Methods for Interactive Exploration of Large-Scale News Streams

Daniel A. KEIM, Miloš KRSTAJIĆ, Peter BAK, Daniela OELKE and Florian MANSMANN
University of Konstanz, Germany
{keim, krstajic, bak, oelke, mansmann}@dbvis.inf.uni-konstanz.de

Abstract. This paper presents a visual analytics approach to exploring large news articles collection in the domains of polarity, spatial and entity analysis. The exploration is performed on the data collected with Europe Media Monitor (EMM), a system which monitors over 2,500 online sources and processes 90,000 articles per day. In the analysis of the news feeds, we want to find out which topics are important in different countries, what is the general polarity of the articles within these topics and how the quantitative evolution of entities that are mentioned in the news, such as persons and organizations, developed over time. To assess the polarity of a news article, automatic techniques for polarity analysis are employed and the results are represented using Literature Fingerprinting for visualization. In the spatial description of the news feeds, every article can be represented by two geographic attributes, news origin and the location of the event itself. In order to assess these spatial properties of news articles, we conducted our analysis, which is able to cope with size and spatial distribution of the data. To demonstrate the use of our system, we also present case studies that show a) temporal analysis of entities, and b) analysis of their co-occurrence in news articles. Within this application framework, we show opportunities how real-time news feed data can be efficiently analyzed.

Keywords. News Feed Application, Sentiment Analysis, Spatiotemporal Analysis, Entity Analysis.

1. Introduction

Excess amount of information is generated each day on the internet, making processing of the content very difficult for the individual. Global news agencies, such as The Associated Press (AP), Reuters and Agence France-Presse (AFP), provide media companies with news reports from all over the world. This content is then duplicated, enriched with commentary and opinion. Additionally, news are filtered according to importance or interest of the editorial team. Besides, local media outlets produce their own local (or global) content having their own point of view, which might be specific to the geographic location of the news source (region, country) or specific to a certain group of people. Furthermore, blogs allowed common people to become active content creators themselves, not just passive readers, thus making the analysis of such amount of information one of today's greatest challenges.

The current paper describes an application aiming to conduct comprehensive analysis of such material. The paper first describes the system that provides the data and

how it is processed for analytic purposes. Second, opportunities for in-depth analysis are shown, taking polarity, temporal and spatial analytic techniques as examples.

2. Europe Media Monitor

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 [6] or Yahoo News [31] show breaking news of the moment sorted by number of sources and categories. Newsmap [28], 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 [20], is an entity search engine, which provides information about different entities (people, places and things) extracted from the news sources.

Europe Media Monitor [1] is a news aggregation system which monitors over 2500 news sources, collecting 80000 - 100000 news articles per day in 42 languages. Websites, which give access to the data collected and processed by EMM are NewsBrief [10] and NewsExplorer [11]. The goal of EMM is to provide assistance to human media monitoring, through automatic analysis and categorization of articles from these sources. In a typical information gathering scenario, journalists try to give the answers to the "Five Ws" questions - "who, what, when, where and why". The EMM system employs various information extraction, clustering and analysis techniques to help the user in answering these questions.

A screenshot of the website, taken on Feb 11, 2010, is shown on Figure 1. The central part of the webpage consists of a list of articles, which are clustered by the EMM system, in each language, into stories which report about the same event. The line graph above the list of articles on Figure 1 shows 10 stories with the largest number of reports in 4 hour time window over the last 12 hours. On the right side of the page, additional features allow the user to get quick access to information on the website, such as toolbox (which provides RSS feed, clusters of articles in KML file or email subscription) and country watch, which gives more information about the country most reported in the news at the moment.

In the lower left corner on Figure 1, a list of categories classified during the information extraction process is shown. The articles are classified into more than 600 categories, which give a high-level overview of the type of the article. The categories are manually classified on the website according to website users' needs into super-categories. These super-categories are displayed on the left side on Figure 1. "EU Focus" tracks categories of high interest like *swine flu* and *financial crisis*, but also EU-centric about European Council and Swedish presidency of the EU. "EU Policy Areas" contain categories such as *agriculture*, *human rights* and *public health*, while "Themes" contain categories like *crime*, *terrorist attack*, *natural disasters* and *security*. In our work, we will be focusing on categories from different super-categories, namely *agriculture*, *security*, *terrorism* and *sports*.

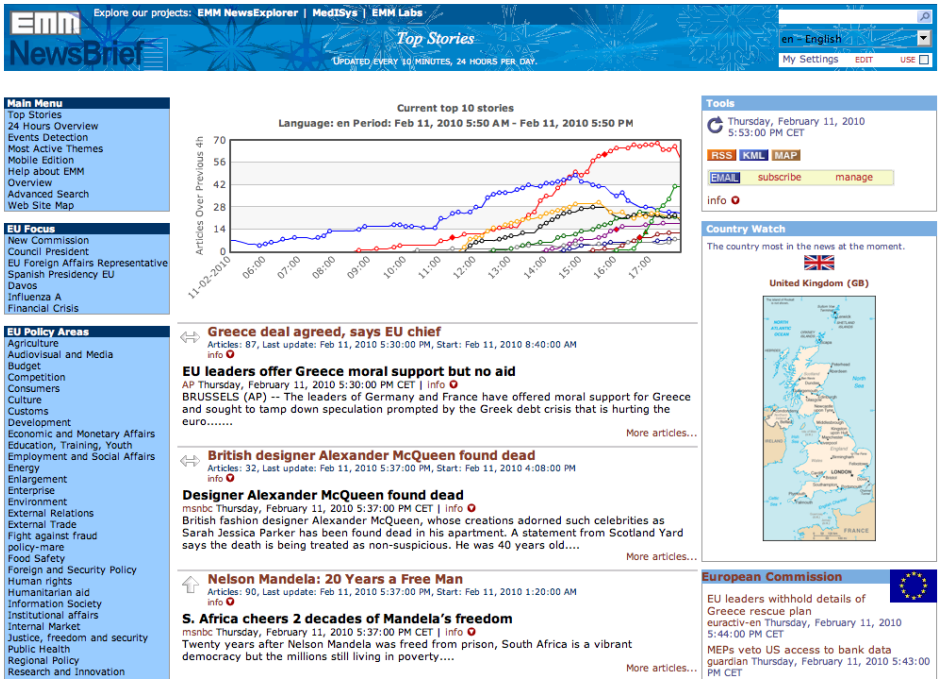

Figure 1. EMM Website snapshot

Each article is enriched with various metadata, such as people, their titles and organizations which are mentioned in the articles. This data, which is extracted in a separate entity recognition process, is available in all languages, as in example with the President of the USA Barack Obama that is shown on Figure 2. In order to give the answer to the question "where?" about the story location, geographical information is also extracted. The disambiguation module in the system uses the meta-information of previously recognized entities, such as names of places, provinces, regions and countries, in order to perform geo-tagging.

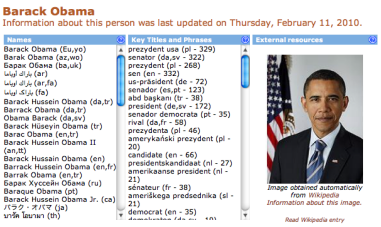

Figure 2. Detail from EMM Website Person Page. Entity Recognition process provides data about people, their titles and usual phrases in multiple languages

Descriptive analysis on the temporal development of topics can be obtained as a part of the website, which can be seen on Figure 3. Red line chart shows the total number of articles in 4 hour time windows, the blue bar chart shows the number of new articles in 10 minute updates and the blue area shows the cumulative sum of articles in the story.

Detailed description of the news streaming system can be found in [17]. The website itself provides many more features for descriptive analysis of the articles. However, in order to extract more useful information from the system, in-depth analysis with more sophisticated analytic methods is needed. The current paper focuses on three application areas, namely in the fields of polarity analysis, spatial analysis and temporal analysis of the news feeds.

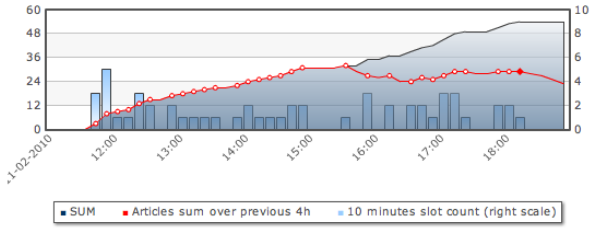

Figure 3. EMM Website Graph - Single story cluster evolution over time

3. Polarity Analysis

Newspapers report about what happens in the world. However, they are not completely neutral in doing so. Instead, there is a bias with respect to *which* news they write about but also with respect to *how* they write about the events.

In this section we are going to present a technique for polarity analysis on news streams. We demonstrate its usefulness on a sample of about 3 weeks of news data (approx. 15000 articles).

3.1. Automatic Techniques

To get a polarity score for each article a basic analysis algorithm was applied. With the help of two lists with signal words (one containing words with a negative connotation and the other with positive ones)[1] each word is classified as positive, negative or neutral[2]. We count the number of positive signal words in an article and subtract the number of negative signal words from it. To improve the accuracy of the method, negation is taken into account. This is done by inverting the value of a word, if in a maximum distance of X words a negation signal word is found (such as “no”, “not”, “without”, . . .). In this case, the parameter X (the maximum distance to the negation signal word) was set to 3, a value that experimentally proved as minimizing the failures.

Usually, the above mentioned technique is used in the context of sentiment analysis. Note that when we apply it to newspaper articles, we do not measure the author’s opinion about the topic directly. However, what we measure is still related to sentiment in the broadest sense. Words with positive connotations produce positive feelings about the topic and vice versa. To account for this difference in the semantics of the score, we call our analysis “polarity analysis” although classical sentiment analysis techniques are use.

[1]The lists of signal words are taken from the General Inquirer Project: http://www.webuse.umd.edu:9090/

[2]Note that the list contains signal words of all parts of speech. That means, not only adjectives, but also nouns, verbs, etc. (e.g. “catastrophe”, “to like”).

In the last years, much research has been conducted in the area of sentiment analysis (see e.g., [32,29]). For document-level sentiment-analysis, which is what we do, classification algorithms are often used (such as Naive Bayes). In case of review analysis (which is one of the easiest domains for sentiment analysis), many algorithms exist that additionally analyze the text with respect to *what* has been commented on (such as [9,24,23]). Those so-called attribute-based approaches first extract the attributes that the writers commented on and afterwards calculate the sentiment scores separately for each attribute. So far, only few approaches exist that work on news articles, as this is a difficult domain. Please refer to [22] for a more comprehensive overview of sentiment and opinion analysis algorithms.

Using the simple algorithm that was described above comes with the advantage that the automatic analysis can be done fast. In our case, this is an important property, because we want to work on a streaming data set. However, note that the algorithm could easily be exchanged in the system, if the analysis task requires a more sophisticated one.

3.2. Visual Representation of the Polarity Analysis results

In analyzing the news feed, we are interested in the question how the different groups (e.g. countries) report on different topics. Do they share the same view on the topic with respect to the polarity of the articles? Are there clear differences between some countries? Does it depend on the topic how much they agree with each other? Which special observations can be made? What is challenging in this case, is that we cannot say clearly, what we are looking for. The fact that our dataset is not static, but that we are working with a data stream, aggravates the problem. Knowing what would be interesting to look at today does not necessarily mean that this would also be a good view for tomorrow's news.

A good way to deal with this problem is to use an expressive visualization technique to represent the result of the automatic algorithm. Thanks to the great capabilities of the human visual system, large amounts of information can be grasped and processed at once if they are visualized. The automatic algorithms in the background make it possible that the tedious work of extracting the polarity of the text is left to the machine. The more demanding work of detecting patterns and anomalies in the data is done by the human analyst when interacting with the visualization.

Several visualization techniques for sentiment and opinion analysis exist. Among them is [19] that represents the result of attribute-based opinion analysis with bar charts. In [4] reviews are clustered according to topic and the average opinion per cluster is visualized in a treemap representation. Morinaga et al. [21] use a 2D scatterplot to display the results of their automatic algorithm. A visualization technique that is able to show the temporal aspect of a data set is introduced in [26]. Note, that all approaches expect for the last one are working on product reviews and not news.

We decided to apply the Literature Fingerprinting technique that was introduced in [13]. The advantage of this pixel-based visualization compared to the previously mentioned ones is that a large amount of values can be shown without the need of aggregation. Furthermore, the inherent hierarchy is clearly visible. In this technique, each score (here the polarity score) is represented by a single pixel and its value is mapped to the color of the pixel. Single pixels are grouped according to a given hierarchy (e.g. first according to topic and within the topics according to the location of the news agency).

3.3. Application

Figure 4 shows a Literature Fingerprint for about three weeks (May 11th - May 28th 2009) of English newspaper articles from all over the world. In the left column each pixel represents the set of news articles for a single country. A block of pixels contains all the articles that belong to a specific topic. (Our topics are *agriculture*, *security*, *sports*, and *terrorism*). Color is mapped to the average polarity score of the articles that are represented by the pixel (see color scale at the right). In the right column, the data are shown on an aggregated resolution level. In this case, each pixel represents a single article and the articles are first grouped according to the country they belong to and then according to the topic they report on. Again, color is mapped to the polarity score but this time it represents the score for a single article.

Looking at the left column of figure 4 it is easy to see that there are clear differences between the topics with respect to their fundamental tone. While *security* and *terrorism* show a negative trend for most countries, the opposite is the case for *agriculture* and *sports*. While some topics show big differences in polarity between the single countries, *agriculture* is a topic that is almost homogeneously seen as positive.

The advantage of the representation in the right column is that not only an average score is depicted. In the higher resolution level it can also be seen how many articles contributed to the average value and how homogeneous the reporting is with respect to the polarity that is expressed.

In the last line of the right column three kinds of patterns can be perceived: countries for which almost all pixels are colored in shades of red (negative), countries which are homogeneously shaded in blue or green, and finally countries in which all colors of our color scale occur. Among the ones whose articles were homogeneously classified as reporting negatively about terrorism are Australia, Croatia, and the Cayman Islands (see enlarged depiction at the bottom of the figure). A closer analysis shows that their articles are primarily about terrorism in other countries. Of course, those terroristic activities are clearly damned. Countries in which a concrete danger of terroristic acts exists, usually show a multi-colored picture in our visualization with an in total negative tendency (see e.g. Great Britain or Israel). The reason for this is that also political speeches or activities (such as cooperations with other countries) that talk about fighting against terrorism are included. This also nicely exemplifies how our algorithm works. The latter articles are dominated by security-related terms, measures against terrorism, and optimistic perspectives for the future and thus our algorithm classifies them as positive, because the connotation of those terms is positive. This means that the used algorithm would not distinguish between an article that agrees with those political speeches and another one that cites them but afterwards disassociates itself from the message. Finally, we were surprised to see that almost all articles in this category of the Syrian Arab Republic and Yemen are clearly classified as positive. Reading through the articles revealed that in those days the foreign ministers of the Islamic countries met. Among other things, they discussed ways to preserve Islamic values and the Islamic culture despite of experienced terroristic activities. For the participating countries this was the major topic in those days and the optimistic tone of the conference (also praising much their own countries efforts and perspectives) explains the large amount of positive reports in the terrorism category.

Finally, we can also analyze the articles across topics. It is interesting to see that Great Britain has about twice as many articles in the category *sports* than in the category

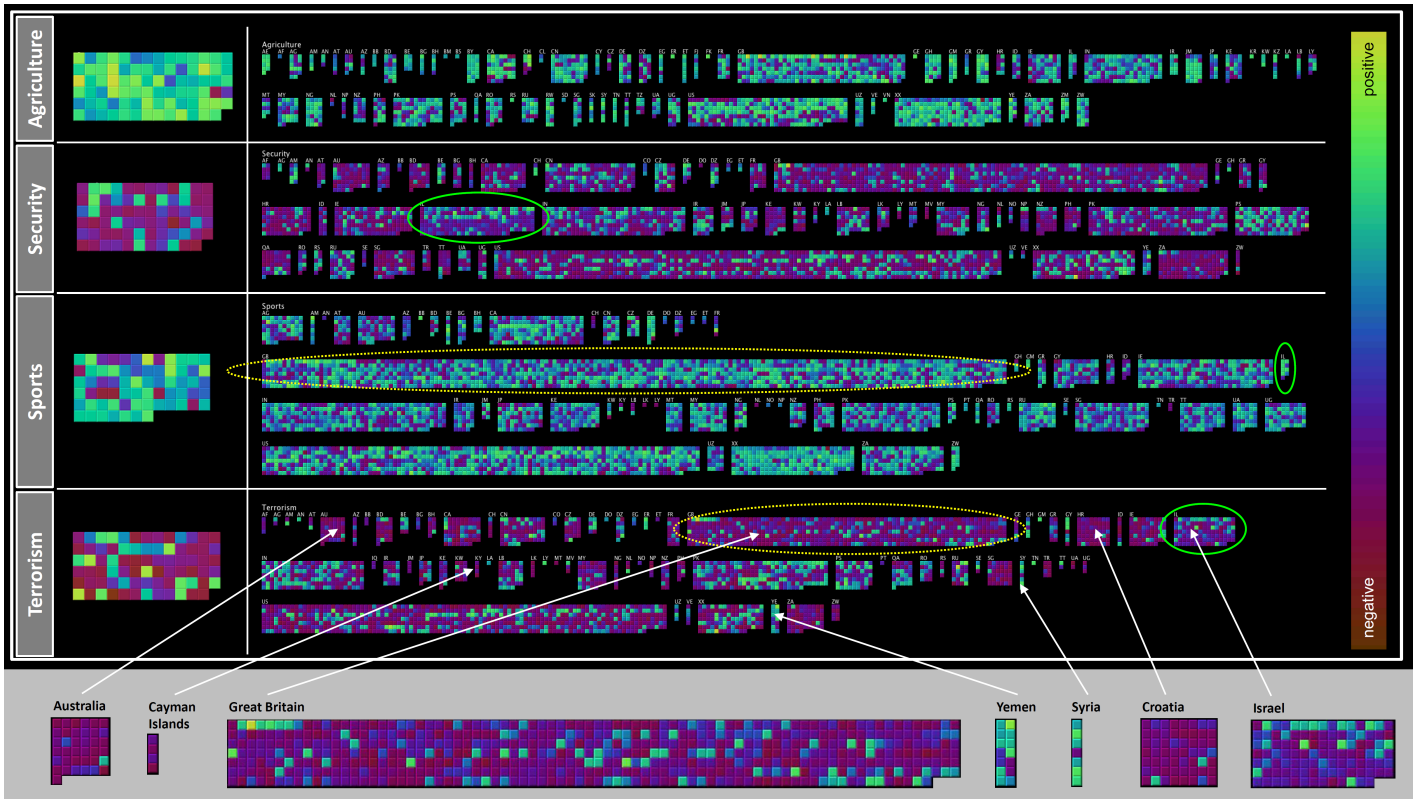

Figure 4.Polarity Analysis on news articles. In the right column, each pixel represents a single article. Its color is mapped to the calculated polarity score. Articles that were published in the same country are grouped together into blocks of pixels. In the left column, each pixel represents the average polarity score for a set of articles from the same country. We display four different news categories.

terrorism (see yellow dotted circles). Opposite to that, Israel has only very view articles in the *sports* category compared to the amount of articles in the categories *terrorism* and *security* (see green circles in figure 4).

4. Spatial Analysis

To the best of our knowledge, only one other geo-sentiment visualization exists [33]. In the paper, Zhang et al. analyze news articles with respect to four categories - joy vs. sadness, acceptance vs. disgust, anticipation vs. surprise, and fear vs. anger. First, a 4D-feature vector with the dimensions mentioned before is created for every article. Next, a summary feature vector is generated for each news site by averaging the values of all the articles from this publisher. This summary vector is then visualized as a line graph that shows the development of the values for this news site over time. Finally, each line chart is placed on a map on the news site's location. The differences to our approach are the following: 1) sentiment is mapped to the location of the news sites, 2) the authors use a different specification of "sentiment" and 3) different visualization technique, which is very coarse, (line chart) is used. Additionally, the authors work on the data collected from 11 news sites only.

In our case, the data provided by EMM can be regarded as an event-based multidimensional dataset, where each event represents one news item with a list of attributes. This dataset contains two geographic attributes. First, the news origin refers to a news agency located in a country or state, from which the news was published. This information is automatically mapped to the location of the news agency. For example, the Associated Press is located in New York City (NY), consequently the origin of their news has the geographical coordinates of New York City / Manhattan. Second, requiring more sophisticated tagging, is the location of the news' topics themselves. For the purpose of geographically tagging the location of a news item, the full text article is scanned for city, state and country names. When such a name is found in the document, its geographic location is automatically acquired from a look-up table. Consequently, one news item could have more than one location, when more distinct places are mentioned. In practice, however, the majority of news items has only one annotated geo-location.

4.1. Application Challenges

The most common approach to visually represent geographic information on a map is to pinpoint - as a single pixel or a small icon - to its location. However, when a large amount of multidimensional event based data has to be represented on a computer screen, this task becomes a challenge. In this case, data has to be shown using single pixels for each event and to map one attribute to the pixel's color [12]. Such pixel-based visualization techniques have a great advantage that they are scalable to the size of the data. A well-known problem is that these techniques often have a high degree of overlap, which may occlude a significant portion of the data values that are presented. Finally, a great challenge in spatially distributed data is the lack of correlation between information content and area size.

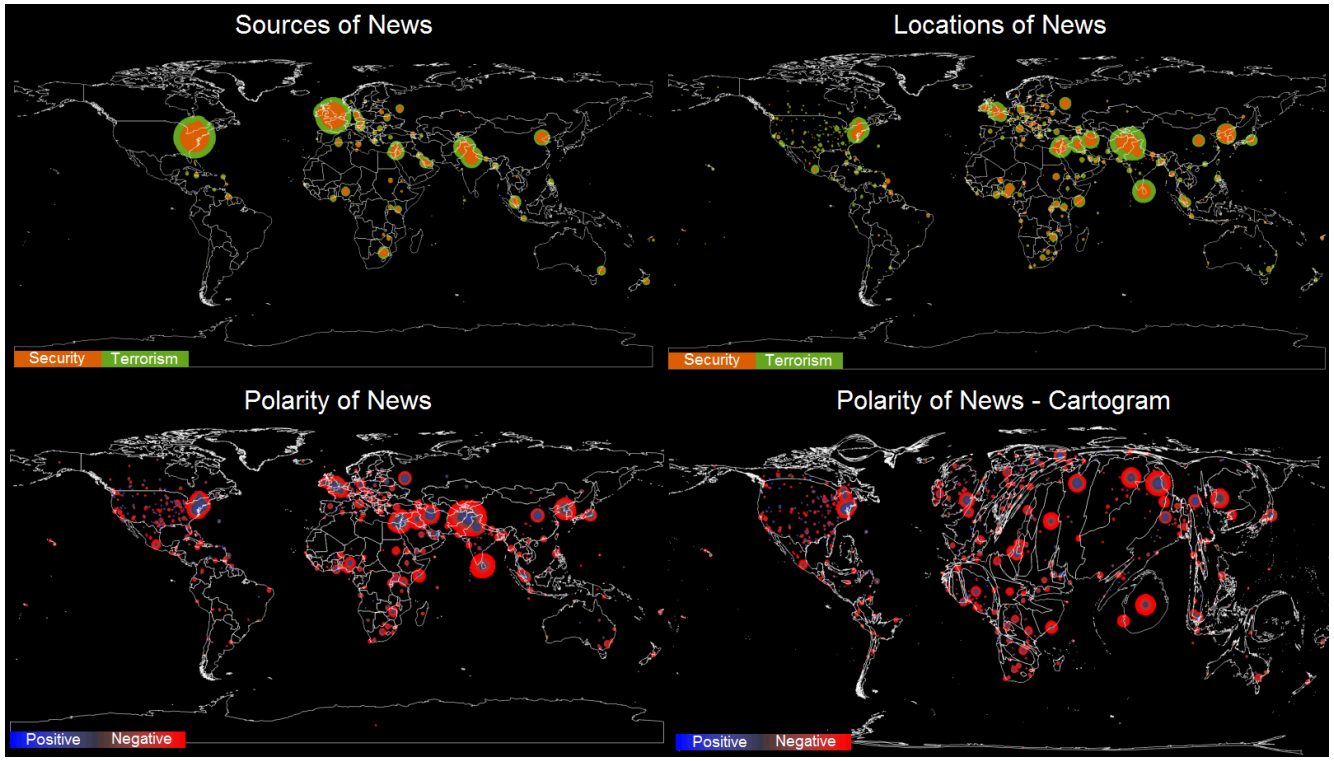

Figure 5.Spatial analysis of news collected from May 11, 2009 until June 7, 2009. The sources (upper left) and the locations (upper right) of news are represented for two categories - Security (orange) and Terrorism (green). The polarity of the news articles (lower left) is shown using bi-polar color map, reaching from blue (positive polarity) to red (negative polarity) on a logarithmic scale. This polarity map is also shown in a Cartogram representation, enhancing regions of importance, where the number of news items corresponds to the area of each country [16].

4.2. Methods

A number of different pixel-oriented visualization techniques have been proposed in recent years and shown to be useful for visually exploring data in many applications. These techniques differ mainly in their approach to arrange individual pixels in relation to each other, and in their choice of shaping the geographic regions to make maximal use of space.

4.2.1. Pixel Placement

In order to avoid overlapping pixels, in the current analysis we used a circular arrangement around the original location taking the given ordering of elements into account [2]. The ordering usually corresponds to the coloring attribute starting with colors that occur least frequently. With this arrangement a natural looking visualization without artifacts is generated. The ordering of elements prevents randomly arranged points that would not benefit the user [2].

4.2.2. Cartograms

Displaying large point sets on conventional maps is problematic. Conventional data-plotting obscures data-points in densely populated areas, while sparsely populated areas waste space and hide the details of information. Moreover, small clusters are difficult to find - they are not noticeable, and are sometimes even occluded by large clusters. A way to obtain more space for regions with a high point density are Cartograms, which distort regions such that their size corresponds to a statistical attribute [3]. In the example of the US elections, the actual size of a state will be rescaled in accordance to the number of its electoral votes. Cartograms usually preserve the topology of the data and the relationships between map regions and data points. They can be created by several algorithms, of which a detailed overview can be found in [25].

4.3. Analysis Results

The spatial aspect of the news was analyzed using the EMM data source with the techniques previously described. The data were obtained in the time period between May 11 and June 7, 2009. Figure 5 represents spatial analysis of news feeds showing the origin (upper left, where the news were published) and location (upper right, where the event mentioned in the news took place) for topics belonging to two categories: *Security* and *Terrorism*. The news originate mainly in Europe and in the US, and are reporting on the US, Europe, but also a lot on the Middle East and Asia. Spatial analysis of news feeds, showing the polarity score of topics related to security and terrorism (bottom left), is shown using bi-polar color map, having red for negative, and blue for positive news with increasing intensity. The news mainly report on Middle East, Central Asia (especially on the events in Sri Lanka) and North Korea in the particular time period. Although the majority of these news is negative in their tonality, there are some positive reports on successes in the fight on terrorism. The Cartogram representation enhances the area of these important locations.

5. Visual Analysis of Entities

Current approaches that deal with the analysis of news lack or have limited possibilities for analysis of dynamic change of the information published on-line. Besides, possibilities for visual exploration of collections of news articles, which would make better use of the 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 [7]. In [30], Wise et al. presented the IN-SPIRE visual analytics system, which uses spatial visualization of the large collection of documents for enhanced analysis. LensRiver [5] 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. [8] proposed to visualize the incremental change in the data by highlighting new (*fresh*) and old (*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 [15]. 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 [14]. Furthermore, the propagation of short quotes over news websites and blogs is analyzed in [18].

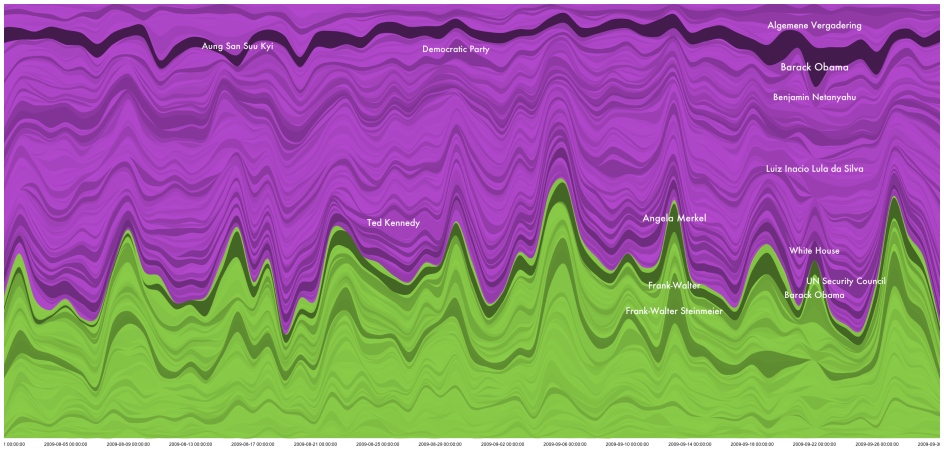

Figure 6. Temporal analysis of entities. The x-axis represents time. Each stream represents a separate entity (person or organization). Height (y-value) of the stream at certain point in time represents share (relative amount). Violet color represents entities mentioned in articles in English, green 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. Stacked graph representation of the data can be found in [17]

5.1. 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 *streamgraphs* is a well known approach. We employ a simple but efficient interactive solution based on NameVoyager [27] to show the initial results of our collaboration and potential for future research.

Figure 6 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, one of the people mentioned the most in the last week of August was Ted Kennedy, former US Senator from Massachusetts, who passed away on August 25, 2009. Another example is Hilary Clinton, who was mentioned very often in early 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 can be performed, either by selecting the appropriate stream from the visualization, or by textual search.

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.

5.2. *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-occurrence pairs are taken into account as well. Interesting subnetworks are shown on Figures 7 and 8.

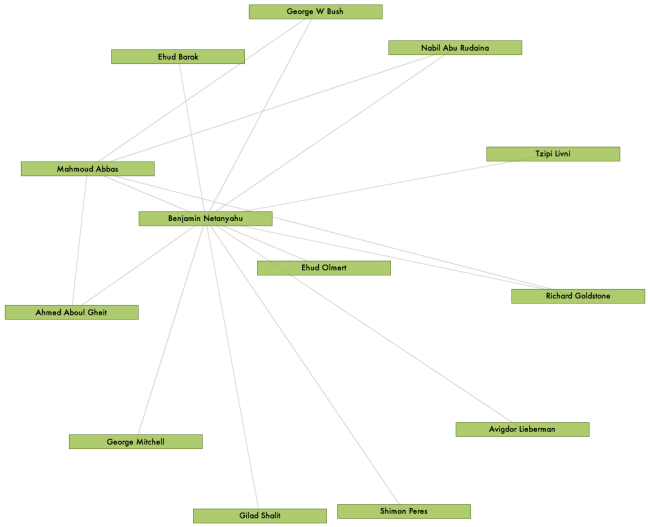

Figure 7. Visual exploration of relationships of news entities: Ehud Olmert Network. The network shows several important Israeli politicians that appeared in the news, but also other political figures involved in the Israeli-Palestinian conflict.

Analysis of network of people around Secretary General of NATO Anders Fogh Rasmussen (Figure 8) gives an overview that is highly related to presidential election in Afghanistan in August 2009, which was characterized by controversy. Therefore, besides Rasmussen and former NATO Secretary General Jaap de Hoop Scheffer, direct connection to Hamid Karzai leads to Abdullah Abdullah (presidential candidates), several UN officials, Independent Elections Commission and US Special Envoy for Afghanistan and Pakistan Richard Holbrooke. Inspection of the Ehud Olmert's network (Figure 7 shows different political figures involved in Israeli-Palestinian conflict that appeared frequently in the news. These are not just Israeli politicians and Palestinian leaders, but also George Mitchell, the American special envoy to the Middle East for the Obama administration, and Richard Goldstone, who headed the United Nations Fact Finding Mission on the 2009 Gaza Conflict. Further analysis of these relationships shows that, for example, Richard Goldstone was frequently appearing in the news in this period, because of the mission's final report, which was released on September 15, 2009.

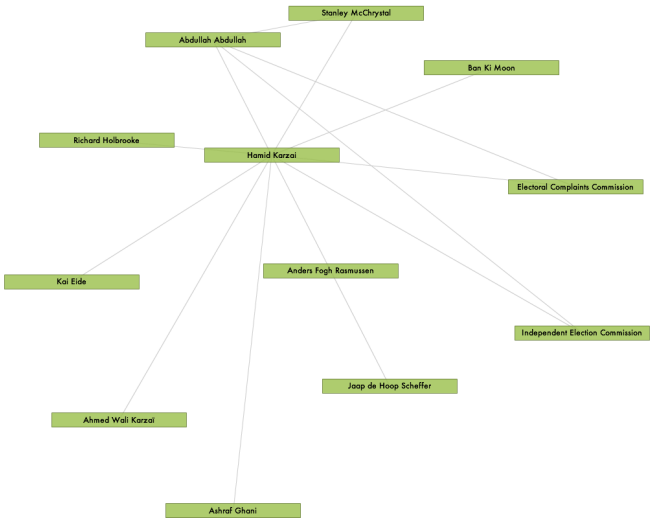

Figure 8. Visual exploration of relationships of news entities: Anders Fogh Rasmussen Network. The figure shows network of people and organizations appearing in the news about 2009 presidential election in Afghanistan.

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.

6. Conclusions and Future Work

The current paper describes an application framework for analyzing real-time news feed data. The data is provided by the Europe Media Monitor (EMM), which collects massive news in real-time and makes it accessible for public use. The aim of our application framework is to allow further in-depth analysis and exploration of these data. In our application we show polarity, temporal and geospatial analysis techniques as examples that are capable of processing such a challenging data source. Polarity analysis showed how to assess the "tonality" of the published news articles using a technique called Literature Fingerprinting. The geo-spatial analysis demonstrated that with the involvement of Pixel Placement and Cartogram techniques many insights can be gained by simply plotting the news articles as single pixels on the display. We have demonstrated the news streaming system's use for temporal analysis of entity occurrences and analysis of relationships among entities using radial tree graph layout. The great challenge for further techniques that will be implemented within this framework is integration in the EMM-platform, making the techniques scalable to cope with large datasets and real-time requirements.

Further research is planned to extend the available analytic and visualization methods with more powerful, efficient and sophisticated ones. Additionally, user interaction techniques should be included that allow a direct brushing and linking between different methods. Our future work in semantic analysis of news will include research on evolution of stories in which the entities (people and organizations) are involved. Understanding their hierarchical and semantic structure is a great challenge in news analysis research. As such, the presented application will be an excellent tool to explore breaking news in many languages and topics by the interested public.

Acknowledgements

This work is funded by the German Research Society (DFG), "Scalable Visual Analytics: Interactive Visual Analysis Systems of Complex Information Spaces" of the Priority Programme (SPP) 1335. The German Research Society (DFG) under grant GK-1042, "Explorative Analysis and Visualization of Large Information Spaces" partially supported this work.

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*, pages 1153–1154. ACM, 2009.

[2] P. Bak, D. A. Keim, M. Schaefer, A. Stoffel, and I. Omer. Spatiotemporal analysis of sensor logs using growth ring maps. In *IEEE Transactions On Visualization And Computer Graphics*. IEEE Press, 2009.

[3] P. Bak, F. Mansmann, H. Janetzko, and D. A. Keim. Density equalizing distrortion of large geographic point sets. In *Journal of Cartographic and Geographic Information Science*, volume 36(3), 2009.

[4] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. Pulse: Mining customer opinions from free text. In *Advances in Intelligent Data Analysis VI*, pages 121–132. 2005.

[5] 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*, pages 123–130. IEEE Computer Society, 2007.

[6] Google. Google news. 2010. http://news.google.com/.

[7] 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.

[8] 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*, page 12. IEEE Computer Society, 2005.

[9] M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[10] JRC. Emm newsbrief. 2010. http://emm.newsbrief.eu/.

[11] JRC. Emm newsexplorer. 2010. http://emm.newsexplorer.eu/.

[12] D. A. Keim, P. Bak, and M. Schaefer. Dense pixel displays. In *Encyclopedia of Database Systems*. Springer, 2009.

[13] D. A. Keim and D. Oelke. Literature fingerprinting: A new method for visual literary analysis. In *EEE Symposium on Visual Analytics and Technology (VAST 2007)*, pages 115–122, 2007.

[14] J. Kleinberg. Bursty and hierarchical structure in streams. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 91–101. ACM, 2002.

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

[16] M. Krstajic, P. Bak, D. Oelke, M. Atkinson, W. Ribarsky, and D. A. Keim. Applied visual exploration on real-time news feeds using polarity and geo-spatial analysis. In *WEBIST 2010: Proceedings of the 6th International Conference on Web Information Systems and Technologies*, 2010. in press.

[17] M. Krstajic, F. Mansmann, A. Stoffel, M. Atkinson, and D. A. Keim. Processing online news streams for large-scale semantic analysis. In *DESWeb 2010: Proceedings of the 26th International Conference on Data Engineering, ICDE 2010*, 2010. in press.

[18] 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*, pages 497–506. ACM, 2009.

[19] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM, 2005.

[20] 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*, pages 161–166, 2005.

[21] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. Mining product reputations on the web. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 341–349. ACM, 2002.

[22] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and TrendsÂő in Information Retrieval*, 2(1-2):1–135, 2008.

[23] A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346. Association for Computational Linguistics, 2005.

[24] I. Titov and R. McDonald. A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In *Proceedings of ACL-08: HLT*, pages 308–316. Association for Computational Linguistics, June 2008.

[25] W. R. Tobler. Thirty five years of computer cartograms. In *Association of American Geographers*, volume 94(1), pages 58–73, 2004.

[26] F. Wanner, C. Rohrdantz, F. Mansmann, D. Oelke, and D. A. Keim. Visual sentiment analysis of rss news feeds featuring the us presidential election in 2008. In *Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW 2009)*, 2009.

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

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

[29] M. Wiegand and D. Klakow. Optimizing language models for polarity classification. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, editors, *ECIR*, volume 4956 of *Lecture Notes in Computer Science*, pages 612–616. Springer, 2008.

[30] 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*, page 51, 1995.

[31] Yahoo. Yahoo news. 2010. http://news.yahoo.com/.

[32] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics, 2003.

[33] J. Zhang, Y. Kawai, T. Kumamoto, and K. Tanaka. A novel visualization method for distinction of web news sentiment. In *10th International Conference of Web Information Systems Engineering (WISE 2009)*, pages 181–194, 2009.