Visual Analysis of Microblog Content Using Time-Varying Co-occurrence Highlighting in Tag Clouds

Steffen Lohmann, Michael Burch, Hansjörg Schmauder, Daniel Weiskopf

VIS/VISUS, University of Stuttgart, Germany

steffen.lohmann@vis.uni-stuttgart.de, michael.burch@visus.uni-stuttgart.de,
schmauhg@studi.informatik.uni-stuttgart.de, daniel.weiskopf@visus.uni-stuttgart.de

ABSTRACT

The vast amount of contents posted to microblogging services each day offers a rich source of information for analytical tasks. The aggregated posts provide a broad sense of the informal conversations complementing other media. However, analyzing the textual content is challenging due to its large volume, heterogeneity, and time-dependence. In this paper, we exploit the idea of tag clouds to visually analyze microblog content. As a major contribution, tag clouds are extended by an interactive visualization technique that we refer to as *time-varying co-occurrence highlighting*. It combines colored histograms with visual highlighting of co-occurrences, thus allowing for a time-dependent analysis of term relations. An example dataset of Twitter posts illustrates the applicability and usefulness of the approach.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Graphical user interfaces (GUI)*

General Terms

Design, Human Factors

Keywords

Tag cloud, co-occurrence highlighting, time-series visualization, histogram, microblogging, microposts, twitter, visual analysis

1. INTRODUCTION

Social networking and microblogging services such as Twitter, Facebook, or Google+ allow people to broadcast short messages, so-called *microposts*, in continuous streams. The posts usually consist of a text message enriched with contextual metadata, such as the author, date and time, and sometimes also the location of origin. While an individual post is small in size (up to 140 characters in Twitter) and of limited information value, aggregated posts of multiple users provide a rich source of time-critical information that can point to events and trends needing attention [6, 12, 15].

AVI'12, May 21-25, 2012, Capri Island, Italy

Accordingly, there is a growing interest in leveraging microposts for analytical tasks in various application areas, such as crisis management, journalism, or politics [12, 13]. However, analyzing the textual content is challenging due to its large volume, heterogeneity, and time-dependence. Methods and tools are needed that assist in the exploration of microposts and support the identification of relevant information.

A straightforward means to visually depict the contents of microposts are tag clouds [16]. They proved useful in summarizing textual content by using Natural Language Processing (NLP) and displaying the extracted terms weighted by their frequency of use. However, tag clouds typically depict the terms and term frequencies summed up over time and provide no information on time-varying term use. Furthermore, they show only the list of terms, without telling how the terms are related to each other. Both would be useful information in the analysis of microposts due to the highly dynamic and interrelated nature of the contents.

To overcome these limitations, we developed an advanced tag cloud visualization tailored to the visual analysis of microposts. In particular, the tag cloud is extended by an interactive visualization technique that we refer to as *time-varying co-occurrence highlighting*. It uses visual highlighting of co-occurrences combined with colored histograms to indicate time-dependent term relations. This way, it is possible to visually explore a large set of terms from microblogs and discover relationships, events, and trends of relevance in a certain context. The associated microposts can additionally be displayed on a map, if they are geocoded.

2. RELATED WORK

There have already been some attempts to use tag clouds for the visualization of microposts. While most works (e.g. [13, 14]) simply process microblog contents with tag cloud generators like *Wordle* [3], others combine the idea of tag clouds with a map-based representation of extracted terms [1, 15]. The microposts' geocoordinates are used to place the terms on a map, a visualization technique that has been called *tag map* in the context of annotated photographs [8]. However, all these approaches use static tag cloud representations that neither provide information on the interrelation of terms nor on their temporal evolution.

The latter is addressed by approaches that encode time information in tag clouds, such as *SparkClouds* [9] or *ParallelTagClouds* [5]. Parallel tag clouds combine the ideas of tag clouds and parallel coordinates to show term frequencies at multiple points in time simultaneously. This technique allows to compare changes in term use,

To appear in:
Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI 2012)
New York, NY, USA: ACM, 2012

but it does not directly visualize the evolution of the terms. This is different in the SparkClouds approach where a sparkline (i.e. a simplified line graph) is added to each tag to explicitly show changes in term use over time. Though useful improvements to tag clouds, neither SparkClouds nor ParallelTagClouds show term relations and are thus not able to visualize time-varying co-occurrences. This is also true for other visualizations that use line graphs or stack graphs to illustrate term evolution in microposts, such as *Topic Streams* [6] or *Twitter StreamGraphs* [4].

Co-occurrence information is, in turn, considered by approaches that visualize term relations in tag clouds, such as *clustered tag clouds* [11] or *tag graphs* [7, 10]. Clustered tag clouds implicitly indicate term relations by distance, e.g. by placing frequently co-occurring terms close to each other. Tag graphs make the term relations explicit by connecting co-occurring terms with edges. However, in contrast to the above works, both approaches do not contain information about time-varying changes in tag use.

The time-varying co-occurrence highlighting presented in this paper combines both directions of extending tag clouds for visual analysis. It integrates information about the time-dependence and co-occurrence of terms in one interactive visualization. It therefore takes into account that both can be important information in the analysis of microblog content and are best explored in an integrated manner.

3. TIME-VARYING CO-OCCURRENCE HIGHLIGHTING

Essentially, our approach is based on two visual features:

- Each term in the tag cloud is enhanced by a histogram that visually depicts the variation in term use over time.
- Whenever a term is selected in the tag cloud, related terms are visually highlighted and co-occurrence information is shown.

Both visual features are integrated as follows:

- Along with the co-occurrence highlighting, those bars in the histogram are colored that represent the point in time of co-occurrence.

Fig. 1 shows a part of an example tag cloud that has been enhanced by time-varying co-occurrence highlighting. In this example, the term ‘center’ has been selected, as indicated by its red color and border. Terms that are used together with ‘center’ in the microposts are highlighted in yellow, with the number of co-occurrences given in parentheses. Terms that are used very often together with ‘center’ are additionally colored in red, such as the term ‘billie’ that co-occurs 275 times with ‘center’ in this example.

Those bars in the terms’ histograms that represent the point of co-occurrence are marked in red, while the other bars remain blue. For instance, in case of the term ‘2011’, only the left bars of the histogram are red, indicating that all co-occurrences of ‘center’ and ‘2011’ appear in the first half of the analyzed time span.

We distinguish between two forms of time-varying co-occurrence highlighting: In the *absolute* variant, each co-occurrence leads to a complete coloring of the corresponding bar in the histogram, as in Fig. 1. In the *relative* variant, the bars are only partly colored, indicating the number of co-occurrences in relation to the term frequency at the corresponding point in time.

Figure 1: Part of a tag cloud enhanced by absolute time-varying co-occurrence highlighting.

Figure 2: Another part of the tag cloud, this time with relative time-varying co-occurrence highlighting.

Fig. 2 illustrates the relative coloring for another part of the tag cloud. This time, the term ’open’ has been selected that co-occurs 565 times with the term ’tennis’ and fewer times with a number of other terms. This is not only indicated by the terms’ font colors but also by the relative coloring of the bars in the histogram (which is only visible for the term ’tennis’). In the following, we describe the creation of the interactive visualization in more detail.[1]

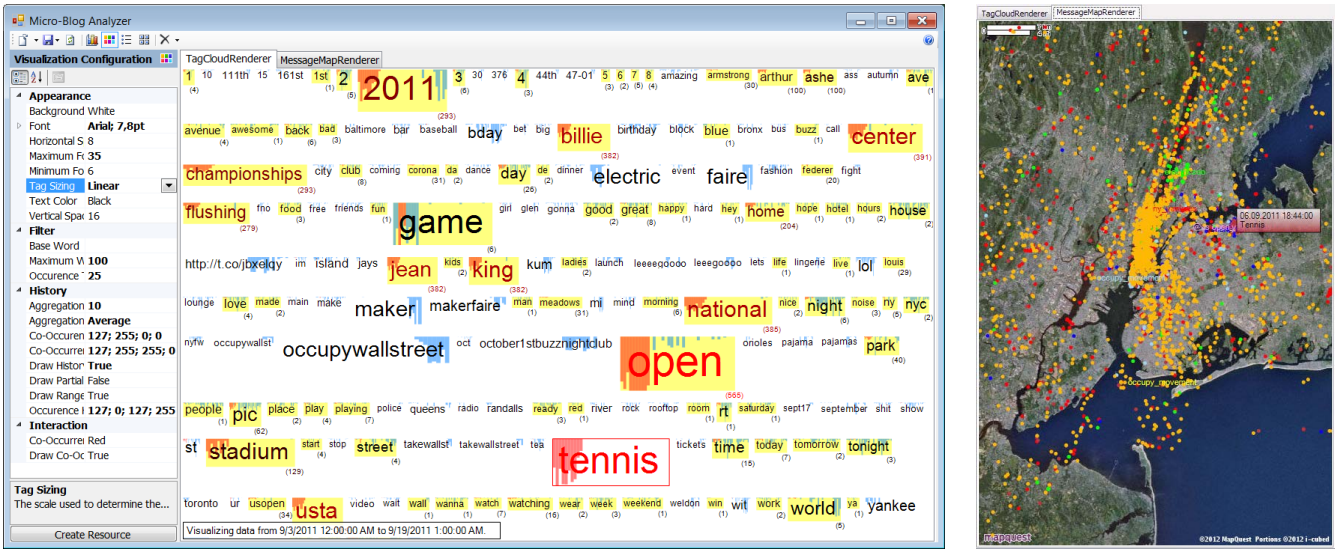
3.1 Retrieval of Microblog Contents

The microblog content was retrieved from a large database of Twitter posts created and maintained by Thom et al. [15]. Twitter is currently the most popular microblog service with more than 140 million users and around 340 million posts per day, according to its own statistics [2]. The database contains all publicly available posts that come with geo-coordinates, which is about 1 to 2 million posts per day. These posts are typically sent from mobile devices, such as smartphones.

For our example case, we used the geo-coordinates of the posts as initial filter and selected a subset restricted to an area of New York. Furthermore, we filtered the posts by date, using a time span from September 3 to 19, 2011. We then applied the *Spatiotemporal Anomaly Detector* offered by Thom et al. to find anomalies in the Twitter data that correspond to certain events. It analyses the spatiotemporal density, term usage, and number of contributing users to identify posts that are likely from local witnesses of an event (see [15] for a detailed description of the approach and the related algorithms).

We found events related to tennis, the New York Yankees, and Electric Zoo for days 3 to 5, and events related to Maker Faire as well as Occupy Movement for days 16 to 18. Because the term ‘party’ occurred most frequently, we used it to pre-filter the microblog

[1] A demo video is available at `http://microposts.visualdataweb.org`



(a) (b)

Figure 3: (a) Visual interface showing a tag cloud where the time-varying co-occurrences of the term ‘tennis’ are highlighted. (b) The geo-information of the posts displayed on a map with colors indicating events the posts likely belong to.

contents. We then removed it from the contents, as it would otherwise be associated with most microposts across all times, thus not adding any value. We finally got a sample of 4,934 posts that we used to examine the applicability and usefulness of time-varying co-occurrence highlighting.

3.2 Tag Cloud Generation and Adaptation

We processed the microblog contents with NLP techniques typically used in tag cloud generation, such as tokenization, lowercase conversion, and stop word removal [3, 6]. That is, we first split the text streams at whitespace characters to get the individual terms, then converted these terms to lowercase, and removed stop words, i.e. common words (such as ‘the’, ‘is’ and ‘at’) that usually do not carry meaning. Finally, we aggregated syntactically identical terms and printed them on the screen, with the font size scaled to their frequency of use. Fig. 3a shows the tag cloud generated for the microblog content we retrieved from the database of Twitter posts.

The visual interface offers several parameters to adapt the visual appearance of the tag cloud to the user’s needs (see Fig. 3a). The users can, for instance, switch between linear and logarithmic scaling of the terms’ font size, depending on the distribution of the terms in the analyzed microblog content. Furthermore, they can adapt the overall size of the tag cloud by specifying the maximum number of terms that are shown, or the minimum number of times a term must occur in the contents to be displayed in the cloud. Finally, they can change the font and background colors, the horizontal and vertical spacing, and the minimum and maximum font sizes of the terms.

3.3 Encoding Time in Tag Clouds

The histograms plotted in the background of the terms show the variation in term use over the analyzed time span. In the example case, they visualize the time span for which we retrieved the microblog contents, i.e. September 3 to 19, 2011, as shown at the bottom of the tag cloud.

The histogram visualization is similar to the idea of SparkClouds [9] introduced in Sec. 2. However, using a histogram with discrete bars instead of continuous sparklines to encode time has some advantages: On the one hand, it allows to color-code certain bars, what is needed to indicate time-dependent co-occurrences. On the other hand, it lets the user specify the aggregation interval, i.e. how much time is represented by each bar in the histogram.

The default aggregation unit is one hour in our implementation. In the example of Fig. 3a, an aggregation value of 10 was chosen, meaning that each bar in the histogram represents 10 hours of microblogging (except from the last bar which is the remainder). Since the aggregation method was set to ‘average’, the presented values are the average occurrences within the 10 hour time frames. As for the tag clouds, one can specify further parameters like the color of the histograms or if the histograms are visible at all.

3.4 Co-occurrence Exploration

Hovering over a term leads to a highlighting of all co-occurring terms in the tag cloud – a key feature of our visualization technique that we call *co-occurrence highlighting*. Drawing connections between co-occurring terms instead, like in the mentioned tag graph approaches [7], could quickly result in a large number of edges creating visual clutter.

We thus decided to use the indirect method of visual highlighting and indicate the degree of co-occurrence by color, i.e. the font color of the highlighted terms correlates with the co-occurrence frequency: The more often a term has been used together with the selected one, the stronger its color. For instance, the term ‘game’ has a black color in Fig. 3a, as it co-occurs only six times with the term ‘tennis’ in the microposts, while ‘open’ has the same red color as ‘tennis’, since the terms are most often used together (565 times in total). The number of co-occurrences is given in parentheses which can be deactivated if not wanted.

Clicking on a term reduces the set of terms displayed in the cloud to those that co-occur with the selected one (including the selected term itself). This allows to explore a certain subset of possibly related terms separated from the others. If there is no distraction by other terms, relevant relationships, events, and trends may be found more easily in the contents.

3.5 Additional Map View

As the microposts' geo-information is not included in the tag cloud visualization, we additionally realized a map view in our visual interface. It is not part of the time-varying co-occurrence highlighting but can fruitfully be used with it. If certain terms in the tag cloud attract the users' attention, they can switch to the map view to see from which region the related microposts have been sent. This allows, for instance, to identify the location of an event that is mentioned in several microposts. Though a one-to-one mapping between the origin of microposts and the location of an event is not possible in many cases, the map view may provide geographic information of interest to the analysis.

Fig. 3b shows the geo-information for the 4,934 microposts we retrieved from the database (see Sec. 3.1). The dots are colored according to the five events detected with the Spatiotemporal Anomaly Detector. Each event has its own color, grouping all microposts that are detected to be about the same event.

4. DISCUSSION AND FUTURE WORK

In this paper, we exploited the idea of tag clouds to visually analyze microblog content. In particular, we presented the novel visualization technique of time-varying co-occurrence highlighting. It supports the visual analysis of highly dynamic and interrelated microblog content by enhancing tag clouds with information about time-dependent term variations and co-occurrences. We showed that the integration of these two key dimensions of microblogging in one visualization can be of value for the interactive exploration of the contents.

A possible drawback of the approach is the increased complexity of the tag cloud visualization. Though we still regard the enhanced tag clouds as relatively easy to understand and use, they are not as simple and puristic as common tag cloud visualizations. Generally, our enhanced tag clouds face the same problem as common tag clouds in that they aggregate terms with identical spellings that may not have identical meanings (e.g. think of the term 'apple' that may denote the fruit or the company, among others). However, in contrast to common tag clouds, the co-occurrence highlighting may help to identify disambiguations (e.g. if both 'fruit' and 'computer' is highlighted along with 'apple'), though it cannot resolve them. These and other issues of the enhanced tag cloud visualization are best investigated in a user study.

Related to this, we plan to more deeply integrate the visual interface with the database of Twitter posts presented in Sec. 3.1. We also plan to apply the visualization technique to contents from domains other than microblogging. Finally, it would be interesting to examine further enhancements to tag clouds and investigate how they support the visual analysis of social media contents.

5. ACKNOWLEDGMENTS

We thank Harald Bosch who helped in retrieving the microblog contents with the Spatiotemporal Anomaly Detector. This work was partially funded by the German Federal Ministry of Education and Research (BMBF) in the context of the VASA project.

6. REFERENCES

[1] Real-time local Twitter trends – Trendsmap. http://trendsmap.com.

[2] What is Twitter? Twitter for Business. http://business.twitter.com/basics/what-is-twitter/.

[3] Wordle – Beautiful Word Clouds. http://www.wordle.net.

[4] J. Clark. Twitter StreamGraphs. http://www.neoformix.com/Projects/TwitterStreamGraphs/view.php.

[5] C. Collins, F. B. Viégas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *Proc. of the IEEE Symposium on Visual Analytics Science and Technology*, VAST '09, pages 91–98. IEEE, 2009.

[6] M. Dörk, D. M. Gruen, C. Williamson, and M. S. T. Carpendale. A visual backchannel for large-scale events. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1129–1138, 2010.

[7] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proc. of the 16th International Conference on World Wide Web*, WWW '09, pages 211–220. ACM, 2007.

[8] A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating summaries and visualization for large collections of geo-referenced photographs. In *Proc. of the 8th ACM International Workshop on Multimedia Information Retrieval*, MIR '06, pages 89–98. ACM, 2006.

[9] B. Lee, N. H. Riche, A. K. Karlson, and S. Carpendale. SparkClouds: Visualizing trends in tag clouds. *IEEE Transactions on Visualization and Computer Graphics*, 16:1182–1189, 2010.

[10] S. Lohmann and P. Díaz. Representing and visualizing folksonomies as graphs – a reference model. In *Proc. of the 11th International Working Conference on Advanced Visual Interfaces*, AVI '12. ACM, to appear.

[11] S. Lohmann, J. Ziegler, and L. Tetzlaff. Comparison of tag cloud layouts: Task-related performance and visual exploration. In *Proc. of the 12th International Conference on Human-Computer Interaction*, INTERACT '09, pages 392–404. Springer, 2009.

[12] A. M. MacEachren, A. R. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. Senseplace2: Geotwitter analytics support for situational awareness. In *Proc. of the IEEE Symposium on Visual Analytics Science and Technology*, VAST '11, pages 181–190. IEEE, 2011.

[13] M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: can we trust what we rt? In *Proc. of the 1st Workshop on Social Media Analytics*, SOMA '10, pages 71–79. ACM, 2010.

[14] A. Pozdnoukhov and C. Kaiser. Space-time dynamics of topics in streaming text. In *Proc. of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '11, pages 8:1–8:8. ACM, 2011.

[15] D. Thom, H. Bosch, S. Koch, M. Wörner, and T. Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In *Proc. of the 2012 IEEE Pacific Visualization Symposium*, PacificVis '12, pages 41–48. IEEE, 2012.

[16] F. B. Viégas and M. Wattenberg. Tag clouds and the case for vernacular visualization. *interactions*, 15:49–52, 2008.