

Interactive Timeline Visualization of Documents

Leandro S. Guedes, Rafael Garcia, Bruno Pagno, Luciana Nedel, João Comba and Carla M.D.S. Freitas

Universidade Federal do Rio Grande do Sul (UFRGS)

Porto Alegre, Brazil

{lsguedes, rgarcia, blpagno, nedel, comba, carla}@inf.ufrgs.br

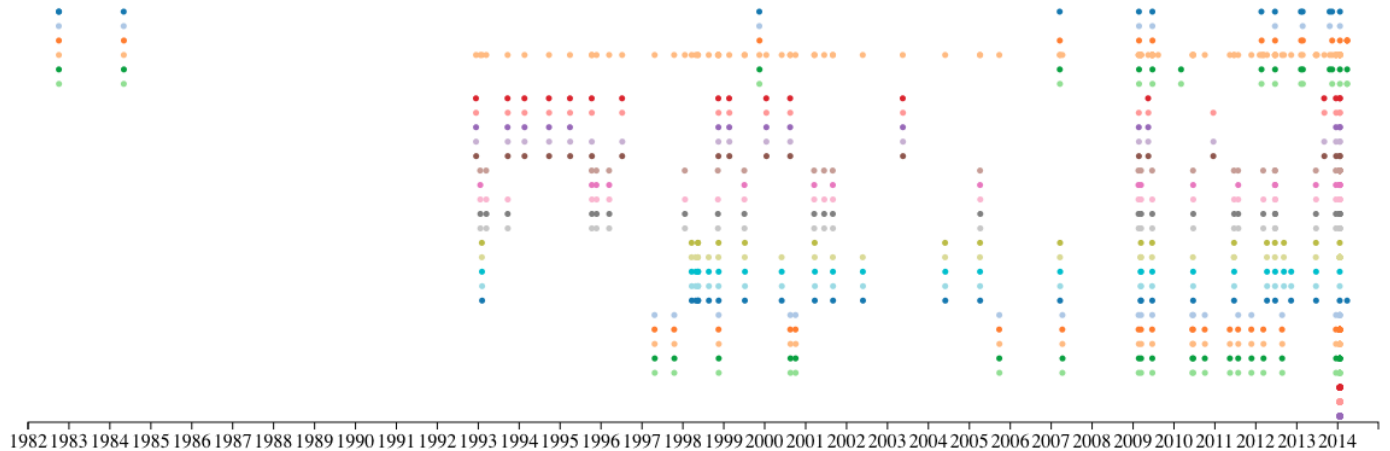


Figure 1. Timeline overview from 1982 to 2014, in this timeline, each line means a different source of documents.

Abstract—In this paper we present an approach for exploring a set of varied documents by means of a data visualization method aiming at enabling the discovery of specific facts for which there might be hints in some of the documents. The approach consists of two phases, the mining of relevant tags in the set of available documents and the actual visualization of the large amount of blog, magazines and newspapers reports. The mining phase transforms raw data into an organized file with tags related to the facts one wants to discover. The visualization phase is based on an interactive timeline that shows the documents over the years. Interactive features allow to query specific periods in the timeline as well as to filter the articles according to the tags found in the previous phase, providing a faster and efficient way of observing the presented data, and facilitating the recognition of the articles that have some interconnection, either by tags in common, source or time proximity.

Keywords—timeline visualization; documents visualization.

I. INTRODUCTION

This work was motivated by the Mini-Challenge 1 proposed as part of the VAST Challenge 2014[1]. The challenge consisted in responding a series of questions about the disappearance of a group of employees of a powerful fictional company called GASTech. The scenario is the fictional country Kronos, and the main suspect in the kidnapping is an organization known as POK (Protectors of Kronos). In order to corroborate the answers, we needed to create visualizations, which of course also allowed the analysis of the data provided within the challenge proposal.

The available documents for solving the problem were a map of Kronos, a chart describing the company organization, a spreadsheet containing the GASTech employees records, a text file containing the headers of all e-mails sent by the employees in the last weeks, short resumes and biographies of some GASTech employees, a document with historical data about Kronos and a large set (844 articles) of news reports of the last decades containing relevant information about the involved organizations.

We tackled the problem in two phases, the first for mining specific data in the set of news articles, and the second performing the exploratory task using the visualization we created. The visualization is basically an interactive timeline showing all documents along the years, separated by source (Figure 1). As we will describe later on, the visualization can be modified by the user through zooming, panning and filtering. We also show a map of the tags relevant for the problem, which were found in all documents, and details on demand in a way similar to a tooltip as well as the content of a selected document in a separate view.

Although they were created for and used specifically in this challenge, the methods presented herein are fully adaptable to any similar problem involving a large amount of data comprising a wide period of time.

In the next sections we briefly comment related work (section II), present our approach (section III) and discuss our results (section IV). Finally, in section V we draw some conclusions and comment on future work.

II. RELATED WORK

Background concepts and techniques for our work can be divided into two classes: document and text mining and visualization of temporal data. However, since our text mining is very simple up to now, we restrain ourselves to comment only visualization of temporal data, more specifically the timeline-based visualizations.

There are several works that use a timeline based visualization for solving specific problems. The need of representing temporal data is quite old. For example, in 1812, Charles Minard made a visual history of Napoleon's Russian Campaign where he illustrated its disastrous result in a graph. This classic example and other historical visualizations can be seen on [2] and on Tufte's classical book [3].

Another well-known example is Themeriver [4], which is a visualization of the thematic variations over time across a collection of documents. The flow is shown within the context of a timeline and a corresponding textual presentation of external events.

In a rather different context, Wongsuphasawat et al. [5] developed LifeFlow, where they introduced an interactive visual overview of event sequences applied to health research. In their tests, novice users were able to rapidly answer questions about the prevalence and temporal characteristics of sequences, find anomalies, and gain significant insight from the data.

Two more recent examples of timeline visualizations are by Craig and Roa-Seiler [6] and Wang and Yuan [7]. Craig and Roa-Seiler [6] used a vertical timeline to support the analysis of human-computer dialogue data, while Wang and Yuan [7] employ a timeline visualization for comparing urban trajectories.

A work closer to ours is by Pak et al. [8]. They also use data mining applied to huge datasets for finding topics that occur in a corpus along seven years. The timeline visualization helps finding sequential patterns formed by temporal sequence of those topics. Unlike our work, however, they show the appearance of topics along time while we decided to represent the news sources along the timeline.

III. DOCUMENTS TEMPORAL VISUALIZATION

The raw dataset provided within the VAST challenge proposal comprises articles with different structures. Most of the files have a header with some information like title, date and source. However, in this dataset, such information is not easily recognized due to the articles not necessarily have all of them. Also, the order and the format in which these items appear in the documents are variable, requiring a more sophisticated parsing strategy. Although all documents are written in english, a reasonable amount of them were originally written in other languages and translated by an automatic translator, making the parsing even more difficult, as this may cause the sentences have a confusing meaning due to translation errors.

The first step was to standardize the articles content, so the timeline visualization could easily handle these files. Although the files have different formats, it was possible to recognize

that they were divided into few groups of articles with similar formats. So, we first manually recognized these formats and implement parsers to read and standardize these documents. They were organized by "content", "id", "authors", "title", "source", "date" and the most important, the "tags" we were in need for solving the motivating problem.

To find the set of tags of each article, we looked for the names of all known people related with some organization (POK and GAS tech in our scenario). These names were taken from the spreadsheet of employee records, from the csv files with the e-mail headers, from the historical reports and from the biographies of the employees. The tags of each article represent all known people mentioned in its content.

By the end of this phase, we have a JSON file as a ready-to-use database.

A. Timeline-based visualization

Our dataset contains a reasonably large amount of articles separated by date, which fits perfectly in a timeline. Each article is represented on the timeline as a circular marker, with the x coordinate being used for its temporal position in the timeline, and the y coordinate serving to separate the articles' sources, facilitating the identification of related reports, since it was known that likely the same source would contain several reports about the same subject, especially in a short period of time.

The colors of the circles also facilitate the identification of the report source, having been assigned a different color for each source. The overview of this timeline is presented in Figure 1. Also, there is an option to show all the tagged articles. In this case, the articles with no associated tags were presented in light blue, and the tagged ones in dark blue. When there are more than one tagged article in the same point, the tagged one comes first. This feature can be observed in Figure 2.

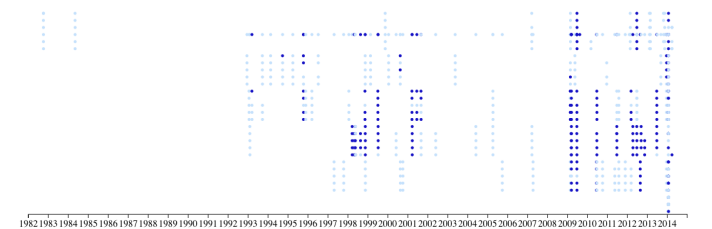


Figure 2. Showing all the tagged articles.

As the articles are distributed over a relatively long period of time, a static timeline would become cluttered, with many elements sharing similar positions. To avoid this, we built a timeline that allows zooming, causing only reports of a given year remain visible. This zoom can be extended, allowing a monthly visualization of the articles. The different zooming levels can be observed in Figure 3.

The identification of the article content is possible by hovering the mouse over the corresponding marker, showing the item number, the name of the newspaper in which it was

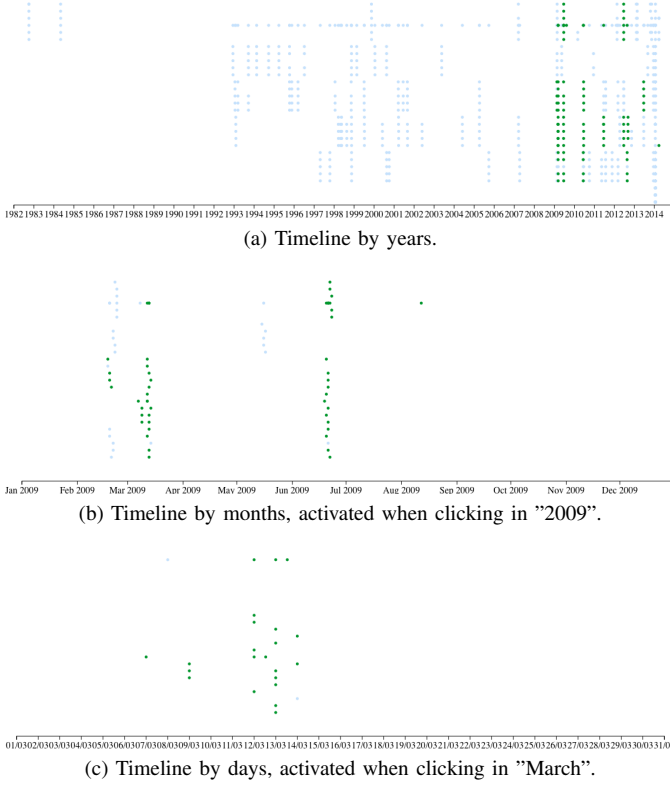


Figure 3. Visualizing articles mentioning "Elian Karel".

published, its tags and the amount of articles posted by the same source on the same day. This feature can be seen in Figure 4.



Figure 4. Identification of the article shown with the mouse over its marker.

B. Map of tags

Since there is a large number of tags, we implemented a map of tags, which is actually a set of buttons. They allow filtering all the reports, making it possible to easily identify those that contain keywords in common. Clicking on the marker of an article causes its content to appear in an information

field on the same page but without disturbing the timeline visualization. The whole page can be observed in Figure 5.

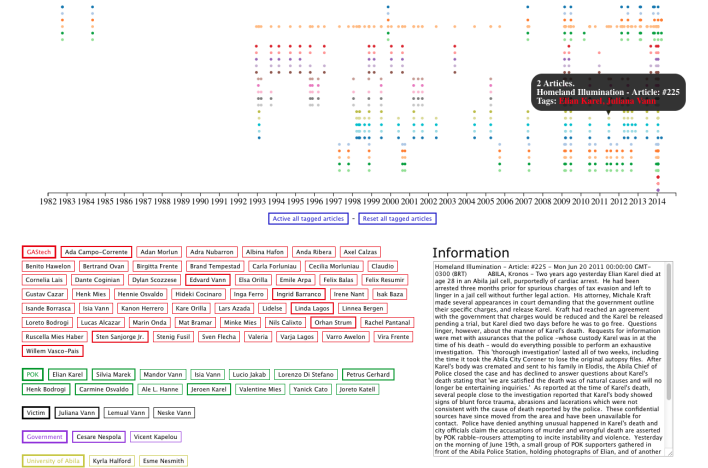


Figure 5. Page overview. Timeline is on top, buttons on the lower left side and article content on the lower right.

Such approach facilitate the observation of key points in the sequence of events of the proposed problem, allowing to quickly identify related articles, and navigate among them, also enabling a better understanding of the events reported in these articles without the need of reading all of them.

IV. RESULTS AND DISCUSSION

Our technique was implemented in Javascript and HTML5, the database being a JSON file. As for the visualization, we adopted the Data-Driven Documents (D3) library [9].

We discuss our results using the motivating problem for showing the positive aspects that the visualization provided in the analysis of the 844 articles.

The articles distribution along the timeline shows that from 1985 to 1992 no articles were produced, as well as in 1983, 2006 and 2008. Also, the first group of articles was published in October 1982. A quick look at the source of these reports shows that all of them were published in international news, allowing the understanding that this is a relevant issue even out of Kronos. When checking the contents of these reports, it is noticeable that they are emphasizing the growth and the economic power of GASTech, crucial information to understand the background kidnapping of its employees. A second group of articles appear after two years, allowing the user to quickly obtain the information that GASTech uses a controversial gas extraction technique, which can contaminate the water.

Filtering the articles by the tags on the map also greatly facilitates decision making, since it allows the user to decide reading only the reports of his/her interest. When selecting items with the tag of some GASTech employees, relevant information can be easily found, as for example, suspicion that, among the kidnapped employees, are Ingrid Barranco, Orhan Strum and Ada Campo-Corrente. An important information is

that the security Edvard Vann went through a long interrogation due to suspicion of involvement with the kidnapping, since he has the same surname of several known POK members.

Clearly, by looking at and interacting with the visualization, we were able to conclude some important facts that would only be possible by reading all the articles to make the adequate connections.

V. FINAL COMMENTS

In this paper, we employed a timeline visualization technique for showing documents containing news articles, which underwent a preprocessing procedure for extracting relevant data, and identifying whether they mention a specific content. Articles were tagged based on this specific content. Our timeline visualization allows to instantly recognize periods with a large quantity of reports, and the associated map of tags allows to filter articles per tag retaining in the timeline only those that comply with the selected tags. Other interactive features provide access to information and content of a selected article.

As future work, we intend to tag and display the employees' e-mails, so we would have better insights without the need of navigating through different visualizations at the same time. This will impose the need of designing a way of showing in the same visualization existing relations between e-mails and news, as for example, co-occurrence of tags and temporal sequence of tags in different sources, just to name a few possibilities.

We also intend to perform tests with users since up to now we were the only users of our technique for solving the problem of the VAST challenge. Another possibility of future work is to apply the technique with a different set of news sources incorporating more sophisticated data mining techniques.

ACKNOWLEDGMENTS

We are grateful to CAPES, CNPq and FAPERGS for funding our current studies.

REFERENCES

- [1] Vast challenge 2014. [Online]. Available: <http://vacommunity.org/VAST+Challenge+2014>
- [2] Timelines and visual histories. [Online]. Available: <http://www.datavis.ca/gallery/timelines.php>
- [3] E. R. Tufte, *The Visual Display of Quantitative Information*. Cheshire, CT, USA: Graphics Press, 1986.
- [4] S. Havre, B. Hetzler, and L. Nowell, "Themeriver: visualizing theme changes over time," in *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, 2000, pp. 115–123.
- [5] K. Wongsuphasawat, J. A. Guerra Gómez, C. Plaisant, T. D. Wang, M. Taieb-Maimon, and B. Shneiderman, "Lifeflow: Visualizing an overview of event sequences," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA: ACM, 2011, pp. 1747–1756. [Online]. Available: <http://doi.acm.org/10.1145/1978942.1979196>
- [6] P. Craig and N. Roa-Seiler, "A vertical timeline visualization for the exploratory analysis of dialogue data," in *Information Visualisation (IV), 2012 16th International Conference on*, July 2012, pp. 68–73.
- [7] Z. Wang and X. Yuan, "Urban trajectory timeline visualization," in *Big Data and Smart Computing (BIGCOMP), 2014 International Conference on*, Jan 2014, pp. 13–18.
- [8] P. C. Wong, W. Cowley, H. Foote, E. Jurrus, and J. Thomas, "Visualizing sequential patterns for text mining," in *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, 2000, pp. 105–111.
- [9] Data-driven documents. [Online]. Available: <http://d3js.org/>