Incremental Visual Text Analytics of News Story Development

Miloš Krstajić, Mohammad Najm-Araghi, Florian Mansmann and Daniel A. Keim
University of Konstanz, Germany

ABSTRACT

Online news sources produce thousands of news articles every day, reporting on local and global real-world events. New information quickly replaces the old, making it difficult for readers to put current events in the context of the past. Additionally, the stories have very complex relationships and characteristics that are difficult to model: they can be weakly or strongly connected, or they can merge or split over time. In this paper, we present a visual analytics system for exploration of news topics in dynamic information streams, which combines interactive visualization and text mining techniques to facilitate the analysis of similar topics that split and merge over time. We employ text clustering techniques to automatically extract stories from online news streams and present a visualization that: 1) shows temporal characteristics of stories in different time frames with different level of detail; 2) allows incremental updates of the display without recalculating the visual features of the past data; 3) sorts the stories by minimizing clutter and overlap from edge crossings. By using interaction, stories can be filtered based on their duration and characteristics in order to be explored in full detail with details on demand. To demonstrate the usefulness of our system, case studies with real news data are presented and show the capabilities for detailed dynamic text stream exploration.

Keywords: News Stream Analysis, Topic Evolution, Dynamic Visualization, Text Analytics

1. INTRODUCTION

Understanding temporal development of unstructured and semi-structured text data streams is becoming increasingly important in many application areas, such as journalism, politics or business intelligence. At the same time, sources of textual data, such as online news providers are creating content in constantly growing amounts. While many automated and visual solutions that deal with the snapshots of information space already exist, few interactive systems can provide an user-friendly environment in which temporal context of these growing data collections can be successfully analyzed. Latest information coming from the news data providers prevails in the news landscape very quickly, putting the prior information out of the picture, even when it is necessary to keep the longer temporal context in mind to understand the current events. News stories that report on real-life events have characteristics that make analysis of this type of data challenging, from both the text mining and visualization perspective. Analyzing the temporal dynamics of the news content has been in focus of computer science researchers for some time. In the area of text data mining, a lot of effort has been put to model topics in evolving document collections. However, little work has been devoted to analyze complex relationships between news stories. In the visualization field, topic evolution is receiving more attention recently, although most of the proposed methods do not scale well when working with data streams. We believe that evolutive aspects of news stories cannot be separated from incremental visualization methods and that these methods must be coupled with interaction techniques that will allow the user to explore the rich content of text data streams in full detail.

In this paper, we propose a novel news stream visual analytics system that integrates topic evolution algorithms with interactive visualization methods at three temporal zoom levels. To support effective analysis of the growing news corpora, we combine interactive methods with incremental visualization of stories development to allow exploration of news data streams from a broader temporal overview to fine-detailed level of a single article that belongs to a particular story.

Further author information: (Send correspondence to M.K.)
M.K.: E-mail: milos.krstajic@uni-konstanz.de
M.N-A.: E-mail: mohammad.najm-araghi@uni-konstanz.de
F.M.: florian.mansmann@uni-konstanz.de
D.A.K.: daniel.keim@uni-konstanz.de

2. RELATED WORK

Research on topic development, visualization and analysis of temporal dynamics in information streams has its roots in text mining, information visualization and time-series analysis. In the field of text data mining, a well-known initiative was Topic Detection and Tracking (TDT),[1] which investigated methods of discovering events in broadcast news streams. Dynamic topic models[2] is another well-known approach, which extends Latent Dirichlet Allocation (LDA) to work with time-stamped data.

Although the challenges were described several years ago,[3] a lot of work in the field of streaming data and text visualization still has to be done. The most popular method to visualize topic trends over time is based on ThemeRiver,[4] which employs stacked graph visualization[5].[6] Fisher et al.[7] presented a keyword tracking tool that calculates the co-occurrences of most important terms in blogs. EventRiver[8] employs bubble-like visualization of news stories extracted from the broadcast news data. Real-time aggregation of news articles into important threads is presented by Krstajic et al.[9] Aigner et al.[10] lately provided an extensive systematic overview of time-oriented visualization techniques.

In the field of visual analytics, the recently published TIARA[11] system integrates interaction methods with text summarization and visualization based on tag clouds and stacked graphs. The work that is most closely related to ours is by Rose et al.,[12] where a similar text flow visualization is used to show evolution of daily news themes over time. Our approach extends this line of research with improved interaction, different levels of detail for different temporal intervals and directly addresses the challenge of working with topics that split, merge and overlap over time.

3. SYSTEM

3.1 System Overview

The goal of our system is to enable the user to understand temporal dynamics of news stories and their relationships. The system must be able to process the growing dataset as fast as possible and help the user to put the new information in the context of the past. These requirements thus pose both technical and analytical challenges in our design.

Our system uses the news stream generated by the Europe Media Monitor,[13] a publicly available online 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. By using URL metadata free text of the article can be easily extracted for further processing in the topic clustering step. In our prototype, we work with all sources that publish information in English (in total around 10,000 articles per day). Focusing on this subset makes the evaluation of the document clustering faster, but the system is designed to easily accommodate data arriving at a higher rate.

3.2 Data Processing

The news articles are collected as XML data, as described by Krstajic et al.[14] Each time-stamped data item contains metadata, such as named entities, which can be people or organizations that are mentioned in the document, or tags that categorize it (e.g. *earthquake*, *sports*, etc). These significant keywords can be used in addition to free text that can be retrieved from the URL of the article in the document clustering step. The document preprocessing module converts the incoming article information to the internal format and sends it as an input to the Story Creator module.

*http://emm.newsbrief.eu/

4. STORY CREATION

We have developed our system with the goal of working with news streaming data, i.e. a sequence of time-stamped documents that arrive continuously over time. Theoretically, the volume and the speed at which the documents arrive are unbounded, and, therefore, it is expected that the algorithms process the document corpora incrementally. In practice, news documents that are similar, i.e. report on a particular real-life event are usually temporally close. Therefore, it is meaningful to process the growing document collection at defined time intervals to create clusters of similar documents, and then sequentially compare the clusters from neighboring intervals to find similar stories that span across a wider time frame. A set of documents in each time interval can be treated as a fixed collection, that can be efficiently processed offline in order to find clusters of similar documents.

4.1 Creation of Daily Clusters

The Story Creator module clusters the news articles in 24-hour time intervals. Since our news data stream source provides data from a large number of news sources, it is expected to have a lot of similar documents, which report on the same real-world event. Document clustering has been exhaustively researched in the field of text data mining, with several main directions, such as dynamic topic modeling based on LDA[2] and topic detection and tracking.[1] The evaluation of document clustering output can be quite exhausting and, although benchmark sets do exist, the results with our real news data were varying, depending on the selected time intervals and the amount of documents in the corpus. However, our visual analytics framework allows easy replacement of the underlying clustering technique. Furthermore, news topics are very often characterized by their braided nature,[15] where topics overlap, split and merge. In this work, we address this challenge by comparing similar clusters from adjunct time intervals to detect overlapping news stories.

The clustering module in our framework is based on Carrot[2],[16] an open source framework for clustering search results. This framework provides two clustering algorithms whose important advantage is that they do not require a pre-defined number of clusters and are very efficient in terms of processing time and computing power. The first algorithm is Lingo,[17] which extracts frequent phrases from documents to produce high quality cluster descriptions (labels). Lingo is based on SVD (singular value decomposition) and uses VSM[18] to create term-document matrix, which is decomposed to create candidate labels and associate documents with the most similar labels. The second clustering algorithm is the Suffix Tree Clustering algorithm (STC),[19] which works under assumption that similar documents are usually expressed using identical phrases. In the world of news publishing, this is very often the case, since many online news portals publish modified versions of the original article, which was created by a global news agency, such as Reuters. Phrase-based approaches, such as STC, have the advantage towards term-based clustering techniques that the phrases are more informative than the

```
Sidi Bouzid (8 documents)

  [19] Tunisian Government: 14 Killed as Rioting Continues
       http://www1.voanews.com/english/news/

  [26] Tunisia 'to respond' to protests
       http://english.aljazeera.net/English
  ....

WikiLeaks' Assange (28 documents)

  [20] Assange: WikiLeaks to speed release of leaked docs
       http://www.ynet.co.il/

  [32] WikiLeaks' Assange faces new court hearing
       http://www.euronews.net/
  ....

Japan (4 documents)
  [ 5] Japan, American Airlines alliance being boosted
       http://thestar.com.my/

  [136] Japan, American Airlines Alliance Being Boosted
       http://www.irishsun.com/
  ....
```

Figure 1. Daily cluster output example. The clustering algorithm produces story labels, followed by the number of documents belonging to each story and a list of document IDs, titles and URLs.

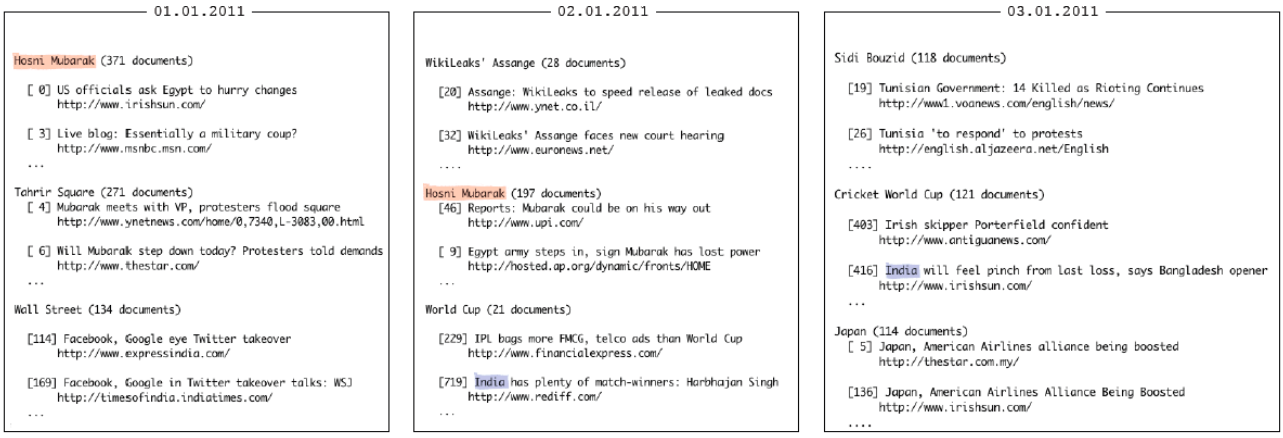


Figure 2. Comparison of daily clusters. For each day, a list of output stories and documents that are assigned to them is produced. The title words, descriptions and cluster labels are compared to detect stories that span over more than one day.

representative (most frequent) set of keywords and can be used to label clusters, which in turn can be a problem with the other clustering techniques. STC has two phases: first, it creates *base clusters*, containing the sets of documents with an identical phrase, and, second, combining the base clusters to form *final clusters*. To assess the quality of clustering results performed by the STC algorithm Stefanowski and Weiss[20] conducted an evaluation.

For each news article, we use its title, summary, entities and tags (categories) as input for the clustering algorithm. The user can refine the input by including or excluding entities or tags in the interaction phase. A short example of a daily cluster output is shown in Figure 1, with three identified stories: *Sidi Bouzid*, *Wikileaks' Assange* and *Japan*. The number of documents assigned to each story is given in the brackets, followed by the list of document IDs, titles and URLs.

4.2 Story comparison

Major stories can span over a longer period of time and understanding the evolution of these stories is one of the challenges that we are dealing with in our work. Since we want to be able to process data incrementally, the stories discovered in each time interval (24 hours), can be compared to the stories from the previous n intervals. Rose et al.[12] use $n = 7$ in their evaluations. In our case, we compare the stories from the current and the previous day, which helps in dealing with visual clutter that arises when $n > 1$. Besides, our experiments showed that, in most cases, stories appear on consecutive days without long breaks between them.

Essential content of each story consists of a set of keywords coming from the story title, description and document title words. We use Jaccard distance to calculate the similarity between stories belonging to neighboring time intervals. An example of the clustering output from three consecutive days is shown in Figure 2.

4.3 Describing Merging and Splitting of Stories

Evaluation of document clustering techniques is a daunting task. Very often, there is no clear cut between stories, and documents that are split into different clusters can be still very related by their content. Furthermore, news stories may disintegrate into two different topics during their evolution over time, or they can dissolve into one single topic (as shown in Figure 3). To address this issue, we have empirically set a threshold for the splitting and merging of news stories. The calculated cluster similarity values are then used for connecting both the most similar stories between consecutive days and also the stories whose similarities are higher than the given threshold. Therefore, when a story splits, each "child" story cluster that evolves from the "parent" story cluster will have the same visual encoding as the "parent". In case of merging of two stories into one, the new story will inherit visual features (namely, color) from the most similar story from the previous day. The details about visual encoding of stories and their evolution are given in Section 5.

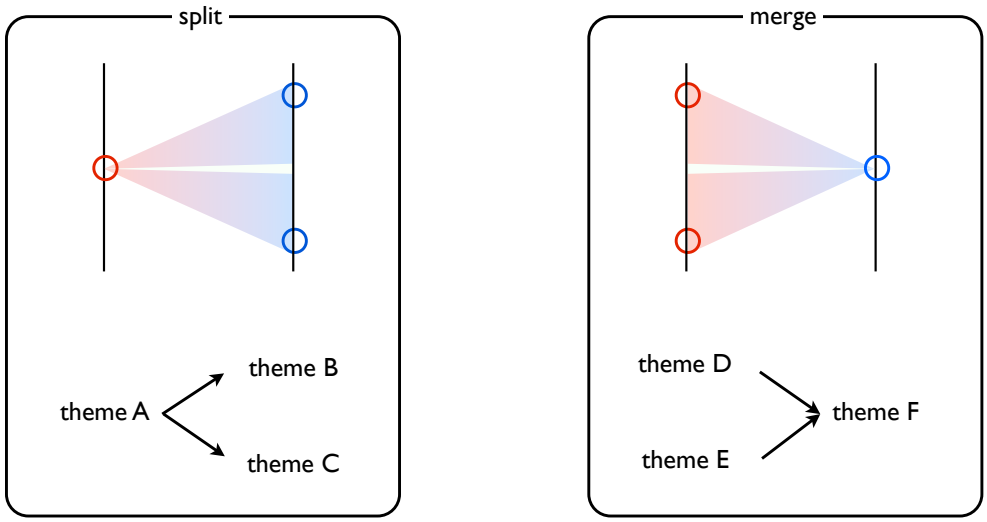


Figure 3. Splitting and merging of news stories. During story evolution over time, theme A can split into two different topics: theme B and theme C (*left*), or two originally disjoint topics, theme D and theme E, can merge into one main story: theme F (*right*).

5. VISUAL DESIGN

The automated layer of our system processes documents in incremental time intervals and produces clusters of documents that represent news stories reporting on a particular real-life event. Each news story is described by a story title (label), most important keywords, people and organizations mentioned in the news, but also by the *strength* of the cluster and the number of documents that belong to the story. The fundamental component of our visual design is a representation of the output for a specific time interval, as shown in the Main View in Figure 4. The visualization places days along the horizontal axis and daily stories are stacked in the single column list and ordered within the list by the strength of the story cluster. Each story is represented by a rectangle in the list, whose height is mapped to the number of documents assigned to the story. This allows the story title and the most important keywords to be displayed inside the rectangle. To provide an overview of the temporal evolution of the stories, we are using two visual clues: a story that evolves during several days is connected with shapes based on Bezier curves and it is colored, while the story that appears for only one day remains grey, which serves a dual function: first, the user can easily distinguish between longer and one-day stories, and, second, the flow of the story can be followed more easily when the interpolating shapes of several different stories overlap. The colors used in our system are selected using the ColorBrewer.[21]

This representation gives a good balance between mid-range temporal evolution, level of detail and importance of news stories. Navigation in this space is possible both in horizontal (time) and vertical (story importance) direction.

5.1 Zoomed View and Monthly View

To explore the contents of a particular story in detail, we developed the Zoomed View, which gives detailed information about the documents assigned to the story. Figure 5 shows Zoomed View for 3 days of a story titled Hosni Mubarak, containing titles of articles that belong to the story, summary and top URLs for each day.

In order to provide a broader temporal context, we designed a Monthly View, which gives less detail about particular stories, but gives an overview in which the analyst can get the first idea about the evolution of the news stories and their relationships. This view, shown in Figure 7, is enhanced by interactive filtering, which is described in more detail in the Section 6.

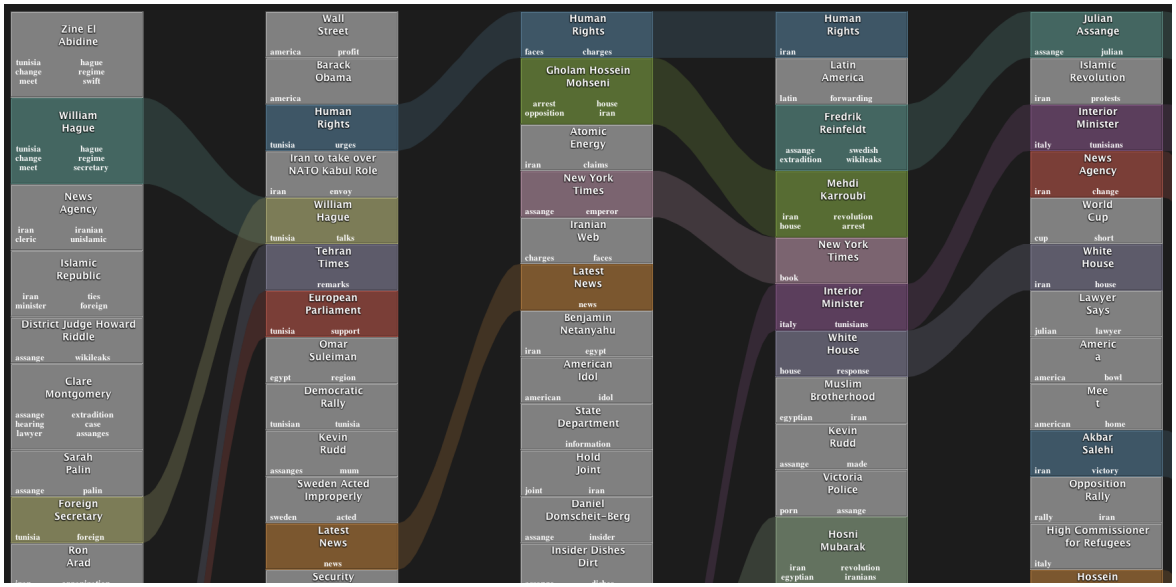


Figure 4. Main View

6. INTERACTION

In the monthly overview, the screen space can become easily cluttered due to the shapes connecting the story clusters and their ranking within the daily listings. A large number of short-lived stories, which might not be relevant to the analyst at this level, aggravate the problem. Since we are developing a system that supports incremental processing and visualization of a data stream, we need a solution that will maintain the layout of the past data regardless of the amount of new information. In order to address these challenges, we developed a filtering mechanism that allows the user to minimize the clutter and detect interesting patterns in this view.

6.1 Filtering

The problem of inter-interval connections clutter can be regarded as a graph layout problem. The daily clusters are the graph nodes and the connections between the clusters are the edges directed from the past to the present.

The user can use the *connectivity* and duration sliders to filter weakly connected and short stories. Using connectivity filter (Figure 6(a)), we can filter out the stories that do not split or merge and therefore keep only the stories with the most braided characteristics on the screen, while the duration filter (Figure 6(b)) allows us to keep the stories that span over multiple days.

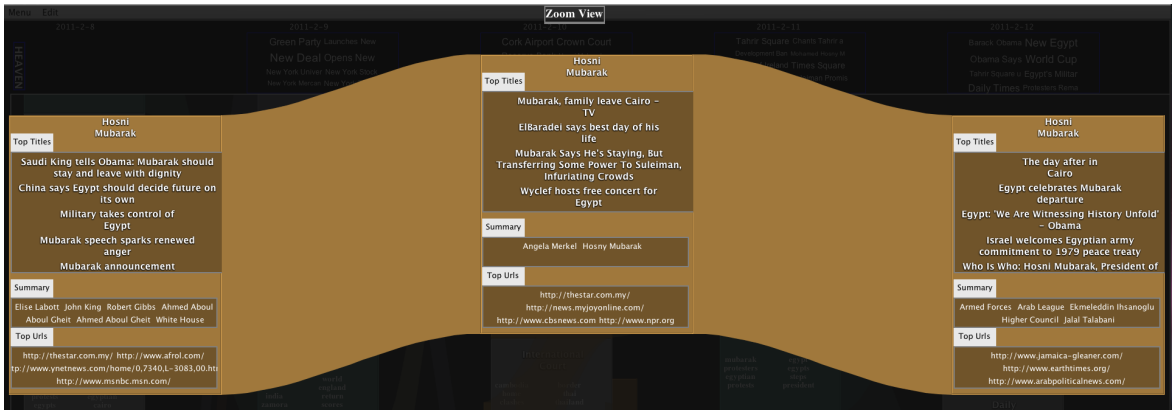
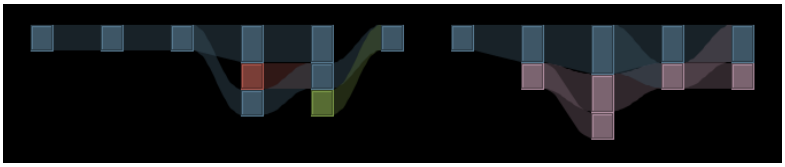


Figure 5. Zoomed View provides fine-detailed analysis of each story. Each daily cluster provides the titles of top 5 documents, short summary of the cluster and the most important URLs.



(a) Connectivity filter. Stories with high similarity remain displayed.

(b) Duration filter. Stories that evolve over 5 days can be easily filtered.

Figure 6. An example that demonstrates the user interaction through filtering. Filtering sliders are used to clean up the clutter in the Monthly View.

To minimize the clutter which is due to the daily cluster connections, we have to find an optimal solution for minimization of edge crossings. Our goal is to optimize a directed graph, where the direction of the edges represents a hierarchy. The hierarchical layout should:

1. have as few edge crossings as possible
2. be presented vertically and in a straight-path
3. have uniformly distributed nodes and avoid long edges

As an additional remark, we have to consider that the first condition can only be achieved by acyclic graphs, which we have in our case. A possible method for such a layout is based on the Sugiyama algorithm.[22] This method consists of four steps, which are difficult optimization problems:

1. *Delete all cycles*: Find a minimal number of edges according to the distance to which the graph is acyclic and switch their direction.
2. *Layer positioning*: Calculate a good allocation of the edges in all layers, so that they are directed upwardly. Replace all edges that go beyond one layer (This step can be skipped, because of our layout)
3. *Minimize the edges*: Calculate for each layer a layout so that the number of crossings is minimal.
4. *Positioning*: Calculate the x-coordinates of the nodes in this way that we have no overlap. (Since our coordinates are fixed, this step can be skipped)

The third phase is NP hard, because changing one layer affects the next layer. To ensure that the story order for each day is unaffected by future, we have reduced the problem to a two layer crossing minimization. This preserves the organization of the topics in the story flow. Therefore, the heuristic starts with the second layer. After sorting the nodes in this layer, we proceed iteratively with the next layer.

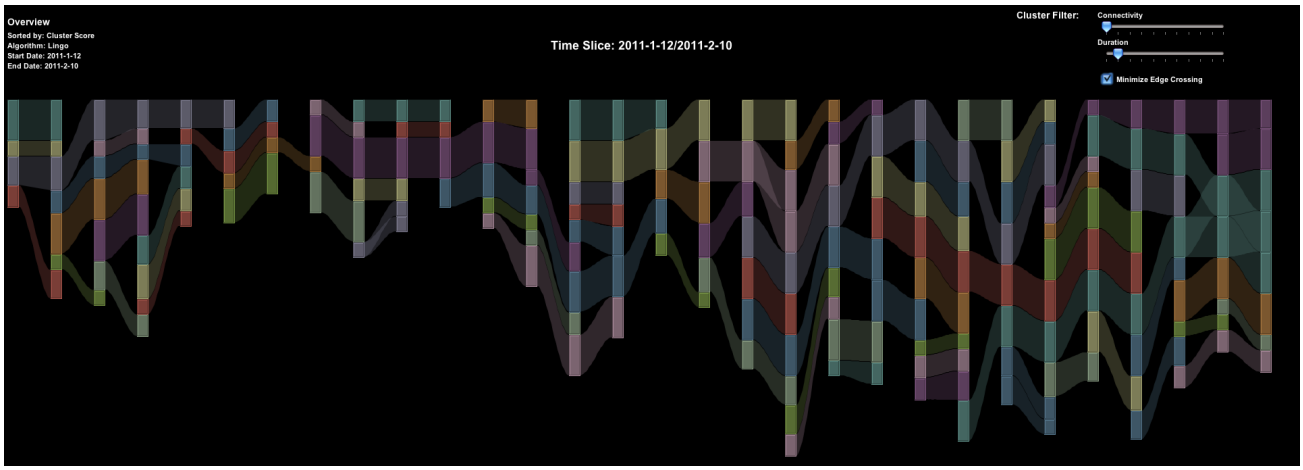


Figure 7. *Overview* of persistent news themes from January 12 to February 10, 2011. The filtering capabilities and edge crossing minimization on the upper right help to bring order to the news landscape.

7. USE CASE: THE ARABIC UPRISING 2011

Our application allows interactive analysis of the large international news landscape with the focus on temporal development of news topics. To demonstrate the added value that our tool gives to a news analyst, we focused our use case on the Arabic uprising in 2010 and 2011. Since December 18, 2010, a revolutionary wave has spread over more than 15 Arabic speaking countries. The Tunisian Revolt was the starting point and devolved, step by step, across other Northern African countries. Even the Arabian Peninsula was not excluded from minor demonstrations up to governmental changes.

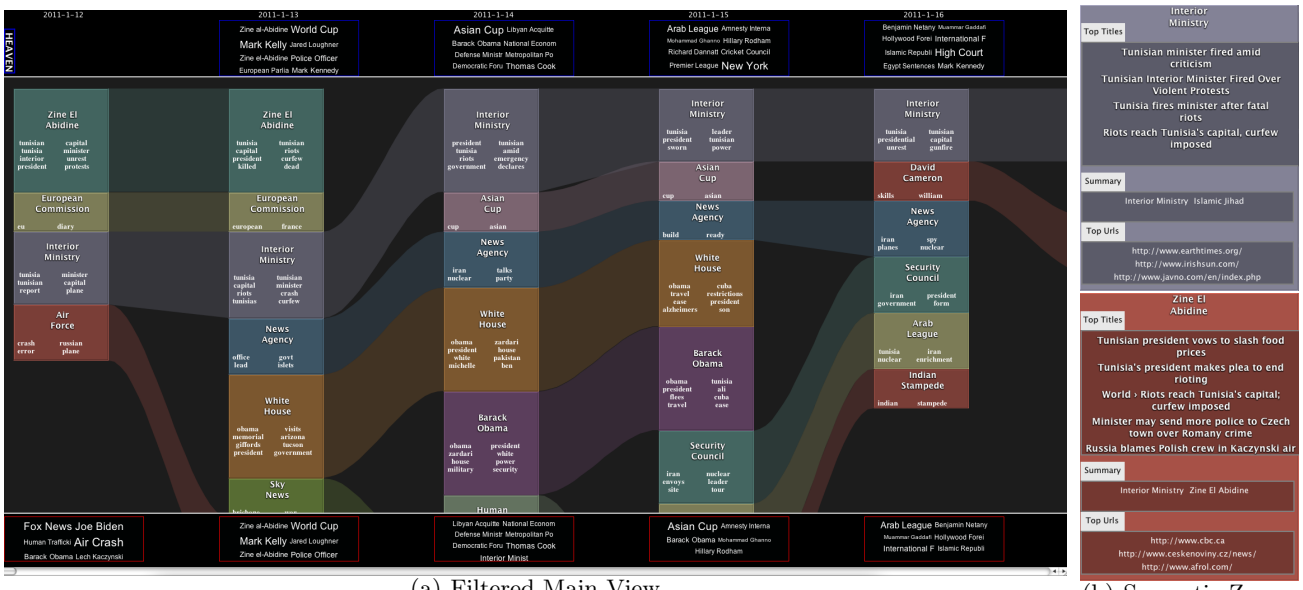
7.1 The Tunisian Riot

Figure 7 gives a visual overview of the persistent themes in the news between January 12 and February 10, 2011. Through filtering and edge minimization, the otherwise complex media landscape becomes easier to interpret due to the fact that many short unrelated themes are discarded and the strict theme ordering criteria according to popularity are relaxed. To avoid a loss of possible interesting stories, it is, for example, recommended to filter out all one-day stories by moving the slider at the top right. The next useful filtering step is to select the 'minimize edges' checkbox. From now on, the user can follow the flow of stories over 30 days, since positioning does not only depend on a theme's popularity for each day, but also on the theme's position on the previous day.

After interactive filtering and automatic layout optimization, the user can follow the ongoing stories in the main view. By looking at the top keywords and entering the zoomed view, the user can identify from the main view, shown in Figure 8(a), that the green and purple-grey colored themes cover the riots in Tunisia. The growing importance of these events in the news is coherent to the reports in the Wikipedia's *Current Events* portal[†]. The portal lists all occurrences in one month and also provides a snippet with a short summary of each event. For the 12th of January 2011, the snippet about the protests states the following: "Tunisia's Interior Minister Rafik Belhaj Kacem is sacked by President Zine El Abidine Ben Ali, who also orders the release of most people detained during recent unrest". At this point in time, Zine El Abidine Ben Ali, the President of Tunisia, is in the focus of the event. From now on, the suspension of the Interior Minister becomes an important event and the main view absolutely conforms with the political events. To discriminate between the events that belong to this long story, keywords and phrases are displayed. In this case, 'minister' or 'president' are characteristic terms. Nevertheless, these topics are still connected, because the protests were against the entire government, including the ministers and Ben Ali.

For detailed analysis, the semantic zoom as shown in Figure 8(b) reveals more information about both themes. This level of detail shows us clearly what we already briefly identified in the main view. The documents, which were assigned to the 'Interior Minister' topic, refer to the event which was mentioned before. The first three

[†]http://en.wikipedia.org/wiki/January_2011



(a) Filtered Main View (b) Semantic Zoom

Figure 8. *Tunisian Revolt* – The filtered main view from January 12, 2011 until January 16, 2011 reveals stories related to the Tunisian Interior Ministry and the Tunisian president Zine El Abidine Ben Ali. Details on demand show top titles, brief summary and most important URLs for each daily story cluster

titles are about the sack of Rafik Belhaj Kacem in the context of violent protests. In contrast, Zine El Abidine contains documents which refer to himself. As we can recognize in Figure 8(b) there is still some noise included - the fourth and fifth document do not belong to the topic but were highly ranked by the "most important title" algorithm.

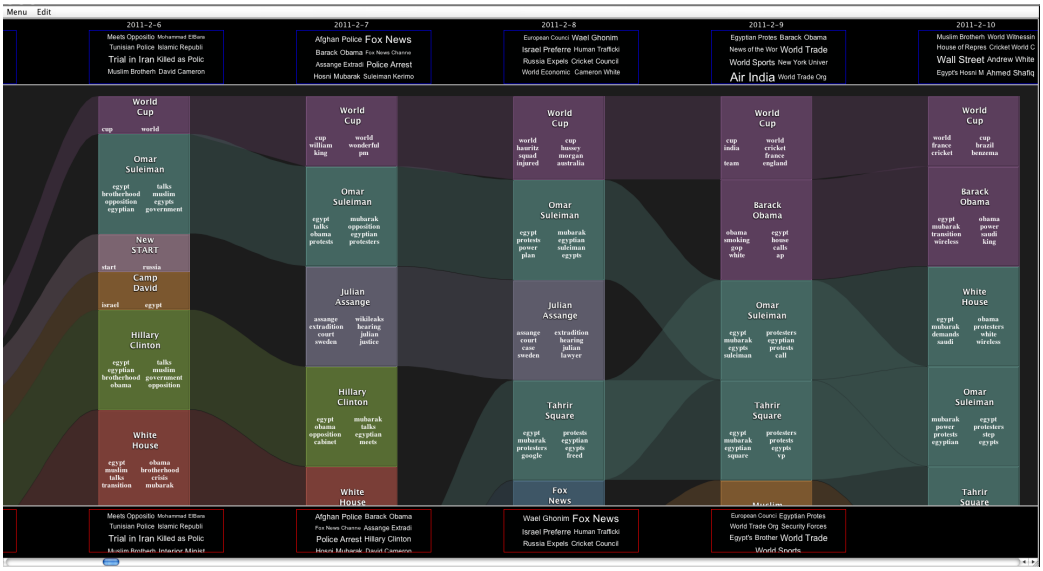
In the second week of the given time window, Tunisia and Zine El Abidine are still a major topic in the news. The *Current Events* portal reports about Tunisia's army firing on the citizens and governmental changes and the dismissal of more ministers from the Constitutional Democratic Rally party that had governed the country. This textual information, represented in a list view where no context is visible, is mapped as an ongoing stream in the main view. Zine El Abidine remains a representative main phrase, and self-explanatory keywords like 'government', 'quit', 'minister', 'fallen', 'victims' and 'revolution' mirror the actual state.

7.2 Riots in Egypt

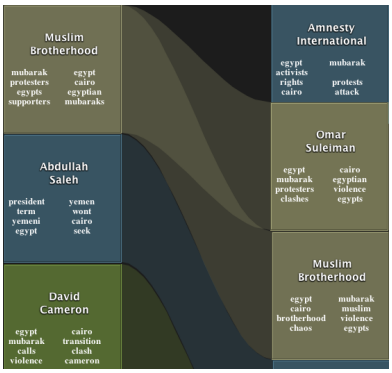
Figure 9(a) shows a filtered view from the 6th until the 10th of February 2011. We can easily identify that the news were dominated by the riots in Egypt. Several different stories reporting on the same topic appear in this view. Omar Suleiman, a former Egyptian army general, and Hosni Mubarak's Vice President, play an important role here. The reason can be identified by reading the top title words. The daily cluster shapes contain the terms about the 'Muslim Brotherhood' and a possible opposition. At this stage, Omar Suleiman and the Brotherhood were discussed as possible successors of Hosni Mubarak. Other protagonists, like Hillary Clinton, or Barack Obama, were also involved in the discussions. Besides, different topics like the cricket World Cup and the story about Julian Assange appear among the most important. By scrolling the main window horizontally back into the past, the user can see that the Tunisia and Yemen protests dominated the streams and built up a large number of stories, similar to what we discovered during the riots in Egypt.

We selected a split of one theme into several ones as one of many interesting patterns of the news flow to demonstrate the effectiveness of our system in describing story splitting patterns. the As Figure 9(b) reveals, the Muslim Brotherhood was displayed as a root theme, which leads to Omar Suleiman as a new ongoing theme.

At that point, severe discussions about the opposition and possible successors broke out. Both the Muslim Brotherhood and Omar Suleiman were part of these discussions. Figure 9(c) shows the result of applying the



(a) Filtered Main View from 6 February 2011 until the 10 February 2011



(b) Splitting of a single story into two different topics on the next day

(c) Zoomed Views of the 'Muslim Brotherhood' and 'Omar Suleiman'

Figure 9. *Riots in Egypt*

semantic zoom on the two themes. The documents assigned to each topic help to understand the causes for this split. As the user can identify, each of them has individual titles like *Door opened for Muslim Brotherhood* and *Suleiman, in new role, counts cost of Egypt turmoil*. These titles discriminate clearly between topics. However, the discussion about the opposition and the possible successors are connecting points for these events. This mixture causes a split which is in this case justifiable. As described in section 4, the splitting and merging of themes is dependent on previously defined similarity threshold. To be more concrete, a split will be defined if the calculated Jaccard similarity coefficient is greater than 30%. However, depending on each concrete use case, this parameter needs to be fine-tuned since it is possible that a split or merge is identified, while it should not, or the other way around.

8. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a visual analytics system for exploration of news stories development and their relationships. Our approach helps in understanding the evolution of long and short stories in a wide time frame, merging and splitting of stories, as well as fine-detailed analysis of story content on three different levels of detail. The incremental processing and visualization of unstructured and semi-structured text data allow the application of the system on the real-world news data streams. The global overview visualization helps in identifying the major news stories over a long period of time and it is enriched with the interaction techniques to filter and re-rank stories on various user-adjustable criteria in order to provide a clutter-free display. Finer-granulated views that correspond to shorter time windows allow analysis of news stories with higher level of detail, up to the textual content of each news article itself. We have demonstrated the effectiveness of our approach on a real-world news stream and described the news stories and their content that were found with our system.

In the future, we plan to refine our document clustering module to enhance the informative context of story labels and test the system with sliding and dynamically adjustable time windows. Additionally, we plan to replace the splitting and merging thresholds, which are currently based on empirically adjusted values to a more refined and less data dependent automatic algorithm. Our research efforts will continue in the direction of integrating incremental text analysis with novel visualization methods that will enable information analysts to analyze and understand growing document collections more effectively and efficiently.

ACKNOWLEDGMENTS

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] Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y., Umass, J. A., Cmu, B. A., Cmu, D. B., Cmu, A. B., Cmu, R. B., Dragon, I. C., Darpa, G. D., Cmu, A. H., Cmu, J. L., Umass, V. L., Cmu, X. L., Dragon, S. L., Dragon, P. V. M., Umass, R. P., Cmu, T. P., Umass, J. P., and Umass, M. S., "Topic Detection and Tracking Pilot Study Final Report," in [*In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*], 194–218 (1998).

[2] Blei, D. M. and Lafferty, J. D., "Dynamic topic models," in [*Proceedings of the 23rd international conference on Machine learning*], *ICML '06*, 113–120, ACM, New York, NY, USA (2006).

[3] Wong, P. C., Foote, H., Adams, D., Cowley, W., and Thomas, J., "Dynamic visualization of transient data streams," in [*IEEE Symposium on Information Visualization (INFOVIS 2003)*], **0**, 13, IEEE Computer Society, Los Alamitos, CA, USA (2003).

[4] Havre, S., Hetzler, B., and Nowell, L., "Themeriver: Visualizing theme changes over time," in [*Proceedings of the IEEE Symposium on Information Vizualization 2000*], *INFOVIS '00*, 115–, IEEE Computer Society, Washington, DC, USA (2000).

[5] Byron, L. and Wattenberg, M., "Stacked graphs–geometry & aesthetics.," *IEEE transactions on visualization and computer graphics* **14**(6), 1245–1252 (2008).

[6] Dörk, M., Gruen, D., Williamson, C., and Carpendale, S., "A visual backchannel for large-scale events.," *IEEE transactions on visualization and computer graphics* **16**(6), 1129–1138 (2010).

[7] Fisher, D., Hoff, A., Robertson, G., and Hurst, M., "Narratives: A visualization to track narrative events as they develop," in [*IEEE Symposium on Visual Analytics Science and Technology, 2008. VAST '08*], 115–122 (2008).

[8] Luo, D., Yang, J., Krstajic, M., Ribarsky, W., and Keim, D., "Eventriver: Visually exploring text collections with temporal references," *IEEE Transactions on Visualization and Computer Graphics* **99**(PrePrints) (2010).

[9] Krstajic, M., Bertini, E., Mansmann, F., and Keim, D. A., "Visual analysis of news streams with article threads," in [*StreamKDD '10: Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques*], 39–46, ACM, New York, NY, USA (2010).

[10] Aigner, W., Miksch, S., Schumann, H., and Tominski, C., [*Visualization of time-oriented data*], Springer (2011).

[11] Wei, F., Liu, S., Song, Y., Pan, S., Zhou, M. X., Qian, W., Shi, L., Tan, L., and Zhang, Q., "Tiara: a visual exploratory text analytic system," in [*Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*], *KDD '10*, 153–162, ACM, New York, NY, USA (2010).

[12] Rose, S. J., Butner, S., Cowley, W., Gregory, M. L., and Walker, J., "Describing story evolution from dynamic information streams.," in [*IEEE Symposium on Visual Analytics Science and Technology, 2008. VAST '09*], 99–106, IEEE (2009).

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

[14] Krstajic, M., Mansmann, F., Stoffel, A., Atkinson, M., and Keim, D., "Processing online news streams for large-scale semantic analysis," in [*1st International Workshop on Data Engineering meets the Semantic Web*], (2010).

[15] Kleinberg, J., "Temporal dynamics of on-line information streams," in [*Data Stream Management: Processing High-Speed Data Streams*], Springer (2006).

[16] Osinski, S. and Weiss, D., "Carrot^2: Design of a flexible and efficient web information retrieval framework," in [*AWIC*], 439–444 (2005).

[17] Osinski, S., Stefanowski, J., and Weiss, D., "Lingo: Search results clustering algorithm based on singular value decomposition," in [*Intelligent Information Systems*], 359–368 (2004).

[18] Salton, G., Wong, A., and Yang, C. S., "A vector space model for automatic indexing," *Commun. ACM* **18**, 613–620 (November 1975).

[19] Zamir, O. E., "Clustering web documents: A phrase-based method for grouping search engine results," tech. rep. (1999).

[20] Stefanowski, J. and Weiss, D., "Carrot and language properties in web search results clustering," in [*AWIC*], 240–249 (2003).

[21] Harrower, M. and Brewer, C. A., [*ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps*], 261–268, John Wiley and Sons, Ltd (2011).

[22] Sugiyama, K., Tagawa, S., and Toda, M., "Methods for Visual Understanding of Hierarchical System Structures," *IEEE Transactions on Systems, Man, and Cybernetics* **11**(2), 109–125 (1981).