EuroVis Workshop on Visual Analytics (2018)
C. Tominski and T. von Landesberger (Editors)

Visual Exploration of Spatial and Temporal Variations of Tweet Topic Popularity

Jie Li [1†] and Siming Chen[2‡] and Gennady Andrienko[2,3] and Natalia Andrienko[2,3]

[1]School of Computer Software, Tianjin University, China
[2]Fraunhofer Institute IAIS, Germany
[3]City, University of London, UK, Austria

Abstract

We present a visual analytical approach to exploring variation of topic popularity in social media (such as Twitter) over space and time. Our approach includes an analytical pipeline and a multi-view visualization tool. As attempts of topic extraction from very short texts like tweets may not produce meaningful results, we aggregate the texts prior to applying topic modelling techniques. Interactive visualisations support detection of burst events in social media posting activities at different locations, show the spatial, temporal, quantitative, and semantic aspects of these events, and enable the user to explore how popularity of topics varies over cities and time. A case study has been conducted using a real-world tweet dataset.

1. Introduction

Topic mining is a common task in social media analysis. Extracting meaningful topics and understanding the variation of their overall popularity as well as the popularity at different locations may be useful for observing processes going on in the society. The potential application value has attracted a considerable amount of research, most of which focuses on analysing the temporal variation of topics [XWW*13, SWL*14, WLY*14, CLWW14] or topic spreading among different user communities [WLC*16].

In this paper, we propose a visual analytics approach to analysing topic popularity variation across different locations and time periods. Topic popularity is a quantitative measure showing what ideas or events gain high attention of public and become actively commented or discussed. As new ideas and events constantly emerge, on the one hand, and public interest to them as well as people's opinions may greatly differ from place to place, on the other hand, topic popularity is a spatio-temporal phenomenon, which needs to be studied using appropriate methods enabling spatio-temporal analysis. Our approach integrates an analytical pipeline and a visualization tool for interactive exploration of topic popularity variation across space and time.

2. Related Work

Existing approaches to analysing collections and streams of geographically referenced social media messages mostly focus on two out of the following three aspects: spatial, temporal, and semantic (i.e., themes and meanings of the messages). Thus, multiple works have been done on analysing spatio-temporal patterns of social media activity [SOM10, CYW*16, BTH*13, MBB*11]. Sakaki et al. [SOM10] constructed an earthquake reporting system by identifying abnormal tweeting activity. Kraft et al. [KWD*13] created a near real-time analysis system for detection of emerging events. Andrienko et al. [FAA*13] extracted tweet burst related to floods in Germany. Furthermore, Andrienko et al. [AAF*15] clustered messages related to a real-world event in space and time and visualized the evolution of the clusters to trace the progress of the event and people's reaction to it.

Another group of related works focuses on the temporal and semantic aspects. Sun et al. [SWL*14, HHKE16] analysed relationships between topics appearing in social media over time. Xu et al. [XWW*13] characterized topic competition in attracting public attentions. Wu et al. [WLY*14] modelled the opinion diffusion phenomena, which represents the temporal variations of topics. Cui et al. [CLWW14] visualized hierarchical and evolving topics in text corpora. Wang et al. [WLC*16] analysed topic transitions between different user communities. Most recent works in visual analytics for social media can be found in [CLY17].

In our work, we address the three aspects, i.e., time, space, and semantics (topics), together. Unlike existing approaches mainly utilizing statistics-based methods [PK11, CTB*13], we provide a multi-view visual analytic approach to exploring topic popularity variation along space and time, enabling interactively drilling down to discover various semantic patterns. Techniques for supporting such kind of analysis have not been proposed before.

† National NSFC project (Grant numbers: 61602340 and 61572348)
‡ DFG Priority Research Program SPP 1894 on VGI, Germany

DOI: 10.2312/eurova.20181105

3. Analytical Pipeline

Our goal is to support interactive visual analysis of the the spatio-temporal dynamics of topic popularity. Our target users are mainly professional analysts, such as news/communication researchers, sociologists, politics researchers, etc.. We have developed an analytical pipeline shown schematically in Fig. 1.

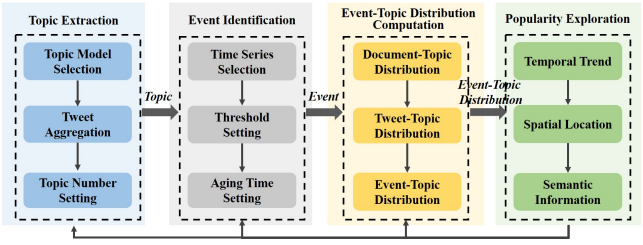

Figure 1: *Analytical pipeline*

- **Topic Extraction**. We utilize LDA (Latent Dirichlet Allocation) [BNJ03] to extract topics from the message texts. As the original messages are too short for reliable topic extraction, we aggregate them by hashtag and time. (Section 4.1).
- **Burst Event Identification**. We extract burst events from time series of tweet counts in different cities. This operation involves several parameters, and the user can interactively test how the settings affect the result. (Section 4.2)
- **Event-Topic Distribution Computation**. Each topic is characterized by a probability distribution over all input documents, i.e., aggregated messages. From these distributions, we derive the topic probabilities for the burst events, which involve subsets of the input documents. Furthermore, based on the number of the original messages associated with each topic, we compute topic popularity scores in each burst event.
- **Popularity Exploration** Analysts can apply various visual components to interactively explore the popularity variation of topics over time and space (i.e., the set of cities).

4. Topic Extraction and Association with Bursts

4.1. Aggregation-based Topic Extraction

LDA and other topic modelling techniques are based on word co-occurrences in texts; therefore, they do not work well with very short text documents, such as tweets, which contain too few words. To overcome this problem, we aggregate the original message texts in longer documents. Generally, messages can be aggregated based on time, user, location, and/or hashtag. For meaningful topic modelling, messages that are put together should have similar semantic contents. We assume that messages with the same hashtag are likely to be semantically related. However, semantic relatedness may be higher in messages posted at close times than in temporally distant messages. Based on these considerations, we chose to aggregate the messages based on the common hashtags and time intervals. The aggregation is illustrated in Fig. 2. Tweets with multiple hashtags are included in multiple documents.

The length of the time interval Δt is generally chosen depending on the expected rate of change. For convenience, usual time units are preferred, such as monthly, weekly, daily, or hourly intervals. It should be taken into account that choosing a longer time interval, such as a week or a month rather than one day or one hour, allows using more tweets for derivation of the LDA model, thus avoiding occasional biases due to sparse text inputs. In this paper, we choose a weekly time interval to balance the quality of the extracted topics and computational efficiency. This choice also determines the temporal scale of the further analysis.

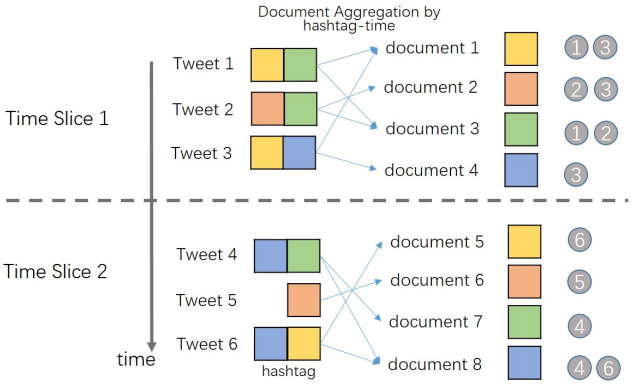

Figure 2: *Illustration of hashtag- and time-based aggregation. A tweet with multiple hashtags is included into multiple documents.*

4.2. Burst Event Identification

A burst event, which is manifested as an abrupt increase of the number of posted messages at some location (city), can be identified by setting a variation threshold value v. Any change of the message number where the change rate is higher than v can be treated as a burst event. The corresponding messages form the body (content) of the burst event. In real life, a burst event can last over some period of time, and the number of messages always fluctuates during this period. It is desirable that such real event is represented by a single computationally extracted event rather than by several events that are disjoint due to drops in the message number. To avoid splitting a burst event, we apply an ageing time threshold: when time intervals between bursts are shorter than the ageing time threshold, they are merged into one event.

4.3. Computation of Event-Topic Distribution

The output of the LDA model contains a document-topic distribution, where each element is a vector of length N (N is the number of topics) representing the probabilities of a document to belong to each topic. Since each document is a combination of multiple messages, we derive from these data the message-topic distribution, which is done as follows. Let d be a document, d_i represent the probability of d to belong to the ith topic, m is a message, and D is the set of documents containing m. The probability of m to belong to the ith topic can be obtained through:

$$m_i = \frac{\sum_{d \sqsubseteq D} d_i}{\|D\|} \quad (1)$$

Using Fig. 2 as an example, $tweet1$ and $tweet3$ only belong to one document (as they have one hashtag); hence, their

message-topic distributions equal the document-topic distribution of *document*1. *Tweet*4 belongs to two documents; therefore, its probabilities on each topic is the average number of those of *document*3 and *document*4.

Having obtained the message-topic distribution, we obtain the probability of an event e to be related to the ith topic through:

$$e_i = \frac{\sum_{m \sqsubseteq e} m_i}{\|e\|} \quad (2)$$

The value e_i reflects the popularity of the ith topic in the event e.

5. Visual Design

As shown in Fig. 3, the user interface includes five components. One of them is the event extractor (Fig. 3a), since only topics occurring in burst events are in the target of the analysis. All extracted topics are vertically aligned in the topic list (Fig. 3d). The topic popularity view (Fig. 3b) and the spatial view (Fig. 3c) show the semantic, spatial, and temporal information for the extracted events. The event content view (Fig. 3e) shows keywords for a selected event. To accommodate a larger number of words, we use a word cloud layout rather than an ordered list. The exact number of occurrences for any word is shown on demand in a pop-up window.

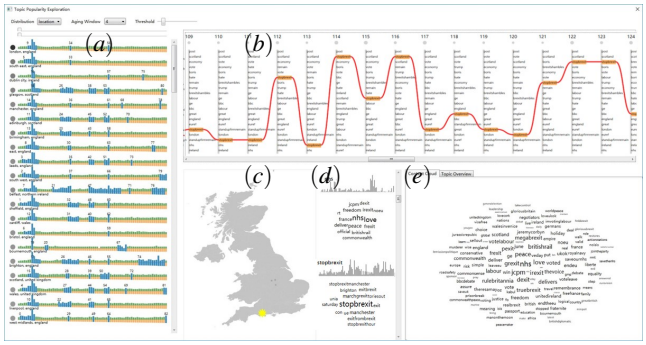

Figure 3: *Visual interface of our approach consists of 5 interrelated components: (a) event extractor, (b) topic popularity view, (c) spatial view, (d) topic list, (e) event content view.*

Event extractor (Fig. 3a) is used to identify burst events in time series of message counts at different locations, which are previously normalized to reduce the impacts of great tweet number differences among the locations. The time series are represented by histograms, which show two types of distributions: the accumulation distribution (yellow bars) and the absolute number distribution (green bars). In the accumulation distribution, each value for a time step t is the sum of all absolute values from t and the preceding time steps. This is used to assess the variation rate of the message number in a way minimising the impact of small fluctuations.

The analyst interactively sets the variation rate and the ageing time thresholds, as introduced in section 4.2. The system finds the corresponding parts of the time series and highlights them in the bar charts (see the blue bars in Fig. 3a). The analyst can interactively explore how changes of the settings affect the result before making the final decision for the event extraction. Each extracted event gets a numeric label that indicates the temporal order of its occurrence based on the time stamp of the earliest message of this event.

The extracted events are represented in the topic popularity view (Fig. 3b) as a sequence of columns composed of vertically aligned words; each column corresponds to one event. The horizontal ordering corresponds to the temporal order of the events. Each word in a column represents a topic; it is the word with the highest TF-IDF score for this topic. The words are put from top to bottom in the order of decreasing popularity of the topics. The analyst can thus judge the relative popularity of the topics in each event and compare this between events. The analyst can also select a topic, which results in generating a trend line showing the popularity variation of this topic along the events. Since the events are associated with specific locations and times of occurrence, the trend line provides information on the spatio-temporal variation of the topic popularity. To see the spatial location of an event in the map view (Fig. 3c), the user clicks on the event's column in the topic popularity view (Fig. 3b). When a topic is selected, the topic list (Figure 3d) is scrolled to the section of this topic, where a word cloud shows the words with the highest probabilities and a bar chart shows the temporal variation of the aggregated message probabilities for the topic. The semantic content of the topic in the selected event is represented by a word cloud in the event content view (Fig. 3e).

6. Case Study

This section demonstrates the effectiveness of our approach by presenting several findings in a collection of tweets related to Brexit.

6.1. Data Preprocessing and Topic Extraction

The data were collected through the Twitter streaming API using a query with the bounding rectangle of Great Britain. We retrieved the tweets containing "brexit" in the message texts or hashtags. We filtered out the tweets with empty hashtag fields, as we use hashtags for text aggregation. The longitude and latitude fields of the tweets are typically empty, but most of them have location names in the location field. We utilized the Google Map API to obtain the coordinates for these location names. We collected about 340,000 tweets from the period March 1, 2016 - October 29, 2017, 78 weeks in total. There are about 6900 different hashtags; we created for each hashtag 78 documents (78 weeks) for topic extraction. To choose a suitable topic number, we conducted multiple LDA trials with different topic numbers and observed whether the extracted topics are interesting, well differentiable, and have sufficient support (number of tweets). Finally, we took 20 representative topics for further analysis.

6.2. Topic popularity variation during the referendum period

The referendum topic is represented by the word 'euref'. We observe that this topic was the most popular before the referendum; see the left part of Fig. 4a. A significant drop of its popularity begins from event 19, whereas the popularity of 'brexitshambles' begins to rise. Events 17 and 18 occurred in Dublin (Ireland) and

Glasgow (Scotland); see Fig. 4b). Here, the bursts of the messages on the topic of the referendum occurred latter than in the other cities. Events 19 and 20 both occurred in Bournemouth (England); see Fig. 4c). This indicates that the changes in the topic popularities after the referendum started from this city.

We also find several topics whose popularities did not greatly change during that period, such as topic 'remain'. By observing its trend line, we find that the popularity rank was between 3 and 7 before the referendum and had only a slight decrease after that, being in the top 7-10 for a long time. As we know, some people turned to support Britain remaining in the EU after the referendum. What we have observed is consistent with the actual development of people's reactions and opinions.

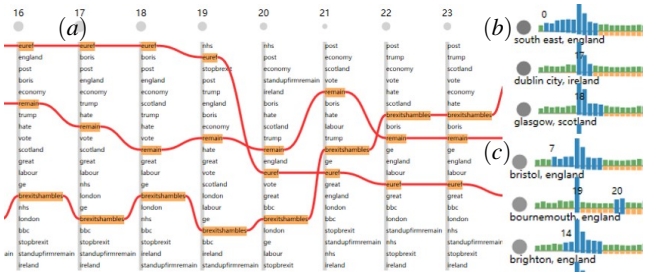

Figure 4: *Popularity variation of three topics during the Brexit referendum period. (a) Popularity trend lines of three topics. (b-c) Cities and time lines of events 17-20, when the popularities changed significantly.*

6.3. When and where Stopbrexit became popular?

We look (Fig. 3b) at the popularity variation of the topic 'stopbrexit', which corresponds to the negative attitude toward brexit. By observing the trend line, we find that 'stopbrexit' was at the bottom of the popularity lists of the events following the referendum. However, along with the development of the Brexit process, its popularity began to rise, and it even reached the top 3 in several events, as can be seen in Fig. 3b. We trace the trend line and select the first great rise of the topic popularity. In the spatial view, we see that this happened in Bournemouth (Fig. 3c) in September 2017. We also find several separated abrupt popularity rises before this time, which also occurred in that city. To verify our findings, we retrieved the related news through Google News, from which we learned that the Liberal Democratic Party Conference on the topic of stopping Brexit was held in Bournemouth during that period. We also found that the headquarters of the company JP Morgan's is in Bournemouth. Their CEO would prefer Britain to stay in EU, since Brexit may seriously affects their business in EU.

6.4. Spatio-temporal and semantic information variation of the topic 'Scotland'

In this example, we examine the spatial, temporal and semantic aspects of the popularity variation of the topic 'Scotland'. By observing the trend line, we find that this topic became popular during the referendum period. We select several events in which the topic had a high popularity and look where these events occurred. Most of them occurred in two big cities of Scotland, Edinburgh and Glasgow; these occurrences are marked with red circles in Fig. 5a. One event occurred in Wales; it is marked by a blue circle in Fig. 5a. The locations of the cities are shown in Fig. 5d. Analysing the semantic information of the selected events, we find "voted" in the word cloud of the event occurring in Wales (Fig. 5b). As we know, although a majority of people supported Brexit, there were several regions, including Scotland and Wales, where more people held an opposite attitude. The term "voted" indicates that they are forced to accept the referendum result. In the word cloud of Scotland (Fig. 5c), the most significant term is "indyref", which is an abbreviation of Scottish Independence Referendum. This represents the Scotland's hope to conduct an independence referendum to become an independent country and stay in the EU.

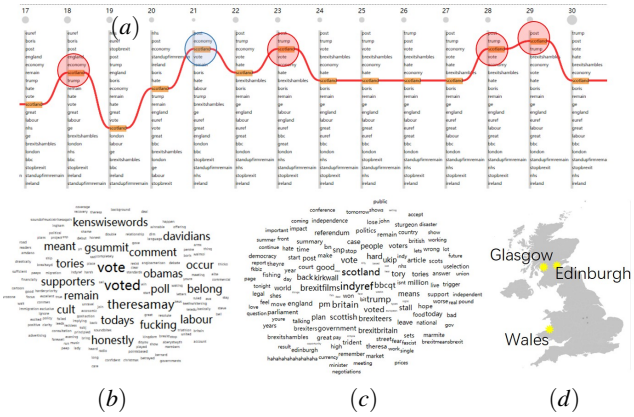

Figure 5: *Popularity variation of the topic 'Scotland'. (a) Topic popularity trend. (b-c) Semantic information of the events occurring in different cities. (d) Geolocation of the cities.*

7. Discussion and Conclusion

We have presented an analytical pipeline and a suite of visual and interactive tools for the exploration of the spatio-temporal variation of topic popularity associated with burst events in social media posting activities. This includes a feasible approach to topic extraction, which tackles the problem of unsuitability of very short texts for topic modelling through hashtags- and time-based text aggregation. We have demonstrated the effectiveness of the proposed analytical tools by examples of exploration of a real-world dataset.

There are several limitations in our current prototype, which will be addressed in the future work. Thus, more efficient support is desirable for determining the suitable number of topics when using the LDA algorithm. For analysing the sensitivity of the event extraction results to the parameters, an overview of the parameter impacts should be provided. A more scalable visual design is needed for dealing with a larger number of topics. We also want to explore using a sliding time window instead of fixed time bins in event extraction. There is also a risk of misinterpreting the topic lists in the topic popularity view. The lists represent the topic ranks rather than frequencies of their occurrence; the latter can be seen on hovering on the topics. We consider enabling explicit choice of the rank or frequency representation, which may reduce the risk of confusion.

References

[AAF*15] ANDRIENKO N., ANDRIENKO G., FUCHS G., RINZIVILLO S., BETZ H.-D.: Detection, tracking, and visualization of spatial event clusters for real time monitoring. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on* (2015), IEEE, pp. 1–10. 1

[BNJ03] BLEI D. M., NG A. Y., JORDAN M. I.: Latent dirichlet allocation. *Journal of machine Learning research 3*, Jan (2003), 993–1022. 2

[BTH*13] BOSCH H., THOM D., HEIMERL F., PÜTTMANN E., KOCH S., KRÜGER R., WÖRNER M., ERTL T.: Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2022–2031. 1

[CLWW14] CUI W., LIU S., WU Z., WEI H.: How hierarchical topics evolve in large text corpora. *IEEE transactions on visualization and computer graphics 20*, 12 (2014), 2281–2290. 1

[CLY17] CHEN S., LIN L., YUAN X.: Social Media Visual Analytics. *Computer Graphics Forum* (2017). doi:org:443/handle/10.1111/cgf13211. 1

[CTB*13] CHAE J., THOM D., BOSCH H., YUN J., MACIEJEWSKI R., EBERT D. S., ERTL T.: Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. 143–152. 1

[CYW*16] CHEN S., YUAN X., WANG Z., GUO C., LIANG J., WANG Z., ZHANG X. L., ZHANG J.: Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 270–279. 1

[FAA*13] FUCHS G., ANDRIENKO N., ANDRIENKO G., BOTHE S., STANGE H.: Tracing the german centennial flood in the stream of tweets: first lessons learned. In *Proceedings of the second ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information* (2013), ACM, pp. 31–38. 1

[HHKE16] HEIMERL F., HAN Q., KOCH S., ERTL T.: Citerivers: Visual analytics of citation patterns. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 190–9. 1

[KWD*13] KRAFT T., WANG D. X., DELAWDER J., DOU W., LI Y., RIBARSKY W.: Less after-the-fact: Investigative visual analysis of events from streaming twitter. In *Large-Scale Data Analysis and Visualization (LDAV), 2013 IEEE Symposium on* (2013), IEEE, pp. 95–103. 1

[MBB*11] MARCUS A., BERNSTEIN M. S., BADAR O., KARGER D. R., MADDEN S., MILLER R. C.: Twitinfo: aggregating and visualizing microblogs for event exploration. 227–236. 1

[PK11] POZDNOUKHOV A., KAISER C.: Space-time dynamics of topics in streaming text. 1–8. 1

[SOM10] SAKAKI T., OKAZAKI M., MATSUO Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (2010), ACM, pp. 851–860. 1

[SWL*14] SUN G., WU Y., LIU S., PENG T.-Q., ZHU J. J., LIANG R.: Evoriver: Visual analysis of topic coopetition on social media. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1753–1762. 1

[WLC*16] WANG X., LIU S., CHEN Y., PENG T.-Q., SU J., YANG J., GUO B.: How ideas flow across multiple social groups. In *Visual Analytics Science and Technology (VAST), 2016 IEEE Conference on* (2016), IEEE, pp. 51–60. 1

[WLY*14] WU Y., LIU S., YAN K., LIU M., WU F.: Opinionflow: Visual analysis of opinion diffusion on social media. *IEEE transactions on visualization and computer graphics 20*, 12 (2014), 1763–1772. 1

[XWW*13] XU P., WU Y., WEI E., PENG T.-Q., LIU S., ZHU J. J., QU H.: Visual analysis of topic competition on social media. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2012–2021. 1