TopicWave: Interactive Visual Analytics of Spatiotemporal Topics Distribution of People's Reactions to Events from Geo-tagged Social Media

Siming Chen
Bonn University/
Fraunhofer IAIS
Bonn, Germany
Siming.Chen@iais.fraunhofer.de

Jie Li
Tianjin University
Tianjin, China
Jie.li@tju.edu.cn

Gennady Andrienko
Fraunhofer IAIS/
City University, London
Sankt Augustin, Germany
Gennady.Andrienko@iais.fraunhofer.de

Nathalia Andrienko
Fraunhofer IAIS/
City University, London
Sankt Augustin, Germany
Nathalia.Andrienko@iais.fraunhofer.de

Abstract

People often post messages in social media to express their opinions and ideas concerning various events happening in the world. It is an interesting but challenging task to understand how the opinions vary across people, locations, and times. We propose ***TopicWave***, an approach combining topic modelling with interactive visual techniques, for exploring the spatial-temporal variation of discussion topics derived from people's reactions to significant events expressed in geotagged social media. To create suitable input data for topic modelling, we aggregate social media messages coming from each user and from each location by time intervals. From topic overview to details, we visualize the evolution of discussion topics, which are represented by significant keywords, for groups of users or locations using a river metaphor. Interactive tools allow the analyst to explore how the popularity of each topic and its semantics (i.e., the representative keywords) vary over the sets of people and locations and evolve over time. We demonstrate the use of the proposed approach by example of analysing geotagged Twitter messages related to the Brexit event.

Keywords: Social Media, Visual Analytics, Topic Analytics, Spatial-temporal Visualization.

Introduction

Messages in social media, such as Twitter and Facebook, are often posted in reaction to some events happening in the world. **People** produce text messages with **semantic** contents (meanings) at different moments in **time** being at specific **locations**. Understanding the variety of people's reactions to significant events is important in many areas, such as journalism, social science, business, and politics. Our goal is to propose a way to analyse the relationships of the message contents to the space, time, and people. It is challenging due to the large volume and multifaceted structure of the data.

Visual analytics is a science aiming at supporting human analytical reasoning by interactive visualisations combined with capabilities of computational processing (Thomas and Cook, 2005). To enable the spatial, temporal, and semantic analysis of event-prompted reactions of social media users, we propose a visual analytics system ***TopicWave***, which combines computational techniques for topic modelling with interactive visual tools enabling the exploration of the variation of the derived topics across the space, time, and set of users.

Recent state of the art in visual analytics regarding analysis of social media data is summarized in (Chen et al. 2017a). Andrienko et al. analyzed social media from spatial-temporal and topic perspectives to investigate users' mobility (Andrienko et al, 2013). Chen et al. proposed a map-based visual metaphor to visualize the event evolution (Chen et al., 2017b). Unlike the previous works, we focus on the evolution of people's reactions to events, and we consider the spatial, temporal, and semantic aspects of this evolution.

1 Data Description and Method Overview

The most important fields in geotagged social media posts are the user identifier (Who), timestamp (When), location (Where), and text, including hashtags (What). The task we aim to support is analyzing how the **semantics** of the texts varies over the set of **people**, **space**, and **time**. Our approach includes three main steps: data processing, topic overview, and reaction exploration (Fig. 1).

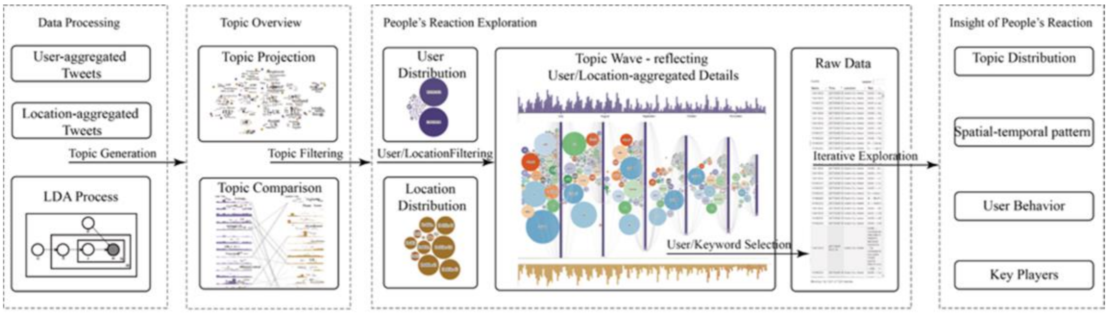
2 Visual Analytics Procedure for Analyzing the Reactions to the 'Brexit' Event in 2017

We use data from Twitter as an example of the social media to illustrate our approach. We crawled 380,000 geotagged messages containing "brexit" in either the message texts or hashtags and posted in UK and Ireland in 2017 to analyze people's reactions to the Brexit event.

2.1 Data Processing: Aggregation and Topic Modeling

Probabilistic topic modelling, such as LDA (Blei et al. 2003), is a class of statistical techniques that discover latent themes from a set of texts. The resulting topics are represented as combinations of significant keywords having high probabilities to co-occur in texts. Input texts for topic

Figure 1: Visual analytics pipeline for people's reactions analysis. TopicWave supports interactive exploration from an overview to details on demand with filtering according to topic, user, location, and keyword.

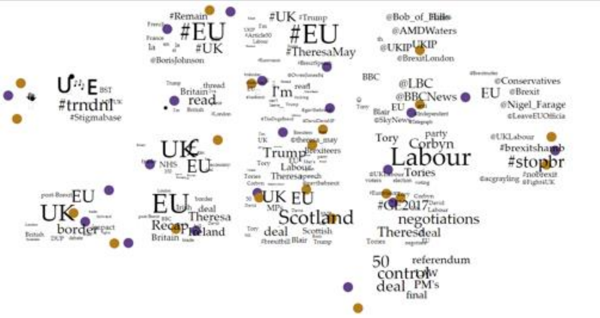


modelling methods need to be sufficiently long. Texts of social media messages, which are usually very short, are not directly suitable. We therefore generate input texts for topic modelling using two types of aggregation: user + time, i.e., by user and time interval, and location + time, i.e., by location and time interval. We generate 20 topics from each set of aggregated texts by means of LDA. For each aggregated text, we obtain the probability of each topic. The extracted topics represent the semantics of the discussions conducted by the users and at the locations.

2.2 Topic Overview and Selection

To support an overall understanding of the extracted topics, we propose two visualizations: topic projection (Fig. 2) and topic comparison (omitted here due to the page limit). For topic projection, we apply the t-SNE projection method (van der Maaten, and Hinton., 2008) to the vectors of keyword probabilities representing the topics. Purple dots in Fig. 2 represent the user-aggregated topics and yellow dots represent the location-aggregated topics. The distances in the projection reflect semantic similarities between the topics.

Figure 2: Topic projection. It shows the most important keywords of each topic and represents semantic similarity between topics through neighbourhood in the projection.



2.3 People's Reaction Exploration

After the analyst selects one or more topics of interest from the overview, the displays show details for these topics. In particular, a temporal display (Fig. 3) shows how the posting of tweets related to the selected topics evolve over time. There are three main parts: user-tweet histogram (top), location-tweet histogram (bottom), and keyword flow (middle), sharing a common horizontal axis representing time. In the histograms, each bar corresponds to one day, and the bar lengths represent the tweet counts. In the keyword flow view, tweets are aggregated by longer time intervals, such as one month, which gives more display space for showing the keywords that occurred frequently in these intervals. Each circle corresponds to one keyword, and the size is proportional to the frequency. On the right of each keyword group, there is a bar representing the distribution of the tweets in the respective time interval over the users (in purple) or locations (in yellow, Fig. 4), depending on the analyst's choice. The bar is divided into segments proportionally to the amounts of tweets from individual users or locations. There are light curved lines connecting bar segments to keywords to show the main themes of the users or locations. Analysts can observe temporal trends regarding the keywords in the discussions and topic-related tweeting activities of the users and locations. For example, in Fig. 4, we see that there was a burst in March, mostly in Scotland region. A new keyword "*ScotRef*" stood out reflecting the independency desire of Scotland people.

Figure 3: Visual design of TopicWave showing the keywords evolution among different users/locations along the time, reflecting the dynamics of people's reactions.

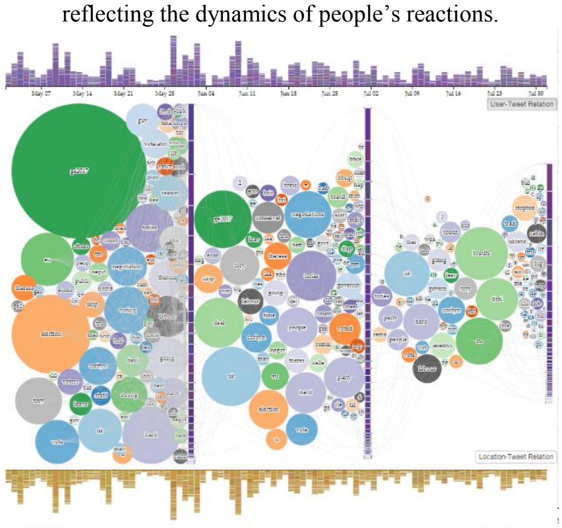
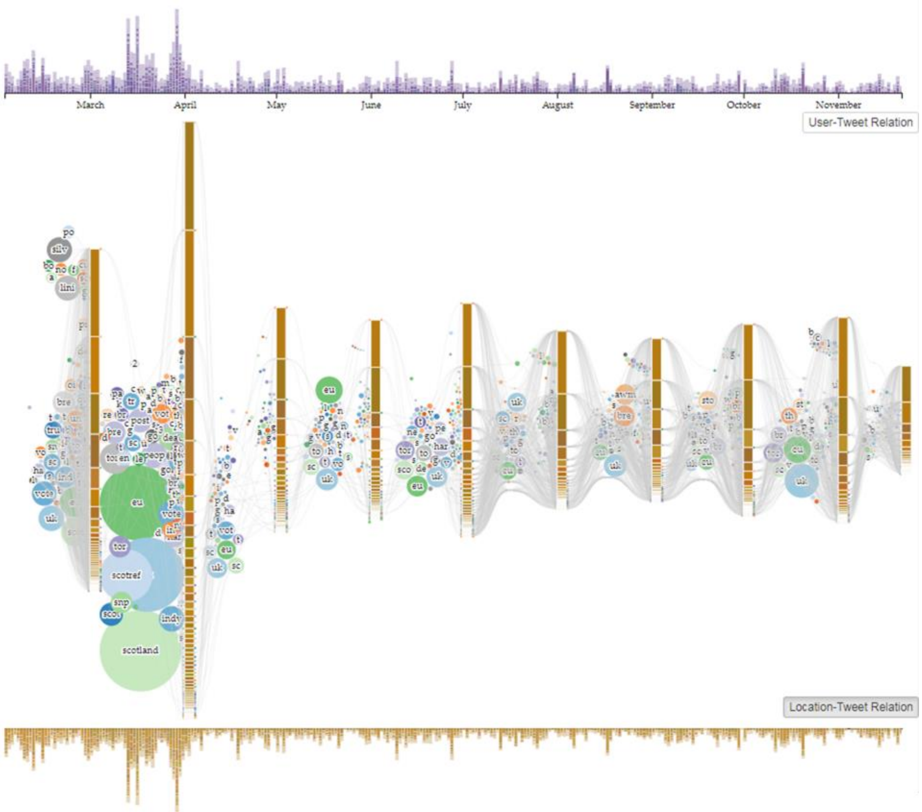


Figure 4: Finding a burst of occurrences of ‘ScotRef’ in the beginning of 2017 in the Scotland region.



Analysts can interactively select not only topics but also subsets of users and/or locations. For example, exploring the topics discussing “*StopBrexit*”, we found a burst starting in the middle of the year (Fig. 5). A plausible guess is that these discussions were triggered by a series of fruitless negotiations between the UK and EU. In the user distribution view, we find that there were several dominating users. We click on one of these users to investigate his/her reactions (Fig. 6) and find an important keyword “*#Standup4FirmRemain*” appearing among the 289 tweets of this user throughout the year. By drilling down to the raw data, we find that this user was actively arguing with others concerning ‘*StopBrexit*’.

3 Conclusion

We present a visual analytics method for analysing how the semantics of people’s reactions varies over space and time. We propose a workflow including a new interactive visualization incorporating different facets of the data: people, semantics (represented by topics and keywords), spatial locations, and time. Interactions enable exploration from gaining an overview to seeing details. We are aware that geotagged social media data may contain fake and wrong geographic positions of the tweets, which may not reflect the actual positions of the users. However, detecting and correcting possible errors in data is out of the scope of our research. We assume that errors occur relatively rarely. A small fraction of erroneous data cannot significantly affect the general spatio-temporal patterns that can be observed.

References

Blei, D. M.; Ng, A. Y. and Jordan, M. I. (2003), 'Latent dirichlet allocation', *J. Mach. Learn*. Res. 3 , 993--1022.

Chen, S.; Chen, S.; Lin, L.; Yuan, X; Liang, J. and Zhang, X. (2017a). E-Map: A visual analytics approach for exploring significant event evolutions in social media. In: *Proceedings. of IEEE Visual Analytics Science and Technology*, Pheonix, 2017.

Chen, S.; Lin, L; and Yuan, X. (2017b). Social Media Visual Analytics. *Computer Graphics Forum*, 36(3), pp.563-587.

van der Maaten, L. and Hinton, G. (2008), Visualizing Data using t-SNE , *Journal of Machine Learning Research* 9 , 2579--2605.

Andrienko, G. L.; Andrienko, N. V.; Bosch, H.; Ertl, T.; Fuchs, G.; Jankowski, P. and Thom, D. (2013), Thematic Patterns in Georeferenced Tweets through Space-Time Visual Analytics, *Computing in Science and Engineering* 15 (3), 72--82.

Thomas, J. J. and Cook, K. A. (2005), *Illuminating the Path: The Research and Development Agenda for Visual Analytics* , National Visualization and Analytics Ctr .

Figure 5: Visualisation of the “StopBrexit” discussion reveals several dominating users and an increase of the occurrences of “*stopBrexit*” at the end of the year.

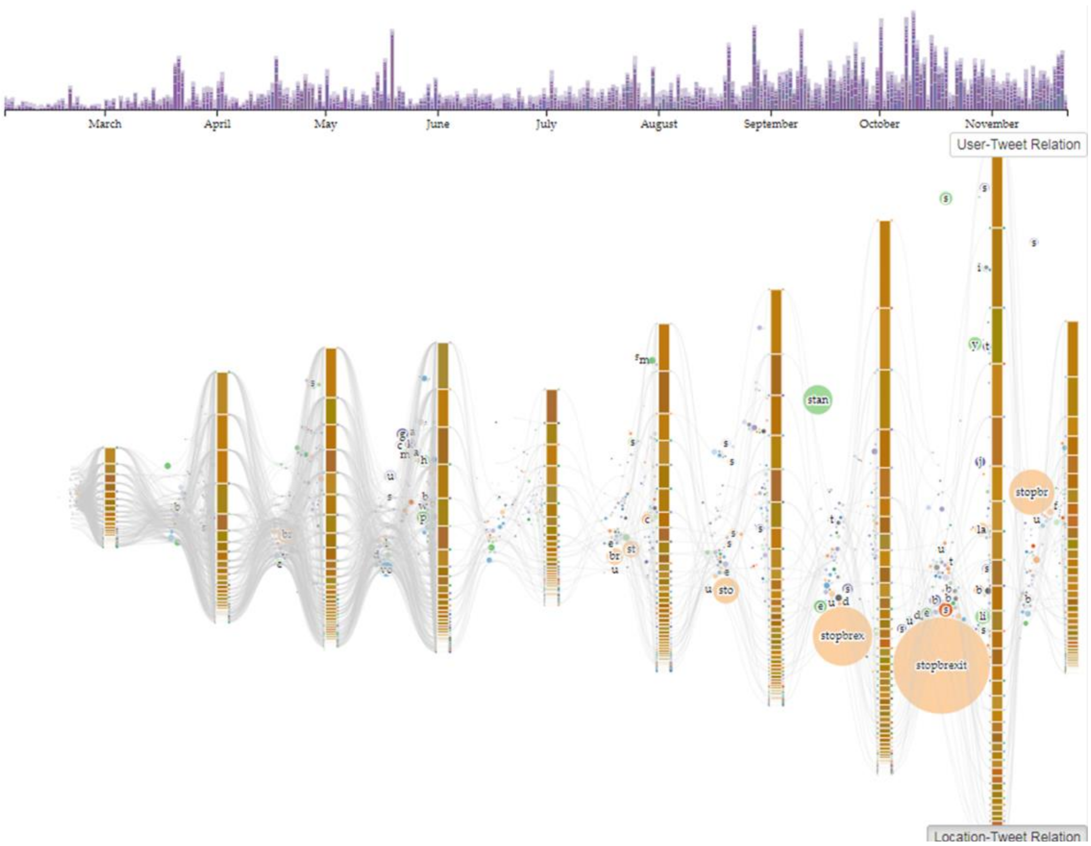


Figure 6: An interactively selected subset of tweets coming from one of the most active users who tweeted on the "*StopBrexit*" topic contains many occurrences of the keyword "*#Standup4FirmRemain*".