is limited to the most prominent named entity(s), while interaction is restricted to simple browsing of the articles with the help of links.

[1]A preview of the system is available at http://evolution.mpi-inf.mpg.de/timelines/

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

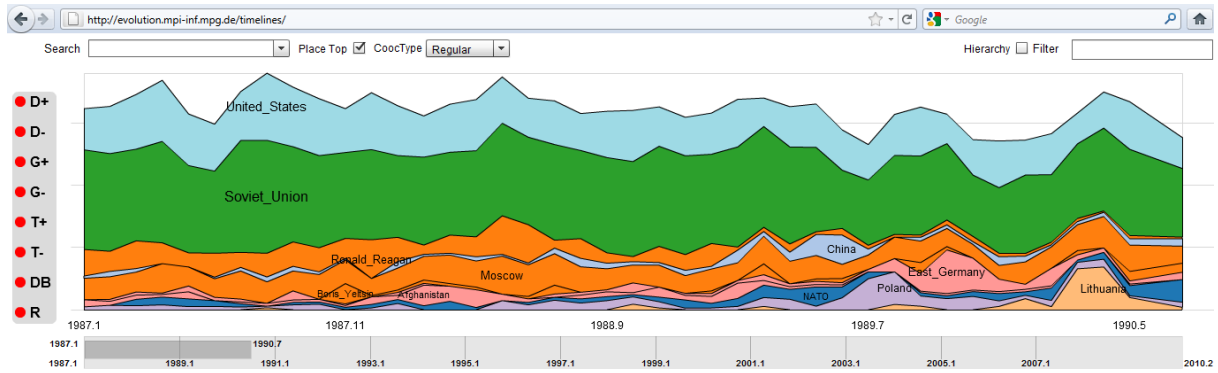

Figure 1: Screenshot of Entity Timelines

We processed the NYTimes collection of news articles to discover and track the evolution of named entities. The collection consists of more than 1.8M news articles published from 1987 to 2007; the extracted DB contains 50M named entities. For a given entity, evolving timelines aggregate named entities co-occurring in the articles and allow to investigate its evolution with the help of stacked displays. Tight connection to the YAGO ontology allows us to both effectively extract named entities and offer a rich data interaction capabilities. Analysts can look at different projections of the data (e.g., evolving timelines for `George_W._Bush` and `Mikhail_Gorbachev`), contrast them (e.g., investigate the timeline of `George_W._Bush` where all named entities from `Mikhail_Gorbachev` are removed from the visualization), drill down/roll up the data based on the YAGO hierarchies (e.g., aggregate individual politicians into left, center, and right wing politicians, etc), or use advanced filtering techniques (include named entities that are a part of a selected grouping or related to a given keyword).

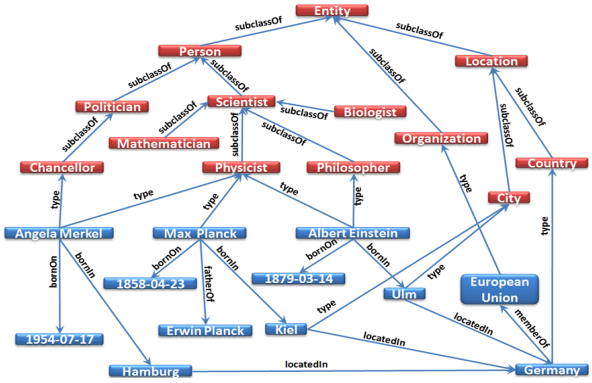

Figure 2: An illustrative example of the YAGO knowledge base

2. YAGO ONTOLOGY

The YAGO ontology [4] is a comprehensive database of human knowledge. It consists of facts extracted from Wikipedia stored in machine readable format. The ontology includes information on millions of individual named entities (e.g., `Max_Planck`, `European_Union`, and `Berlin`), hundred of millions facts about them (e.g., `Max_Plank hasNationality Germany`, `bornOn` 1858-04-24), has hundred thousands of classes/types (e.g., `Politician` is a subclass of `Person`), and has an overall accuracy of around 95%. The ontology forms a directed graph (Figure 2) and can be used for numerous information retrieval and data analysis tasks. In this paper we use the link structure between the entities for named entity disambiguation (the bottom of the figure) and the subclass hierarchy for advanced grouping and filtering (the top of the figure). For these tasks any general purpose ontology like DBpedia or FreeBase can be used. However, since entities were extracted from a large text corpora (and therefore, small named entity disambiguation errors could accumulate into large overall errors), we opted for an ontology with a (probably) lower coverage but very high precision.

3. NAMED ENTITY EXTRACTION, DISAMBIGUATION, AND CANONICALIZATION

Named entities are predefined categories such as the names of persons (e.g., Mikhail Gorbachev), organizations (United Nations), locations (Moscow), expressions of times (Morning), quantities (a pound), monetary values (a buck), percentages (a half), etc. Extraction of named entities is a tough problem due to different spelling and errors (Michail vs. Mikhail), different formatting (Gorbachev, Mikhail) and different meanings (pound may refer to the mass or currency), or they may be vaguely expressed and therefore hard to identify (capital of the country). A number of different attempts exist to extract named entities from (English language) texts using grammar based or statistical methods. Our work is the first attempt to perform the extraction and analysis on a large scale data and connect them to the entities in an ontology.

Conceptually, our named entity finder processes the English text in five stages: extraction (stage 1 and 2), disambiguation (stage 3 and 4) and canonicalization (stage 5). In the first stage we identify words that are interesting and may be a part (substring) of the named entity. This can be done using a dictionary, pattern matching, or the Stanford Entity Recognizer (NER). In the second stage we identify all super-strings that the interesting word is a part of and are either YAGO named entities or found in the Wikipedia anchor texts pointing to articles describing YAGO entities. This produces a list of candidates, where each interesting word is mapped to one or more named entities. In the third stage we identify the interesting words that have only one named entity associated with it (i.e., do not require disambiguation). We call them *final entities*. In the fourth step we filter out and disambiguate the remaining named entities with the help of the final entities and the YAGO ontology. The disambiguation score of a candidate entity increases if there is a link between the candidate and the final entity in the ontology. Fi-

nally, for each interesting word we pick the entities with the highest score.

As an example, consider sentence "At Waterloo Napoleon faced his final battle against Wellington and Blücher". With the identified and extracted named entities, the sentence becomes "At Battle_of_Waterloo Napoleon_I_of_France faced his final battle against Arthur_Wellesley,_1st_Duke_of_Wellington and Gebhard_Leberecht_von_Blücher". The words "Waterloo, Napoleon, Wellington, Blücher" are identified as interesting. Battle_of_Waterloo, Waterloo,_Ontario, Waterloo_(ABBA_song), Waterloo,_Iowa and other 207 named entities form a set of candidates for Waterloo. Similarly, Wellington, Wellington_Rugby_Football_Union, Wellington_Firebirds, Arthur_Wellesley,_1st_Duke_of_Wellington and other 205 named entities form a set of candidates for Wellington. YAGO link structure between the candidates allows us to identify Battle_of_Waterloo and Arthur_Wellesley,_1st_Duke_of_Wellington as the best matching named entities.

4. ENTITY TIMELINES

We visualize the evolution of named entities as a timeline using the paradigm of stacked areas [3] and river metaphor [1]. Given a named entity as a query, we visualize other entities co-occurring in the same Web page, document, or tweet as stacks (polygonal areas) in XY coordinates. The time is mapped to the X coordinate, while the Y coordinate (the area of the stack) depicts how prominent the associated named entity is to the query.

Figure 1 shows a screen shot of the entity timeline for Mikhail_Gorbachev (i.e., the area of the stack depicts the frequency of the co-occurring entities in the articles over time). The tool clearly identifies key events and their relevance to the named entity. For example, Mikhail_Gorbachev is associated mostly with Soviet_Union and United_States in time window 1987–1990. At the beginning of the time window Ronald_Reagan is prominently seen in the timeline, since both politicians were the key figures in the ending of the Cold War. Further, one can see the end of Ronald_Reagan's and the beginning of the George_W._Bush presidency; activity of East_Germany (fall of Berlin Wall and reunification of Germany), followed by the independence of Lithuania and the other Baltic states. Using the visual analytics tool one can also see such events as the collapse of the Soviet_Union, the increase of the weight of Russia, the rise of new politicians (presidents Boris_Yeltsin and Vladimir_Putin), 9/11, diagnose of the Alzheimer's disease and death of Ronald_Reagan, and raise of political influence of the Russian oil company Gazprom.

The tool allows numerous interactions with the visualization and multiple displays over the data. Features of the tool span from basic navigational interactions, like selecting timelines and windows for different named entities, to more advanced, like contrasting or drill down/roll up of timelines. Both single (one timeline) and multiple linked displays (two timelines aligned vertically) are supported. Due to a more efficient use of space, single display visualizations are more appropriate for detailed analysis of an evolution of one named entity, while linked, side-by-side displays are more appropriate for comparisons and contrasting.

4.1 Contrasting Timelines

Pre-processing of the documents and canonicalization of named entities allows effective comparison and contrasting of two timelines. Let us consider the timelines of Jacques_Chirac (President of France during 1995–2007) and Gerhard_Schröder (Chancellor of Germany and the leader of the Social Democratic Party during 1998–2005). Both timelines share a significant number of named entities, e.g., Russia, Kosovo, UK, and NATO (Figure 3, top), indicating the similarities of German and French politics. However there are a number of entities that are prominent in Chirac's timiline but not in Schröder's (e.g., Austria, Palestine, Libya, and Syria, see the middle in the figure), and prominent in Schröder's but not in Chirac's (Saxony, Social_Democrats, VW, and Tony_Blair) The contrasting functionality allows to emphasize these similarities and differences in one visualization.

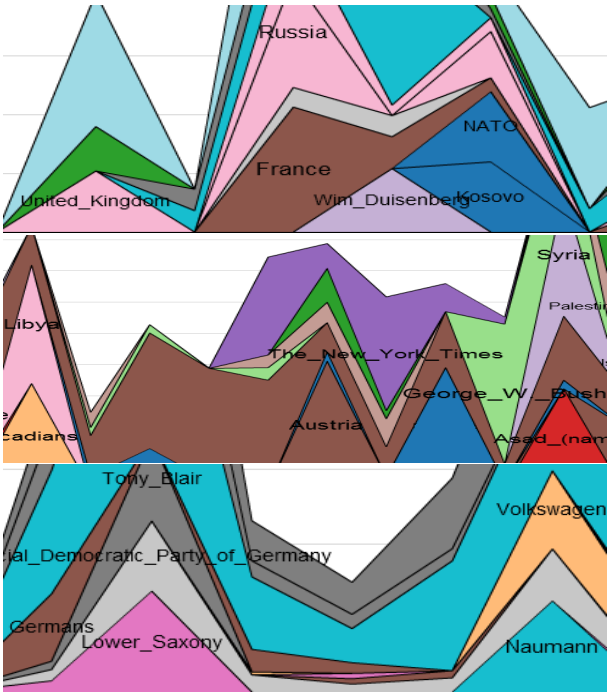

Figure 3: J. Chirac vs. G. Schröder: intersection (top), Chirac\Schröder (middle) and Schröder\Chirac (bottom)

Conceptually, contrasting functionality takes two sets of time series (e.g., the time series of Chirac and the time series of Schröder) and performs a general set operation over the series applying a "join" function to the counts. This can be the set union (all tuples from both subsets are grouped by co-occurring named entity and month, where the corresponding counts are summed up), set intersection (the minimum or maximum is taken from the corresponding counts), set difference (the count of one is subtracted from the other), etc. Since the ranges of the counts may vary significantly (e.g., the timeline of a popular entity may exhibit counts in the range of thousands, while the timeline of a less popular entity be in the range of tens) we rescale the counts for a better comparison.

Contrasting timelines can be visualized in single or multiple displays. In a single display the contrast (union, xor, etc.) is shown, while in multiple displays the standard timelines are placed side-by-side. Multiple displays share the visual style: the colors, order of the co-occurring entities and the position and the size of time window are the same. Moving the mouse over a stack area highlights the corresponding stacks in the timelines of both entities.

4.2 Semantic Drill Down/Roll Up Using Type Hierarchies

YAGO ontology has a comprehensive type system and it stores a rich set of attributes of entities. For example, YAGO knows 450 facts about `Mikhail_Gorbachev`, 779 facts about `Albert_Einstein`, and 885 facts about `United_Arab_Emirates`. YAGO canonicalization of named entities facilitates easy integration of the information systems and the use of the information available in the ontology.

We use the YAGO type hierarchies to define and compute semantic drill down and roll up over the named entity domain. Similar to the horizontal drill down/roll up, where for each named entity the counts of a range of months were summed up, semantic drill down/roll up sums up the counts of named entities belonging to the same group for each month. For example, since according to the YAGO ontology `Mikhail_Gorbachev` is a member of the type hierarchy `politician` → `leader` → `person` → `entity`, therefore if the semantic roll up is applied, then the frequency (counts) of `Mikhail_Gorbachev` are added together with the frequency of the other `politicians`.

Semantic roll up for `Gerhard_Schröder` during (top) and after retirement from politics (bottom) is illustrated in Figure 4. Roll-up shows the change: previously Schröder was affiliated with `country` and `state`, while afterwards groupings related to `companies` are more prominent.

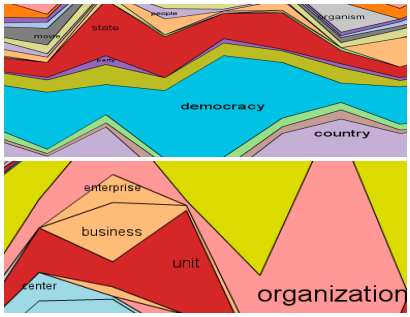

Figure 4: The roll up of `Gerhard_Schröder`

4.3 Ranking Type Hierarchies and Advanced Filtering

YAGO offers multiple ways to roll up/drill down hierarchies for almost all entities. For example, `Mikhail_Gorbachev` has the following

```
Soviet politician→politician
   →leader→person→entity
Russians of Ukrainian descent→person→entity
Russian atheist→atheist→disbeliever
   →nonreligious person→person→entity
Moscow State University alumni→alumnus
   →scholar→intellectual→person→entity
```

and other 15 hierarchies, `Watergate_scandal` has five, `Soviet Union` has 17, `David_Bowie` has 44, and `Albert_Einstein` has 94 type hierarchies. In general, the rich hierarchy system gives a possibility for customized and fine-adjusted aggregations. Since our timelines consists of multiple (thousands of) named entities, using all hierarchies of a named entity for grouping would result in a lattice that is similar to the one of multidimensional hierarchical heavy hitters [2]. Instead, here we aim to find *the best* hierarchy for each named entity and incorporate all these hierarchies into one (Figure 5).

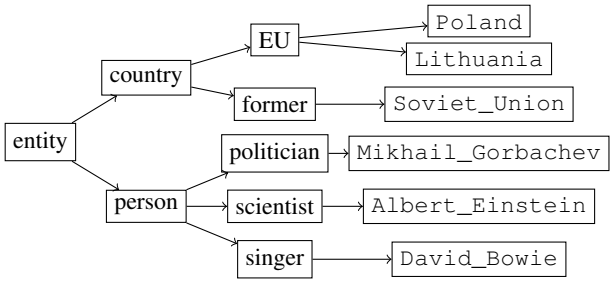

Figure 5: A (simplified) hierarchy for `Lithuania`, `Poland`, `Soviet_Union`, `Mikhail_Gorbachev`, `Albert_Einstein`, and `David_Bowie`

We identify the best hierarchy for a named entity using a number of heuristics. First, we investigated the evolution of Wikipedia categories of named entities in the corresponding Wikipedia articles and picked the hierarchies with the earliest and most long-lasting categories in the Wikipedia history. This allowed us to focus only on the most prominent hierarchies of the named entities. Then we carefully applied an additional filtering step using the term frequency scores of nouns and words of the abstract from the Wikipedia article (content up to the table of contents). We also tried to exploit the inverse document frequencies (selects very specialized paths like `People_from_Stavropol`→`person`→`entity` for `Mikhail_Gorbachev`) and number of named entities in the group of the hierarchies (suggests considerably different paths for similar named entities like `Poland` and `Lithuania`) but the results were less effective.

The computed hierarchy is used in semantic drill down/roll up as well as for advanced filtering. In the latter case the analyst may select a (number of) sub-hierarchies and filter out all named entities that fall outside the selected hierarchy.

Acknowledgments

This work is supported by the 7^{th} Framework ICT programme of the European Union through the large-scale integrating project on Living Knowledge. We thank K. Berberich, S. Bedathur, E. Lewis-Kelham, J. Hoffart, A. Anand, and M. Spaniol for valuable discussions and comments.

5. REFERENCES

[1] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.

[2] A. Mazeika, M. H. Böhlen, and D. Trivellato. Analysis and interpretation of visual hierarchical heavy hitters of binary relations. In *ADBIS*, pages 168–183. 2008.

[3] L. Shi, F. Wei, S. Liu, L. Tan, X. Lian, and M. X. Zhou. Understanding text corpora with multiple facets. In *IEEE VAST*, pages 99–106, 2010.

[4] F. M. S. , G. Kasneci, and G. Weikum. YAGO: a core of semantic knowledge. In *WWW*, pages 697–706, 2007.

[5] T. Tylenda, M. Sozio, and G. Weikum. Einstein: physicist or vegetarian? Summarizing semantic type graphs for knowledge discovery. In *WWW*, pages 273–276. ACM, 2011.