Vortrag gehalten bei: Task-Driven Analysis of Social Media : The 2nd Workshop on Interactive Visual Text Analytics. Part of the VisWeek 2012, October 14-19, 2012, Seattle

Topic Tracker: Shape-based Visualization for Trend and Sentiment Tracking in Twitter

Franz Wanner[1], **Andreas Weiler**[2], **Tobias Schreck**[1]
[1]Data Analysis and Visualization Group, University of Konstanz
[2]Database and Information Systems Group, University of Konstanz
{firstname.lastname}@uni-konstanz.de

ABSTRACT
In recent years there has been a continuous development of social media services on the web. Unprecedented success and active usage of these services result in massive amounts of user-generated data. Visual representation of these large amounts of unevenly distributed time series data is a challenging task, especially while preserving access to individual data points. Our hypothesis is that shape-based visual representations have advantages over established time series compression visualizations like Two-Tone Pseudo Coloring or line graphs. In this paper we present a shape-based visualization for trend and sentiment tracking of user-defined topics in the Twitter data stream. We use glyphs to visualize the appearance and sentiment of tweets on a timeline and enable analysts to keep track of the trend of their defined topic and the corresponding sentiment expressed by the Twitter users.

Author Keywords
Topic and Sentiment Tracking;Shape-based Visualization;Twitter

ACM Classification Keywords
H.3.3 Information Systems: Information Search and Retrieval

INTRODUCTION

Time series data is appearing in many different areas and applications, such as in analysis tasks for monitoring log data or in tracking scenarios. Often data is sampled in regular intervals resulting in evenly distributed data points, which means that the data items are equidistant over time. Nevertheless, unevenly distributed time series data, like data from social media streams, are common in many applications and more challenging to deal with.

As social media services changed the way of communications on the Internet and play an increasing role in our daily life, it was only a question of time until it became a source for information gathering. With over 200 million registered users and about 220 million messages per day Twitter became the undisputed number one in social microblogging nowadays. Unfortunately, the vast amount of data and the high variability in the quality of user-generated data is obstructive for analysis tasks. A challenge is, to find a way to visualize the often vast amounts of user-generated data adequately to the analyst or information seeker and guide the perception to interesting periods in time. In this case adequate means to save space without loosing too much information at once and to support the user fulfilling and solving analytical tasks. In our data interesting points or episodes are regions showing a high density of data points. For example, our system supports an analyst with the task to keep track of a specific topic (e.g., trend of a cinema film) and the sentiment of the reactions to the topic by offering an overview about the frequency of tweets and resulting patterns for negative or positive sentiment.

DATA SOURCE

Twitter provides an API [11] for direct access to the public live stream of Twitter messages ("tweets") for application developers. As we are connected with the so-called "gardenhose" access level, we are able to consume and collect 10% of the public live stream. Figure 1 displays statistics about the amount of incoming tweets per hour for a representative sample of days. Consequently, we obtain more than 30 million tweets per day with an average of more than one million tweets per hour.

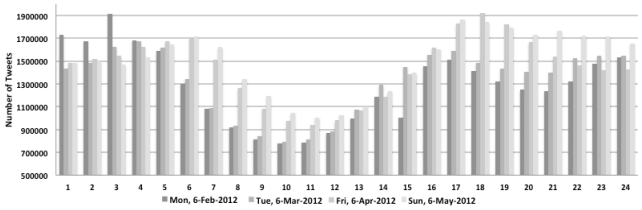

Figure 1. Amount of incoming tweets per hour on four exemplary days.

The data of Twitter consists not only of the tweets themselves, but also contains a very large amount of meta-data on the tweet (e.g., count of retweets, geographic location) and on the user's profile (e.g., count of followers). Each tweet object is streamed in the semi-structured JSON format containing 67 data fields. To cope with the massive data stream and to facilitate the access to the data we built our system on top of an extended version of a native XML database [24]. With this solution we are able to keep up with the live data stream and can store all incoming tweets in proper time.

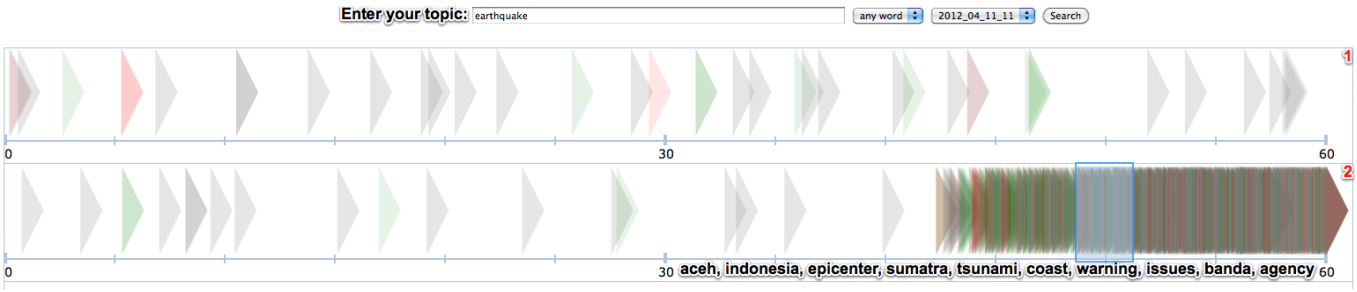

Figure 2. Trend of the topic with the keyword "earthquake" for the hours during the time of an earthquake happening in Aceh, Indonesia. The first line shows the hour before the earthquake (07:00 - 08:00 AM UTC) and the next line (2) the hour of the earthquake event. On top of the visualization is the search panel for entering topic keywords, choosing the time frame and options such as "any word" or "all words". By selecting a range of tweets the ten most frequent common terms of that area are displayed below the visualization (2).

SYSTEM DESIGN

Our approach is a system for tracking the trend and sentiment of self-defined topics by analysts, information seekers or normal users. We implemented a web frontend to enter topic terms, set several options and execute the analysis. For each analysis a new line of results is added to the frontend and therefore it is easy to keep track of previous results and to compare results with different keywords or timeframes. To obtain the results we use the above-mentioned native XML database and the full-text functionality of XQuery. For result presentation we use a shape-based visualization [20], in which each data item is represented by a shape object.

To retain the original timeline of the tweet objects, we also use a timeline as the fundamental drawing object and plot the resulting tweet objects at the position corresponding to their creation date onto the timeline. Due to the fact that the result of an analysis can be very large and we want to preserve access to individual data points we choose an equal-sided triangle with its peak to the right to represent single tweets in the visualization. With this shape it is possible to have an overlap between the objects, but also the possibility to recognize single data objects in the timeline. The shape also expresses the short lifetime of tweets and the temporal overlapping of the incoming messages. The filling color of the shape signifies the sentiment (red = negative, green = positive, grey = neutral) of the tweet and the opacity of the shape expresses the value of the sentiment. The value of the sentiment is calculated by using an analysis tool from Thelwall et al. [19], which analyzes the tweets and returns values from -5 (extremely negative) to 5 (extremely positive). The sentiment results are converted to opacity values with a multiplying factor of -0.1 for negative and 0.1 for positive sentiment. By selecting a range of tweets the system analyzes the corresponding tweets and displays the ten most frequent common terms below the result line. As the topic terms are always very common and frequent in the resulting tweets, they are excluded from the frequent common terms list.

USE CASES

Since the amount of tweets increases daily it becomes a crucial task to keep track of specific topics. For example, our system supports users, which are interested in the emergence of natural disasters (e.g., earthquakes), the popularity of a movie, or further events such as sport events.

Topic Trend Tracking

Tracking of Topic Trends is a task that needs to be done in various fields of analytics. For example, an analyst who wants to keep track of natural disasters appearing in the Twitter stream needs an appropriate tool, which displays a compact overview of all appearances of the topic in the data. Another example use case of our system, is a brand manager of a company who is interested in mentions of their company name or new products of their company. By using our system it is an easy task for the manager to keep track about the defined topic. Additionally it is possible to define topics of rival companies or products and compare the results with each other.

Figure 2 shows the resulting visualization of a scenario, in which we keep track of a topic about natural disasters. The topic is described by a single keyword "earthquake" and we choose two hours of Twitter data. The first line shows the hour from 07:00 to 08:00 AM (UTC) on April 11, 2012 with only a few appearances of the topic. We can derive that at this time no earthquake happened or at least not many users reported about it. The second line shows the hour from 08:00 to 09:00 AM. We can discover a high occurrence of tweets in the last 15 minutes of this hour. This reflects the happening of an earthquake in Aceh (Indonesia) at 8:38 AM on that day. By selecting a range of tweets in the second line (from about 8:49 to 8:51) the following terms are shown: "aceh, indonesia, epicenter, sumatra, tsunami, coast, warning, issues, banda, agency". Hereby we can be sure that the shown tweets report about the above mentioned earthquake. These frequent terms support the user in defining a more specific topic such as "earthquake, indonesia, aceh", by recommending more terms which can be added to the topic definition. By activating the full-text option "all words" the system only returns tweets containing all defined terms.

Topic Sentiment Tracking

Since the large amount of tweets represents a high variety of opinions for a topic, we combine the tracking of a topic trend with the tracking of the sentiment. An analyst who detects interesting occurrences of tweets for their defined topic is also mostly interested in the reactions and emotions about the topic. For example, a brand manager can get an overview about reactions to a certain product and compare the results with reactions to a product from a rival company.

We present a scenario in which we keep track of a topic about

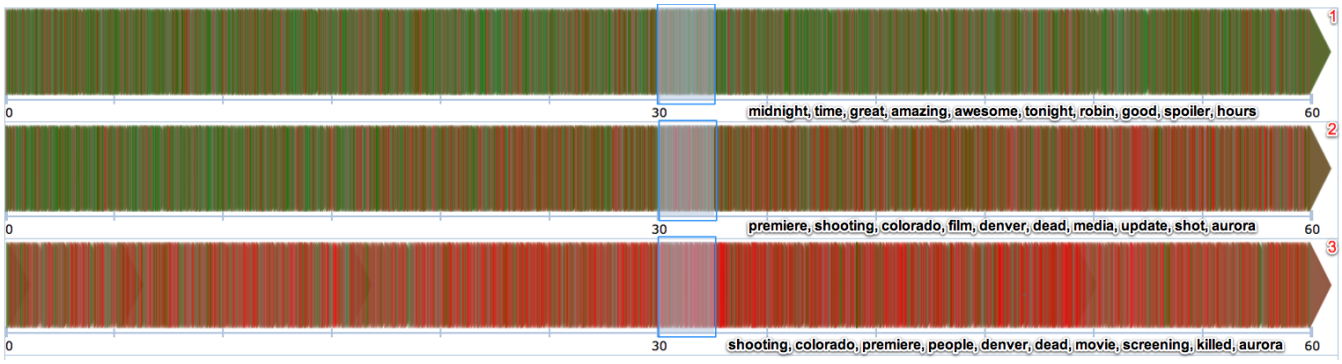

Figure 3. Trend of the topic with the keywords "dark knight", "dark knight rises" or "batman" for the hours after the film premiere of the new batman movie "The Dark Knight Rises" in Aurora, Colorado. The first line shows the hour after the premiere start (07:00 - 08:00 AM UTC) and the next two lines (2 & 3) the following two hours.

a new cinema movie, which is described by three keywords "dark knight", "dark knight rises", and "batman". We choose the full-text option "any word" to find all occurrences of the three search terms. Figure 3 shows the resulting visualization for three hours of Twitter data. The first line shows the hour from 07:00 to 08:00 AM (UTC) on July 20, 2012 with a majority of positive (green colored) sentiment. In the second line, which shows the hour from 08:00 to 09:00 AM, the sentiment changes after roughly half an hour from positive to a more negative one. The third line (09:00 to 10:00 AM) shows a majority of negative sentiment. Through the help of the range selection (3 minutes from the beginning of the second half of the hour) in the three lines and the view of the most frequent common terms we get an idea about the change of the sentiment. The ten most frequent common terms for the three hours are the following:

1. **Hour:** midnight, time, great, amazing, awesome, tonight, robin, good, spoiler, hours
2. **Hour:** premiere, shooting, colorado, film, denver, dead, media, update, shot, aurora
3. **Hour:** shooting, colorado, premiere, people, denver, dead, movie, screening, killed, aurora

From the list of terms we can derive that the tweets in the first hour are mostly about the positive reactions to the premiere of the movie. Since there is also some criticism about the movie, we also obtain some negative parts in the visualization. The terms in the second hour change from reactions about the movie to reactions about a shooting taking place during the film premiere (6:38 AM UTC) in a cinema in Aurora, Colorado. This trend continues for the third line, which consists almost entirely of messages with negative sentiment. It can therefore be concluded that reactions to the shooting form the majority in the third hour.

RELATED WORK

In the 18th century William Playfair started the ongoing success story of visualizing time series with a visualization of trade time series [16]. Since the data volume increased during the last decades and nowadays almost all data has a time component, much research has been carried out in the field of time-based data visualization techniques ([4], [3], [12]).

In multidimensional data visualizations, glyph and shape based representations are known for years ([22], [23]). Ward also defines glyphs as graphical entities, which are able to map more data values onto their visual attributes. These attributes can be shape, size, color, and position [22]. They are therefore suitable for quantitative data values, what Aigner et al. state in [2]: they "can be directly mapped to visual variables such as position, length, and orientation". In such cases glyphs can be very useful, because of many mapping possibilities. Hao et. al wrote in [7] that for the visual exploration of time series considering the visualization method, any technique, including glyph representations, is possible. Glyphs are furthermore used in the network area to visualize network traffic [5]. In [9] a clock metaphor is used to monitor large IP spaces. More sophisticated glyphs for network service and maintenance are shown in [14]. Suntinger et al. created a "spherical event glyph" to encode event attributes in a visual representation [17]. Uncertainty visualization, a topic which is ongoing was already addressed in 2005 by Aigner et al. using glyphs [1]. Further usage of glyphs can be found in [18]. Starting and ending time, as an extension of [15], are encoded in a glyph [10]. An interesting approach using shape-based glyphs can also be found in [13].

An approach for visual density estimation and event detection in news streams is shown in [8]. The authors try to capture the problem of high density and overplotting via an importance function. As graphical representation they use circles. The output of the importance function is mapped on the opacity and size of the filled circles. Another application in the area of news streams can be found in [20] and [21]. Fekete and Plaisant in 2002 also stated, that a smooth shading of rectangles helps "to distinguish items in dense visualizations" [6].

CONCLUSIONS AND FUTURE WORK

In this paper, we demonstrated a system that enables analysts and information seekers to explore and browse contents of large amounts of tweets. In addition to the exploration task, the system supports users in keeping track of the trend and sentiment of their self defined topics. Since topics can be defined by an enumeration of keywords there is no domain limitation for the analysts.

Future work includes the live update of the visualization directly connected to the live Twitter stream and the implemen-

tation of correlation with other streaming sources such as RSS feeds or further news sources. In this context, it would also be possible to calculate the opacity value of shapes with other dimensions then the sentiment. Further ideas are to extend the system with zooming to provide a more detailed view to the users. Since the amount of tweets can be very high and the visualization can become cluttered we also plan to add a filter function to display only a certain type (e.g., only negative or positive sentiment) of tweets. Further options could be derived from the meta-data of tweets, such as the source of the tweet (e.g., mobile phone or web), the geographic region of the tweet or the type of the tweet (e.g., retweet, direct message or standard message). For a more precise search and to improve the results more full-text options, such as fuzzy search or the exclusion of negative terms can be added. The importance of the exclusion of negative terms can be derived from the second use case, in which the results shifted from tweets about the original topic (reactions to a movie premiere) to a different topic (reactions to a mass shooting).

REFERENCES

1. Aigner, W., et al. Planninglines: Novel glyphs for representing temporal uncertainties and their evaluation. In *Information Visualisation, 2005. Proceedings. Ninth International Conference on*, IEEE (2005), 457–463.
2. Aigner, W., et al. Towards a conceptual framework for visual analytics of time and time-oriented data. In *Simulation Conference, 2007 Winter*, IEEE (2007), 721–729.
3. Aigner, W., et al. Visual methods for analyzing time-oriented data. *IEEE Transactions On Visualization and Computer Graphics* (2007), 47–60.
4. Aigner, W., et al. *Visualization of time-oriented data*. Springer-Verlag New York Inc, 2011.
5. Becker, R., et al. Visualizing network data. *Visualization and Computer Graphics, IEEE Transactions on 1*, 1 (1995), 16–28.
6. Fekete, J., and Plaisant, C. Interactive information visualization of a million items. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*, IEEE (2002), 117–124.
7. Hao, M., et al. Multi-Resolution Techniques for Visual Exploration of Large Time-Series Data. In *Eurographics/IEEE-VGTC Symposium on Visualization, 23 - 25 May 2007, Norrkoeping, Sweden* (2007).
8. Keim, D. A., et al. CloudLines: Compact Display of Event Episodes in Multiple Time-Series. *IEEE Transactions on Visualization and Computer Graphics 17* (2011), 2432–2439.
9. Kintzel, C., et al. Monitoring large ip spaces with clockview. In *Proceedings of the 8th International Symposium on Visualization for Cyber Security*, ACM (2011), 2.
10. Kosara, R., and Miksch, S. Visualizing complex notions of time. *Studies in Health Technology and Informatics*, 1 (2001), 211–215.
11. Krikorian, R. Developing for @twitterapi (Techcrunch Disrupt Hackathon).
12. Muller, W., and Schumann, H. Visualization methods for time-dependent data-an overview. In *Simulation Conference, 2003. Proceedings of the 2003 Winter*, vol. 1, IEEE (2003), 737–745.
13. Overby, D., et al. Interactive visual analysis of location reporting patterns. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, IEEE (2009), 223–224.
14. Pearlman, J., et al. Visualizing network security events using compound glyphs from a service-oriented perspective. *VizSEC 2007* (2008), 131–146.
15. Plaisant, C., et al. Lifelines: visualizing personal histories. In *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground*, ACM (1996), 221–227.
16. Playfair, W., and Corry, J. *The commercial and political atlas*. Printed for J. Debrett; GG and J. Robinson; J. Sewell; the engraver, SJ Neele; W. Creech and C. Elliot, Edinburgh; and L. White, Dublin, 1786.
17. Suntinger, M., et al. The event tunnel: Interactive visualization of complex event streams for business process pattern analysis. In *Visualization Symposium, 2008. PacificVIS'08. IEEE Pacific*, IEEE (2008), 111–118.
18. Tekusova, T., and Kohlhammer, J. Applying animation to the visual analysis of financial time-dependent data. In *Information Visualization, 2007. IV'07. 11th International Conference*, IEEE (2007), 101–108.
19. Thelwall, M., et al. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology 61*, 12 (2010), 2544–2558.
20. Wanner, F., et al. Large-scale Comparative Sentiment Analysis of News Articles (InfoVis 2009). Presented at the poster session at IEEE Information Visualization Conference 2009 (IEEE InfoVis 2009), Atlantic City, USA, 2009.
21. Wanner, F., et al. Visual Sentiment Analysis of RSS News Feeds Featuring the US Presidential Election in 2008. In *Proceedings of the IUI'09 Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW 2009)* (2009).
22. Ward, M. A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization 1*, 3-4 (2002), 194–210.
23. Ward, M. Multivariate data glyphs: Principles and practice. *Handbook of Data Visualization* (2008), 179–198.
24. Weiler, A., et al. Towards an advanced system for real-time event detection in high-volume data streams. In *Proceedings of the 5th workshop for Ph.D. students on Information and Knowledge Management*, PIKM '12, ACM (2012).