Network of the Day: Aggregating and Visualizing Entity Networks from Online Sources*

Darina Benikova, Uli Fahrer, Alexander Gabriel, Manuel Kaufmann,
Seid Muhie Yimam, Tatiana von Landesberger, Chris Biemann
Computer Science Department, TU Darmstadt, Germany
www.tagesnetzwerk.de

Abstract

This software demonstration paper presents a project on the interactive visualization of social media data. The data presentation fuses German Twitter data and a social relation network extracted from German online news. Such fusion allows for comparative analysis of the two types of media. Our system will additionally enable users to explore relationships between named entities, and to investigate events as they develop over time. Cooperative tagging of relationships is enabled through the active involvement of users. The system is available online for a broad user audience.

1 Introduction

The constantly growing interest in social media raises a need for new tools enabling wide audience to analyze and explore the available data. Our work addresses this need via the interactive online visual system *Network of the Day* (Netzwerk des Tages). It combines information extracted from the social media platform Twitter and online newspaper articles. *Network of the Day* offers a transparent exploration of current media to politically interested non-experts.

The visualization shows the most important current entities discussed in online media in a compact and interactive form. The presented data is kept up to date on a daily basis. We present the media data in several interlinked views. First, we extract and show the relationships between entities (i.e., persons and organizations) in a network. Interaction with this network enables the users to tag the relations between entities, which creates additional semantics in the data. Second, a line chart shows the occurrences of most popular entities for the respective day over the past months. This offers the possibility to spot the development of important topics over time. Third, this enables the user to compare commonalities and differences of the two media. Finally, the user can search for entities of her interest in order to gain information on media developments, which are of relevance to her.

2 Related work

Summarizing and extracting information from media databases has been a task of great interest in natural language processing, as the amount of information is too large to be processed by humans without automatic aids.

In recent years, the possibilities of opinion expression or social-media communication have increased, resulting in a surge of sentiment analysis tools (Pang and Lee, 2008). Especially there is a need for filtering and exploring events and opinions in high-volume social media data.

The visualization of social network data, Twitter data and news has gained importance. Several approaches have been developed. TextViz[1] provides an overview of text visualization techniques from various areas. Most relevant to our work are the visualization of word co-occurrence in Twit-

[1] http://textvis.lnu.se

ter messages and visualizations of relations between named entities. For example, Phrase Nets (Van Ham et al., 2009) show co-occurrence of words as a network, however they do not allow for exploring time dependent changes. On the contrary, Topic Competition (Xu et al., 2013) shows the development of word and topic frequencies over time. However, the relationships between topics and entities are not visible. A further relevant work by Biemann et al. (2004) shows paths through networks extracted from news. While this software is interactive, relations between entities cannot be labeled interactively and developments over time are not shown.

In this work, the social media communication is represented by the Twitter[2] platform. Meckel and Stanoevska-Slabeva (2009) investigated the reflexion of politics upon Twitter. *Twitterbarometer* [3] is a tool developed by the Buzzrank company which measures the political mood in real time by capturing tweets related to parties – as indicated by hashtags – and classifying them as positive or negative.

3 Description of main components

This section presents the main components of the project. We first describe the data sources, their deployment and their processing. We then present two main components of the project – the *Twitter contrast analysis* and *Network of Names*. These components form a basis of the new system presented in Section 4.

3.1 Data Sources

The data sources used in our system are online news from "Wörter des Tages" and online messages from Twitter.

3.1.1 Online News

The project "Wörter des Tages"[4] (Quasthoff et al., 2002) serves as our source of daily news articles. Frequently appearing words are extracted daily by a text mining suite from daily newspapers and news services.

The project "Wörter des Tages" extracts its data mostly from German online sites, resulting in a daily dataload of approximately 20,000 - 50,000 sentences. The texts are segmented and indexed, the terms are quantitatively acquired and statistically significant co-occurrences are computed. The main parameters for the term selection are the frequency in the current daily corpus, the frequency in the already mentioned reference corpus "Deutscher Wortschatz" and the factor of relative frequencies between the two corpora of the term (Quasthoff et al., 2002).

3.1.2 Twitter

We download Twitter data using its public Streaming API[5] that gives developers access to Twitter's global stream of Tweet data. This stream is filtered according to previous selected most important keywords, i.e. as extracted by (Quasthoff et al., 2002).

3.2 Basis Software Components

Two recent works form the basis of this project: Fahrer's implementation (2014) of a Twitter contrast-analysis, which shows words frequently co-occurring with search terms and the work of Kochtchi et al. (2014), which visualizes the relationships between people and organizations using online newspaper articles as a source. Both projects provide full provenance information, i.e. users are not only able to see and manipulate the display of automatically extracted relationships, but also to access the text sources from which the relationships are extracted.

3.2.1 Twitter contrast-analysis

The component by Fahrer (2014) provides a contrastive co-occurrence analysis that contrasts two separate keywords regarding their strongly associated words in Twitter messages. For example, Figure 1 shows a contrastive analysis for the keywords *Brüderle* and *Trittin*, who are prominent German politicians from two different parties. The left side of the graph shows words only co-occurring with the keyword Brüderle and the right side shows only co-occurring words with Trittin. The overlap in the middle indicates words that are co-occurring with both terms. Results

[2] http://www.twitter.com

[3] http://twitterbarometer.de

[4] http://www.wortschatz.uni-leipzig.de/wort-des-tages/

[5] https://dev.twitter.com/docs/api/

show that the overlap in the contrast analysis gives a sensible reflection of main political events. Furthermore, most of the relevant newspaper topics regarding the contrastive analysis are reflected in Twitter.

The data for a study on the German parliament election was collected from Twitter between August 2, 2013 and October 9, 2013. Overall a corpus of 10,524,367 Twitter messages was collected. For the tokenization, the Twitter tokenizer from Gimpel et al. (2011) was employed. To determine the words strongly co-occurring with a given word the log-likelihood measure (Dunning, 1993) was applied to rank the vocabulary according to descending values (Fahrer, 2014).

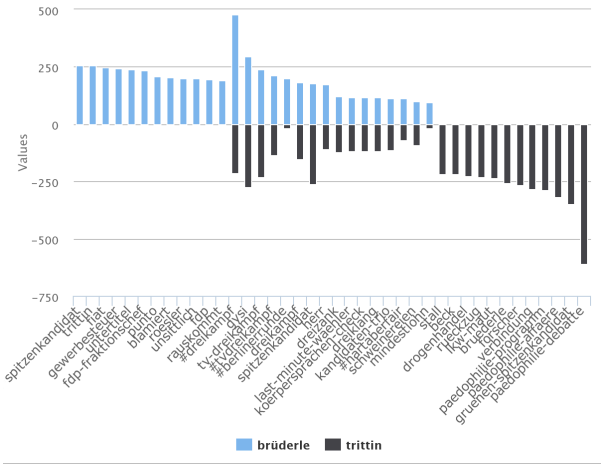

Figure 1: Sample contrastive analysis with the search terms “Brüderle” (light bars) and “Trittin” (dark bars) with 40 result terms, cf. (Fahrer, 2014)

3.2.2 Network of Names

The second basic component is the exploration of relationships between named entities presented by Kochtchi et al. (2014). This interactive system derives a social network graph from information extracted from online publications of newspaper articles.

The visualization enables to explore and investigate the relationships between people and organizations of public interest, reflecting the interaction between public protagonists and the influence of their surroundings, sociality and public policy. Kochtchi et al. (2014) used the Leipzig Corpora Collection (Richter et al., 2006), containing about 70 million of sentences extracted from German online newspapers between 1995 and 2010, as the text source of his project. In the course of preprocessing, Kochtchi et al. (2014) extracted Named Entities using the Stanford Named Entity Parser (Faruqui and Padó, 2010; Finkel et al., 2005) and calculated normalized PMI scores (Bouma, 2009) of co-occurrence. The Network of Names component offers the possibility of collaborative social tagging. By clicking on the edges between entities, users can enter a relation label of this relationship. The users base these labels on the sentences containing the two entities. The sentences are shown in an extra frame next to the relationship. While the Network of Names was a static visualization of a large corpus, we use parts of this technology to create daily networks and components display changes over time.

4 Combination of social-media and computer-mediated communication

The main goal of “Network of the Day” is to present current main topics and their relationship on the basis of combining *online news* and *social media*. The combination represents the contrast of the presentation of events by the German online media and the reaction to the situation of a part of the German online Twitter community.

Figure 2 illustrates the visualization for networks extracted from daily news. Our visualization comprises four main parts, which are interactively linked: daily network, social tagging, time line and twitter contrast analysis.

Networks are constructed on a daily basis, represent important events of the day, and can be visually compared to networks from the past. Each network shows the relationships between the most important persons and organisations of the day. Entities are nodes and their co-occurrence is denoted by edges. The user can select entities from the graph and their most important co-occurring terms over time. The network is clustered with the Markov Cluster Algorithm (van Dongen, 2000), and clusters can be unfolded and collapsed by clicking on them. Cluster labels are the most central three nodes within a cluster that are calculated using the Pagerank algorithm (Page et al., 1999). We use a flexible force-directed layout for

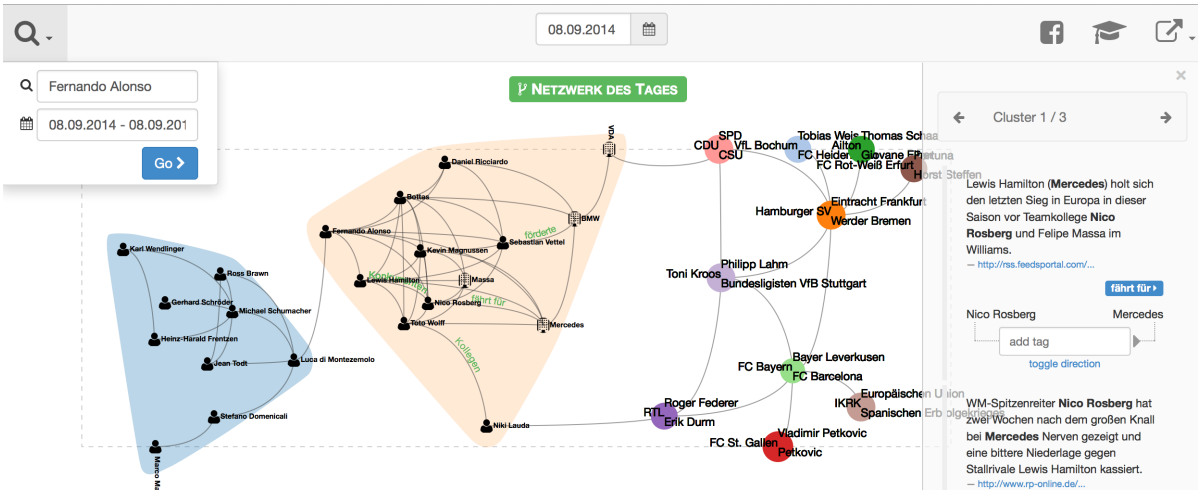

Figure 2: Visualization of a Network of the Day for September 8, 2014 after a search for ”Fernando Alonso”. Two clusters about motor sports are unfolded, the sources for the link between ”Nico Rosberg” and ”Mercedes” are shown and their relation is labeled as ”fährt für” (drives for).

the graph rendering that is implemented using the D3.js[6] JavaScript visualization library.

Clicks on links result in the display of source sentences, which are linked to the original online articles. Users can tag relationships of entities using the *interactive social tagging component*, see right side of Figure 2. Further, selecting an edge also invokes a contrast analysis of the two connected entities based on Twitter data, cf. Section 3.2.1 (not shown due to space constraints). The search mask allows the user to search for entities of her choice in arbitrary time spans, and to obtain a detailed analysis. This allows for user specific exploration of current and past social media.

The dynamics of word frequency over time is exemplified in Fig. 3 and displayed below the network. Initially, it shows terms that were popular on the respective day, but arbitrary terms from the network can be selected, and compared in the frequency diagram.

5 Outlook and Further work

Network of the Day offers a transparent aggregation of current media to laymen interested in politics and other daily affairs. Moreover, it offers them the possibility to collaboratively tag interesting relationships. Very importantly, the visualization provides full provenance, as original sources are linked.

[6] http://d3js.org/

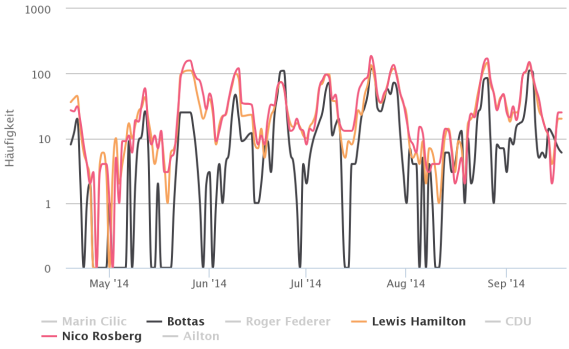

Figure 3: Frequency diagram of trending terms on September 8, 2014, reflecting the bi-weekly schedule of Formula 1 races.

By extracting the current information on relations, people, organizations and events from Twitter, the result of this project may be used in political education or serve voters as an overview. In this study only a comparison of data containing the search terms, as described above, may be provided. In a further study, a direct comparison of entities such as persons, organizations and events, appearing in both Twitter and online newspaper articles may be conducted.

The software is available as an online website[7], and is expected to be finalized in October 2014.

[7] available on http://maggie.lt.informatik.tu-darmstadt.de/nod/ via http://tagesnetzwerk.de/

Acknowledgements

"Netzwerk des Tages" (Network of the Day) is funded by BMBF via a grant from Hochschulwettbewerb 2014[8].

References

Chris Biemann, Karsten Böhm, Gerhard Heyer, and Ronny Melz. 2004. Automatically building concept structures and displaying concept trails for the use in brainstorming sessions and content management systems. In *Proceedings of I2CS*, Guadalajara, Mexico. Springer LNCS.

Gerlof Bouma. 2009. Normalized (Pointwise) Mutual Information in Collocation Extraction. In *Proceedings of the International Conference of German Society for Computational Linguistics and Language Technology*, pages 31–40, Potsdam, Germany.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Uli Fahrer. 2014. Contrastive Co-occurrence Analysis on Twitter for the German Election 2013. In *GI-Edition: Lecture Notes in Informatics*, pages 257–260, Potsdam, Germany.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a German named entity recognizer with semantic generalization. In *Proceedings of KONVENS*, pages 129–133, Saarbrücken, Germany.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, Michigan, USA.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proc. ACL-HLT-2011*, pages 42–47, Portland, OR, USA.

Artjom Kochtchi, Tatiana von Landersberger, and Chris Biemann. 2014. Networks of Names: Visual Exploration and Semi-Automatic Tagging of Social Networks from Newspaper Articles. *Computer Graphics Forum*, 33(3).

Miriam Meckel and Katarina Stanoevska-Slabeva. 2009. Auch Zwitschern muss man üben: Wie Politiker im deutschen Bundestagswahlkampf "twitterten". *Neue Zürcher Zeitung*.

[8] `http://www.hochschulwettbewerb2014.de/`

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November. Previous number = SIDL-WP-1999-0120.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Uwe Quasthoff, Matthias Richter, and Christian Wolff. 2002. " Wörter des Tages"-Tagesaktuelle wissensbasierte Analyse und Visualisierung von Zeitungen und Newsdiensten. In *ISI*, pages 369–372.

Matthias Richter, Uwe Quasthoff, Erla Hallsteinsdóttir, and Chris Biemann. 2006. Exploiting the Leipzig Corpora Collection. In *Proceedings of the IS-LTC 2006*, pages 68–73, Ljubljana, Slovenia.

Stijn van Dongen. 2000. *Graph Clustering by Flow Simulation*. Ph.D. thesis, University of Utrecht, Utrecht.

Frank Van Ham, Martin Wattenberg, and Fernanda B Viégas. 2009. Mapping text with phrase nets. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1169–1176.

Panpan Xu, Yingcai Wu, Enxun Wei, Tai-Quan Peng, Shixia Liu, Jonathan JH Zhu, and Huamin Qu. 2013. Visual analysis of topic competition on social media. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2012–2021.