News Visualization Based on Semantic Knowledge

Sebastian Arnold, Damian Burke, Tobias Dörsch,
Bernd Loeber, and Andreas Lommatzsch

Technische Universität Berlin
Ernst-Reuter-Platz 7, D-10587 Berlin, Germany
{sarnold,damian.burke,tobias.m.doersch,bernd.loeber,
andreas.lommatzsch}@mailbox.tu-berlin.de

Abstract. Due to the overwhelming amount of news articles from a growing number of sources, it has become nearly impossible for humans to select and read all articles that are relevant to get deep insights and form conclusions. This leads to a need for an easy way to aggregate and analyze news articles efficiently and visualize the garnered knowledge as a base for further cognitive processing.
The presented application provides a tool to satisfy said need. In our approach we use semantic techniques to extract named entities, relations and locations from news sources in different languages. This knowledge is used as the base for data aggregation and visualization operators. The data operators include filtering of entities, types and date range, detection of correlated news topics for a set of selected entities and geospatial analysis based on locations. Our visualization provides a time-based graphical representation of news occurrences according to the given filters as well as an interactive map which displays news within a perimeter for the different locations mentioned in the news articles. In every step of the user process, we offer a tag cloud that highlights popular results and provide links to the original sources including highlighted snippets. Using the graphical interface, the user is able to analyze and explore vast amounts of fresh news articles, find possible relations and perform trend analysis in an intuitive way.

1 Introduction

Comprehensive news analysis is a common task for a broad range of recipients. To overlook the overwhelming amount of articles that are published in the Web every hour, new technologies are needed that help to classify, search and explore topics in real time. Current approaches focus on automated classification of documents into expert-defined categories, such as politics, business or sports. The results need to be tagged manually with meta-information about locations, people and current news topics. The simple model of categories and tags, however, is not detailed enough to suit temporal or regional relationships and it cannot bridge the semantic gap that the small subset of tagged information opens. The challenge for machine-driven news analysis consists of two parts. First, an extractor needs to be able to identify the key concepts and entities mentioned in the documents and to find the most important relationships between them. Second, an intuitive way for browsing the results with support for explorative discovery of relevant topics and emerging trends needs to be developed.

We present a semantic approach that abstracts from multi-lingual representation of facts and enriches extracted information with background knowledge. Our implementation utilizes natural language processing tools for the extraction of named entities, relations and semantic context. Open APIs are used to augment further knowledge (e.g. geo-coordinates) to the results. Our application visualizes the gained knowledge and provides time-based, location-based and relationship-based exploration operators. The relationship between original news documents and aggregated search results is maintained throughout the whole user process.

In Section 2, we give an overview on existing projects of similar focus. Our knowledge-based approach and the implementation is introduced in Section 3. The user interaction and visualization operators are discussed in Section 4. We conclude in Section 5.

2 Related Work

We start with an overview on existing projects in the field of semantic news visualization. The following projects are related to our approach on a conceptual or visual level.

MAPLANDIA[1] visualizes news for a specific date beginning in 2005 on a map. The system uses the BBC news feed as its only source to deliver markers and outlines for the countries that were mentioned in the news on a specified date. Additionally, it offers a list of the news for the day. However, by using only one marker the application is unable to visualize news on a more detailed and fine-grained level. MAPLANDIA also does not offer any possibility to limit the displayed visualizations to a certain region of interest. The application offers news in only one language and source for a specific day.

The idea behind the SPIGA-SYSTEM [3] is to provide businesses with a multilingual press review of news from national and international sources. Using the Apache UIMA framework, the system crawls a few thousand sources regardless of the language used. After a fitting business profile has been created, the system clusters information and visualizes current trends.

3 Implementation of Knowledge Extraction

In this section, we explain our semantic approach to news aggregation. In contrast to classical word- or tag-based indexing, we focus on semantic features that we extract from daily news documents. To handle the linguistic complexity of this problem, we utilize information extraction techniques for natural language [2]. The knowledge extraction pipeline is shown in Fig. 1. It consists of a periodic RSS feed crawler as source for news documents[2] and the language components for sentence splitting, part-of-speech (POS) tagging, named entity recognition (NER) and coreference resolution. We utilize a Stanford CoreNLP named entity recognition pipeline [1] for the languages English and German. The pipeline periodically builds histograms over the frequency of named entity occurrences in all documents. Using a 3-class entity typification (*Person*, *Organization*, *Location*) we apply special treatment to each of the entity types.

[1] `http://maplandia.com/news`

[2] In our demonstrator, it is configured to use feeds from `http://www.theguardian.com`

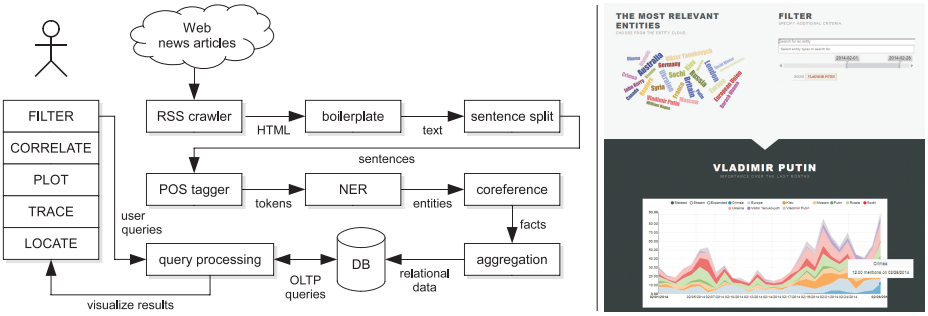

Fig. 1: The figure shows the architecture and a screenshot of our system. The system architecture is divided in user operators FILTER, CORRELATE, PLOT, TRACE and LOCATE (shown left) and the knowledge extraction pipeline (shown right). The screenshot of the Web application visualizes news entities related to *Vladimir Putin* utilizing the operators CORRELATE, FILTER and PLOT.

Person and *Organization* names are normalized to obtain representative identifiers from different spellings. *Location* names are resolved to geo-location coordinates using Google Maps API.[3] The results are processed and aggregated into relations of the form $\mathit{MentionedIn}(\text{ENTITY}(\mathit{type}), \text{DOCUMENT}(\mathit{date}), \mathit{frequency})$. To get a comprehensive view on the relevant information, we create histograms over these relations and store the distributions in a relational database. The data is processed using relational statements to fit the type of user query that is requested from the frontend. To increase processing performance of the information in Web-scale, we partly precompute these statements. The results are then visualized to allow easy recognition of trends and relationships.

4 Demonstrator and Benefits to the Audience

In this section, we present our user interface which has a specific focus on a simple workflow.[4] Our aim is to relieve the user from having much work with sorting and filtering tables and instead allow exploratory search [4] inside the data set. We present the results in a clean web interface that establishes an interaction flow from top to bottom. Our system offers several operators to explore the temporal, geographical and relational distribution in the news corpus.

A user session starts with the FILTER operator that is visualized as a tag cloud showing the most mentioned entities in a given time range in a size proportional to their frequency (e.g. location *Russia*). The selection can be influenced by further filter settings, such as type and date restrictions. Clicking on an entity within the tag cloud will trigger the CORRELATE operator, which offers related entities to the selected one in the tag cloud (e.g. location *Ukraine*, person *Vladimir Putin*). This is done by intersecting the most relevant documents for a given entity and time range and picking the top mentioned named entities in these articles. Selecting different items will further narrow down the

[3] `https://developers.google.com/maps/`

[4] An online demo is available at `http://irml-lehre.aot.tu-berlin.de`

results. Both the selected and the correlated entities are then displayed in a time-based PLOT with the time range on the x-axis and the frequency of occurrences on the y-axis. To instantly reveal the relationships and importance of co-occurrent entities, one can modify the display style (e.g. stacked, expanded or streamed). To get more detailed information about specific data points, the user can hover the cursor above them to trigger a TRACE operation. Then, more details about the selected tuple (ENTITY, *date*) are revealed: counts and snippets of the news articles that mention the selected entity and links to trace back the original documents.

The LOCATE operator focuses on geographic locating of selected entities and their relations. The operator works by computing a set of bounding coordinates which are used to query the database for possible locations. Using the same FILTER interface, a location of interest can be selected from the tag cloud or by entering its name in the text field. By utilizing a slider to set a search perimeter, the user is able to further focus on the regions around the selected location. After selection, a world map will display locations mentioned in the matching articles. By clicking the markers on the map, a balloon listing shows up to ten headlines and links to the respective articles. This allows the user to gain an overview of connections and associations of different countries and locations.

5 Conclusion and Future Work

The application allows the user to quickly visualize and analyze vast amounts of news articles. By the use of interactive elements such as graphs and a world map the user is able to check hypotheses and draw conclusions in an explorative and playful manner. This greatly reduces the cognitive load for the user as he or she is able to find the relevant facts fast and browse the underlying news articles to get further information from the original source. In our conducted user studies we observed that a streamlined interface with the options at the top and the results below was most appealing to users, and fewer options led to a more intuitive experience. The presented application is realized as a prototype and will be expanded by further development. A broader range of information can be achieved by including more news sources, implementing extended language support (e.g. multi-lingual knowledge extraction) and expanding the features of the FILTER operator (e.g. including sentiment selection). A deeper enrichment of knowledge can be achieved by linking the detected entities to additional knowledge sources (e.g. DBpedia or freebase) and using context information to extract more language features (e.g. sentiments, quotations, relations).

References

1. D. Cer, M.-C. de Marneffe, D. Jurafsky, and C. D. Manning. Parsing to stanford dependencies: Trade-offs between speed and accuracy. In *7th Intl. Conf. LREC 2010*, 2010.
2. R. Grishman. Information extraction: Techniques and challenges. In *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*. Springer, 1997.
3. L. Hennig, D. Ploch, D. Prawdzik, B. Armbruster, H. Düwiger, E. W. De Luca, and S. Albayrak. Spiga - a multilingual news aggregator. In *Proceedings of GSCL 2011*, 2011.
4. G. Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.