Processing Online News Streams for Large-Scale Semantic Analysis

Miloš Krstajić #[1], Florian Mansmann #[2], Andreas Stoffel #[3], Martin Atkinson *[4], Daniel A. Keim #[5]

#*University of Konstanz*
Germany
[1]milos.krstajic@uni-konstanz.de
[2]florian.mansmann@uni-konstanz.de
[3]andreas.stoffel@uni-konstanz.de
[5]daniel.keim@uni-konstanz.de

**EC Joint Research Centre*
Ispra, Italy
[4]martin.atkinson@jrc.it

Abstract— **While Internet has enabled us to access a vast amount of online news articles originating from thousands of different sources, the human capability to read all these articles has stayed rather constant. Usually, the publishing industry takes over the role of filtering this enormous amount of information and presenting it in an appropriate way to the group of their subscribers. In this paper, the semantic analysis of such news streams is discussed by introducing a system that streams online news collected by the Europe Media Monitor to our proposed semantic news analysis system. Thereby, we describe in detail the emerging challenges and the corresponding engineering solutions to process incoming articles close to real-time. To demonstrate the use of our system, the case studies show a) temporal analysis of entities, such as institutions or persons, and b) their co-occurence in news articles.**

I. Introduction

Tens of thousands of news articles are published every day on the internet by thousands of news sources. Acquiring and reading all this information by human has thus become practically impossible. Human's internal filtering mechanisms, which are created according to their preferences, interests and beliefs, narrow down this vast amount of information to an acceptable subset. As a result, people usually track only a small number of stories from a selected set of news sources, which they consider trustworthy.

On the other side of this relationship are news providers. In today's fast paced world, each news provider publishes, on a daily basis, the amount of information that is sufficient for an average *news consumer*. Latest news replace the older ones at higher speed than ever and, therefore, tracking the temporal aspect of the story becomes a hard task.

On the one hand, the vast amount of breaking news, and sources that provide them, require automated methods that will facilitate processing of the created information by the user in order to regain the overview. But on the other hand, sharing of the acquired data has to be realized with caution. Certain issues, such as legal or security related, define new constraints. Sharing the data between companies, academic institutions or other organizations can be a delicate issue, dependent on challenges in legal, security and/or political spheres.

In our concrete case, we obtained a live stream of online news articles, which are continuously collected by the European Media Monitor (EMM) [1]. An interesting aspect of this data is the fact, that the articles are written in a multitude of languages. EMM thereby focuses on entity extraction and news clustering approaches in order to cope with this multilingual corpus of news articles. In contrast to EMM, our system is designed to support the user in the following two tasks using visual representations:

A. Since the languages are annotated for each article, by means of our system the user can investigate the *temporal and quantitative occurrence of entities*, i.e., certain persons or institutions, in different languages.
B. The system can be used to investigate real-world entity relationships and their volatility by assessing *co-occurrences of entities* over time.

This paper is structured as follows: Section II presents related work in the field of news aggregation and visualization. Next, our news streaming system is described in detail in Section III. The system is then (cf. Section IV) demonstrated in two real-world case studies of temporal and cross-language comparison of entity occurrences. Finally, Section V summarizes our contributions and gives an outlook to future work.

II. Related Work

Currently, there are several approaches that deal with the analysis of news articles. A large audience uses so-called news aggregator systems, which provide latest articles clustered into groups of similar stories from different sources reporting on the same event. Publicly accessible aggregators such as Google News [2] or Yahoo News [3] show breaking news of the moment sorted by number of sources and categories. Newsmap [4], which uses news aggregated by Google, shows the data visually encoded into a TreeMap visualization, based on the amount of news in each cluster and category, to which the cluster belongs to. A major drawback of these news aggregators is that they are dealing only with the latest news, i.e. they provide the data for a specific (current) point in time, there are no possibilities for temporal analysis (or it is limited) and they don't give much semantic information about the events.

The TextMap website, based on Lydia [5], is an entity search engine, which provides information about different entities (people, places and things) extracted from the news sources.

The Europe Media Monitor (EMM) [1] is a multilingual news aggregator system, which also provides information about entities, such as people, organizations and geographical location mentioned in the news. Websites, which give access to the data collected and processed by EMM are NewsBrief [6] and NewsExplorer [7].

All of the above mentioned approaches lack or have limited possibilities for analysis of dynamic change of the information published on-line. Also, possibilities for visual exploration of collections of news articles, which would make better use of human visual system in detecting trends, patterns and relationships in the news space, are also limited. One of the first approaches that used visualization to depict temporal evolution of themes within collection of documents is ThemeRiver [8]. In [9], Wise et al. presented the IN-SPIRE visual analytics system, which uses spatial visualization of the large collection of documents for enhanced analysis. LensRiver [10] extends the river metaphor from ThemeRiver into an analytical system for temporal analysis of unstructured text retrieved from video broadcast news. It deals with evolution of themes over time, their hierarchical structure, and employs different visual analytics techniques to perform the analysis. Hetzler et al. [11] proposed to visualize the incremental change in the data by using highlighting of new (*fresh*) and old and probably irrelevant (*stale*) documents.

Temporal analysis of news is not just a question of visual depiction of news over the time domain, but also a fundamental problem in textual data mining. An issue of considerable interest is analysis of news articles as document streams that arrive continuously over time. Each stream is not only an independent sequence of documents, but it also exhibits braided and episodic character [12]. Moreover, in today's news reporting, most attention is paid to breaking news about the latest events, which are characterized by fast growth of amount of information until a certain peak is reached, and fading of interest afterwards. A formal approach to model *burst of activity* of topics appearing as document streams is presented in [13]. Furthermore, the propagation of short quotes over news websites and blogs is analyzed in [14].

III. News Streaming System and Data

A. System

Europe Media Monitor (EMM) [1] is a news aggregator, which collects news articles from over 2,500 sources in 42 languages. These hand-selected sources include media portals, government websites and commercial news agencies. EMM processes 80,000 - 100,000 articles per day, enriching them with various metadata on which we perform our analysis.

Figure 1 presents the overall system architecture. Data retrieval is realized as an extension to the existing EMM web service architecture. Incoming data is handled by Java servlet, which is responsible for two-way communication with the other side from which it receives HTTP post requests. The servlet is designed as an external EMM processing node connected to the EMM pipeline, whose web service architecture is briefly described in [1]. This push technology allows for immediate retrieval of the new data as soon as it becomes available. Additionally, this architecture also allows us to send the results of our processing back, where it can be included in the EMM system. Similar nodes exist in the EMM processing chain and each of them performs a specific processing task on incoming data, which is retrieved from the upstream node, and outputs processed data to the next node on the downstream. Queueing and scheduling of jobs is implemented in every node to ensure successful transfers without loss of data. The processing node thereby sends back HTTP response status code 200 to the node on the upstream to acknowledge that the data was successfully received, understood, and accepted. This simple modular architecture allows for easy implementation of new modules (nodes) in the system, or replacement of the old ones.

Incoming HTTP post requests are Unicode XML files containing semantically annotated information in the metadata by modules on the upstream. Very often, several XML documents with news metadata are bundled within one incoming XML file. We perform splitting of the articles into separate files based on the unique article id, which is more suitable for our analysis. The XML metadata is also transformed directly in the servlet to our standardized internal XML format, which is used in our group for the analysis of text.

Our system is running on Solaris 10 on Sun Fire X4600 M2 x64 server, with 4 AMD Athlon Opteron 8384 CPUs and 32 GB RAM. The storage is realized through a Sun Storage Tek J4400 Array with 32 1 TB hard drives. Apache Tomcat 5.5.27 is used as a servlet container and MySQL 5.4.2-beta.

B. Data

Incoming XML file consists of semantically enriched metadata that is of great interest for our analysis, such as entities, categories, geo-tags, URL of the article, source, publishing time, date and language. An example file is shown in Figure 2.

People and organizations mentioned in the article are extracted in the entity recognition process [15], which relies heavily on multilingual Named Entity Recognition and Classication (NERC) and cross-lingual information aggregation. They are provided as `<emm:entity>` element, together with additional attributes: *id*, *type*, *count*, *position*, and *name*. Content from the `<emm:entity>` tag and *name* attribute show, respectively, the name of the person or organization, as it appears in the entity database and the name variant. The main reasons for these variations are morphological variants,

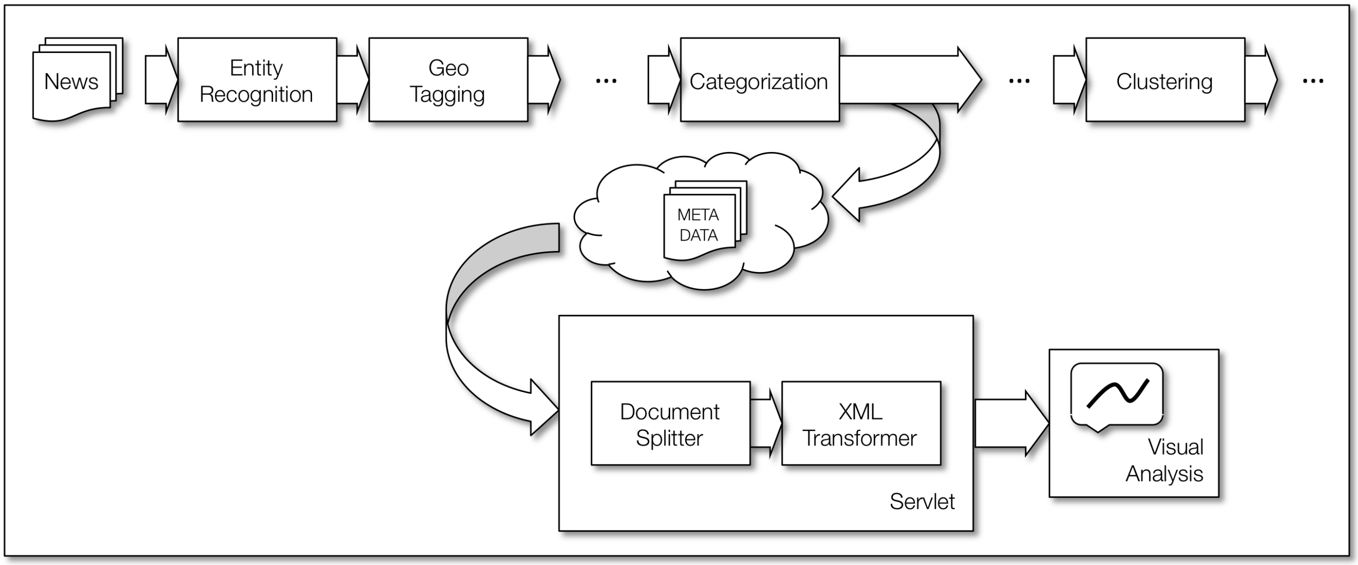

Fig. 1. News streaming system. News articles are semantically annotated in the pipeline of processing nodes in EMM system and XML metadata is sent in HTTP posts. On our side, the incoming data is processed within a servlet designed as an external EMM node.

```
<item emm:id="bbc-78ee9c05b6c74cf842c409f7c3467c1a">
  <link>http://news.bbc.co.uk/2/hi/asia-pacific/8353451.stm</link>
  <pubDate>2009-11-10T19:19+0100</pubDate>
  <source url="http://news.bbc.co.uk"
        country="GB"
        rank="1">bbc</source>
  <iso:language>en</iso:language>
  <category emm:rank="1"
        emm:score="18"
        emm:trigger="warned[2]; deadly[2]; war[1]; fire[2];
        incidents[1]; warns[1]; protesting[1]; nuclear[3];
        warship[1]; weapons[3]; battles[1]; armed[2]; ">
          Security</category>
  <category emm:rank="1"
        emm:score="44"
        emm:trigger="North Korea[5]; Pyongyang[1]; ">
          NorthKorea</category>
  <emm:entity id="1510"
        type="p"
        count="1"
        pos="469"
        name="Barack Obama">Barack Obama</emm:entity>
  <emm:georss name="South Korea"
        id="192"
        lat="37.5424"
        lon="126.935"
        count="1"
        wordpos="23"
        class="0">South Korea</emm:georss>
</item>
```

Fig. 2. Semantically annotated metadata. Entities, categories and geographic locations extracted from the news articled are passed as content in appropriate elements. Each metadata is enriched with additional information, such as trigger words, scores, type, etc.

repeated use of the name in the text, spelling mistake or adaptation of the name to local spelling rules. All articles are classified into different categories based on combinations of trigger words which are provided as a value of the <emm:trigger> attribute within the <category> element. Association of a news article to certain categories gives additional semantic information that can be analyzed further. This is also a logical basis for building a hierarchy over a large collection of news articles in the future. Every article has a <language> element, which is added in the language detection process. Semantic analysis of news articles across multiple languages gives the analyst an overview of what's being talked about in different countries. Geoinformation about location mentioned in the articles is provided within <emm:georss> element. The method for geolocation recognition and disambiguation in the free text is described in [16].

Incoming data has to be transformed to our *internal XML format*. Our standardized format is created to streamline structural analysis of any textual data by applying uniform attributes to logical structural elements of text and its meta information. It is used to ensure consistency of work within the group, thus requiring an additional data preprocessing step for the news analysis system. Therefore, news articles metadata belong to its <header> element. Semantic information about entities, countries, categories is converted to <DOC_ATTR> elements within the header. Text of the article, which could be retrieved using URL information from the metadata, would belong to <level> element within <body> of our XML. It should be annotated with attribute *type* with value *section*. Further processing of the text would split the structure of the text into other types, such as *paragraph*, *sentence*, *phrase* or *token*.

In our analysis, we have worked with 5 months of data, which was collected between June and October 2009. In total, 1,736,246 articles were collected, with 4,980,972 entities records and 8,847,596 categories records in English, German and French. There are 87,472 unique entities and 1,110 unique categories in the database in this period. In total, our collection is created from 979 news sources.

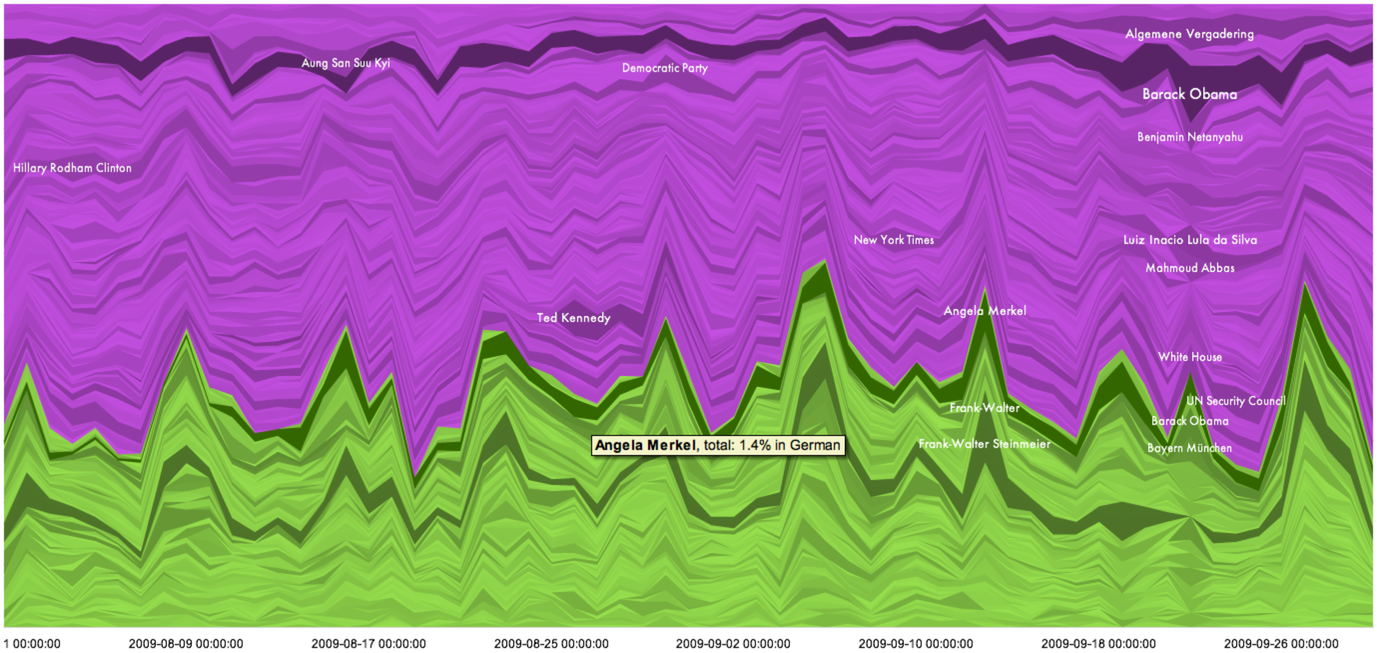

Fig. 3. Temporal analysis of entities. The x-axis represents time. Each stream represents a separate entity (person or organization). Height (y-value) of the stream at certain point in time represents share (relative amount). Violet color represents entities mentioned in articles in English, green represents German. Saturation is mapped to the total number of entities in the whole period. Entities mentioned more than 20 times per day are shown. Data for August and September 2009 are shown, with daily aggregates.

IV. Case Study

A. Temporal Visual Analysis

In analyzing the temporal aspect of the news feed, we are interested in the question how did the popularity of people mentioned in the articles evolve over time. Which people and organizations were most talked about in the news appearing in different languages? Are we able to identify people who are constantly in the news? Who are the people that had only temporary popularity in some period of time? How can the amount of information about these entities be compared? Understanding temporal aspect in the analysis of entities is an interesting challenge and when we are dealing with such an amount of information, visualizing the results helps in its processing and gathering new insights.

Visualizing the temporal data using stacked time series graphs is a well known approach. We employ a simple but efficient interactive solution based on NameVoyager [17] to show the initial results of our collaboration and potential for future research.

Figure 3 shows *streams* of entities (people and organizations) that were mentioned in the news articles in August and September 2009, from news sources from all over the world that publish in English (violet) and German (green). Streams are created from daily counts of entities in the news articles. For improved visibility, only entities that were mentioned at least once more than 20 times per day are shown. Saturation determines overall count for each entity and is used to distinguish very popular entities in the whole period from the less popular ones. The total daily amount of entities is normalized in order to show *share* of each entity per day. We think that using share as a relative measure gives a better overview of entity popularity over time. It is also possible to visualize absolute daily values for specific analyst tasks. Quick observation of visualization with absolute values can reveal interesting patterns, such as low traffic on weekends.

Looking at the visualization it is easy to spot few highly saturated streams. Another visual clue is given with text labeling of streams; stream, which has a daily share greater than a certain value, is labeled with the name of the entity. The label is positioned on the middle of the highest value in the stream. Inspection of highly saturated streams shows that the person most mentioned in the news articles in English is Barack Obama. In case of articles in German, most mentioned persons are Angela Merkel, Frank-Walter Steinmeier and Barack Obama. Text labeling of entities reveals several other entities that stand out, and some of them have only a temporary popularity. For example, Hilary Clinton was often in the news in early August while Ted Kennedy, US Senator from Massachusetts, who passed away on August 25, 2009, was one of the people mentioned the most in the last week of August.

Interactive features of the tool allow the user to get additional information about the dataset. For example, values that describe entities, such as total amount or percentage of news articles in a specific language are provided in the tooltip. Besides, in the analysis of entities, Visual comparison of a certain entity across languages (Figure 4) can be performed by selecting the appropriate stream from the visualization, or by textual search. In case of Barack Obama, it is easy to notice

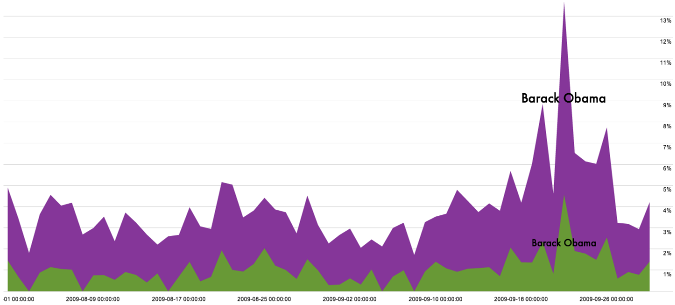

Fig. 4. Visual comparison between languages for a single entity. Barack Obama in news in English (violet) and German (green)

a similar pattern in both languages, which shows increasing amount of news about him, with peaks toward the end of September 2009.

Since visualization of large collection of documents gives a good overview of the whole dataset and its characteristics, detailed information about a specific news articles can be provided on demand. Direct links to news sources can be established, in case the analyst wants to read the full text of the article.

B. *Exploring relationships*

In analyzing the relationships between entities, the first question is which entities appear together in the news. How do the networks of certain people or organizations look like? How can we benefit from visual analysis of this dataset? A good way to show this information is to visualize graphs of personal networks that could be interactively explored. As an example, we have built a network of 1,000 most frequent co-occurrences of persons during 4 months. Our tool uses the well-known radial tree layout to display the graph and allows interactive exploration of each subtree. By selecting a node, the surrounding nodes are rearranged on the same distance from the focus node, but their co-occurence pairs are taken into account as well. Interesting subnetworks are shown on Figures 5, 6 and 7.

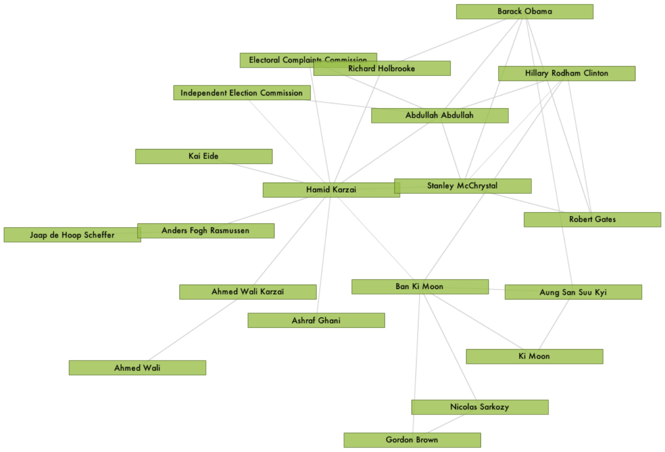

Fig. 5. Hamid Karzai Network. The network shows people appearing in the news articles covering elections in Afghanistan: UN and NATO officials, US representatives, but also political rivals.

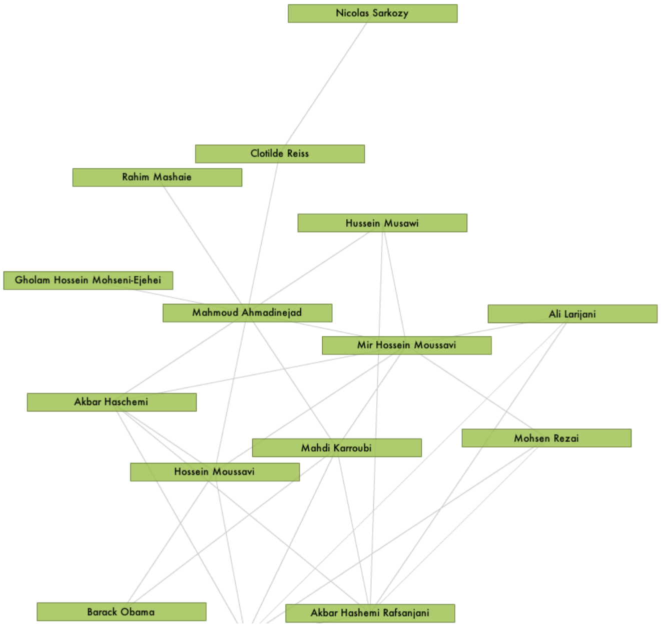

Fig. 6. Mahmoud Ahmadinejad Network. The figure shows imperfections in the dataset, since Ahmadinejad's political opponent, Mir-Hossein Mousavi, appears several times under similar names.

Analysis of network of people and organizations around Hamid Karzai (Figure 5) gives an overview that is highly related to presidential election in Afghanistan in August 2009, which was characterized by controversy. Therefore, besides Karzai's political opponent Abdullah Abdullah, several UN and NATO officials, Independent Elections Commission and US Special Envoy for Afghanistan and Pakistan Richard Holbrooke can be found appearing frequently together with Karzai in the news. Other direct connections from the child nodes can be immediately revealed, such as Richard Holbrooke and Barack Obama, or Anders Fogh Rasmussen and Jaap de

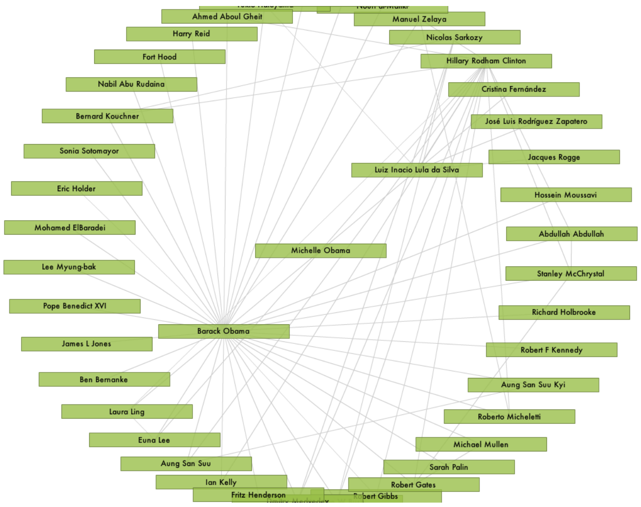

Fig. 7. Michelle Obama Network

Hoop Scheffer, current and former Secretary General of NATO (respectively).

Michelle Obama's network (Figure 7) includes direct edges to her husband, Barack Obama and Luiz Inacio Lula da Silva, President of Brazil. Analysis of temporal dynamics of both Lula da Silva and Obama reveals that they were mentioned very often in the news in the end of September 2009. Further inspection showed that Michelle Obama appeared together with Lula da Silva in the news about the 2016 Summer Olympics bids, where both took significant part (Obama supported Chicago as a candidate city, while Lula da Silva supported Rio de Janeiro).

These examples show strength of parallel use of visual analysis for exploration of entity relationships and for temporal analysis of entities, even on this initial stage, when simple techniques are employed. Also, it shows some weaknesses. First of all, there is no distinction between frequent co-occurrences and non-frequent ones. Further analysis should distinguish between frequent and rare pairs and also take into account differences across languages and sources. Also, graph layout should be improved to accommodate the networks of nodes with high number of edges, since radial tree layout produces overlapping, thus making the exploration of relationships difficult.

V. Conclusions

In this paper, we presented a data sharing and analysis system that is in use between the European Commission's Joint Research Centre and the Data Analysis and Visualization Group at the University of Konstanz in Germany. We identified several tasks in the analysis of news articles and methods for exploration of semantically annotated news data and demonstrated them in two case studies:

1) The first case study demonstrated the system's use for temporal analysis of entity occurrences over a time period of two months including cross-language comparison of entity occurrences.
2) The second case study dealt with the analysis of relationships among entities, which we realized using a radial graph layout.

In particular, these case studies showed several possibilities of using visualization to facilitate the exploration of large collection of news articles using semantic metadata.

Our future work in semantic analysis of news will include research on evolution of stories in which the entities (people and organizations) are involved. News stories are built around events and can have complex properties. Events can be disjoint in time, they can evolve from other events, or merge into the same story. Therefore, news stories can also merge, morph, divide, vanish and reappear. Understanding their hierarchical and semantic structure is a great challenge in news analysis research. Future analysis of EMM data stream will take into account categories as a part of the hierarchical structure of news, and also further comparison across languages and sources will be performed. We will also continue our research in the direction of performing the analysis in near real-time. The borderline between historical and real-time data analysis raises a lot of issues in both textual data mining and visualization research fields.

Acknowledgements

This work was partially funded by the German Research Society (DFG) under grant GK-1042, "Explorative Analysis and Visualization of Large Information Spaces".

References

[1] M. Atkinson and E. Van der Goot, "Near real time information mining in mulitlingual news," in *WWW '09: Proceedings of the 18th international conference on World Wide Web*. ACM, 2009, pp. 1153–1154.

[2] (2009) Google news. [Online]. Available: http://news.google.com/

[3] (2009) Yahoo news. [Online]. Available: http://news.yahoo.com/

[4] M. Weskamp, "Newsmap," *Webdesigning Magazine*, June 2004, http://www.newsmap.jp.

[5] L. Lloyd, D. Kechagias, and S. Skiena, "Lydia: A system for large-scale news analysis," in *String Processing and Information Retrieval: 12th International Conference, SPIRE 2005, Buenos Aires, Argentina, November 2-4, 2005: Proceedings*, 2005, pp. 161–166.

[6] (2009) Emm newsbrief. [Online]. Available: http://emm.newsbrief.eu/

[7] (2009) Emm newsexplorer. [Online]. Available: http://emm.newsexplorer.eu/

[8] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "Themeriver: Visualizing thematic changes in large document collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 9–20, 2002.

[9] J. Wise, J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, "Visualizing the non-visual: spatial analysis and interaction with information from text documents," *Information Visualization, IEEE Symposium on*, vol. 0, p. 51, 1995.

[10] M. Ghoniem, D. Luo, J. Yang, and W. Ribarsky, "Newslab: Exploratory broadcast news video analysis," in *VAST '07: Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 123–130.

[11] E. G. Hetzler, V. L. Crow, D. A. Payne, and A. E. Turner, "Turning the bucket of text into a pipe," in *INFOVIS '05: Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*. Washington, DC, USA: IEEE Computer Society, 2005, p. 12.

[12] J. Kleinberg, *Temporal Dynamics of On-Line Information Streams*. Springer, 2006.

[13] ——, "Bursty and hierarchical structure in streams," in *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2002, pp. 91–101.

[14] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2009, pp. 497–506.

[15] R. Steinberger and B. Pouliquen, "Cross-lingual named entity recognition," *Linguisticae Investigationes*, vol. 30, no. 1, pp. 135–162, January 2007.

[16] B. Pouliquen, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fuart, W. Zaghouani, A. Widiger, A.-C. Forslund, and C. Best, "Geocoding multilingual texts: Recognition, disambiguation and visualisation," in *Proceedings of LREC-2006*, Sep 2006.

[17] M. Wattenberg, "Baby names, visualization, and social data analysis," in *INFOVIS '05: Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*. Washington, DC, USA: IEEE Computer Society, 2005, p. 1.