Tile-Based Spatio-Temporal Visual Analytics via Topic Modeling on Social Media

Minsuk Choi*, Jaeseong Yoo*, Ashley S. Beavers†, Scott Langevin^, Chris Bethune^,
Sean McIntyre^, Barry Drake†, Jaegul Choo*, Haesun Park‡
*Korea University, ^Uncharted Software Inc., †Georgia Tech Research Institute, ‡Georgia Tech

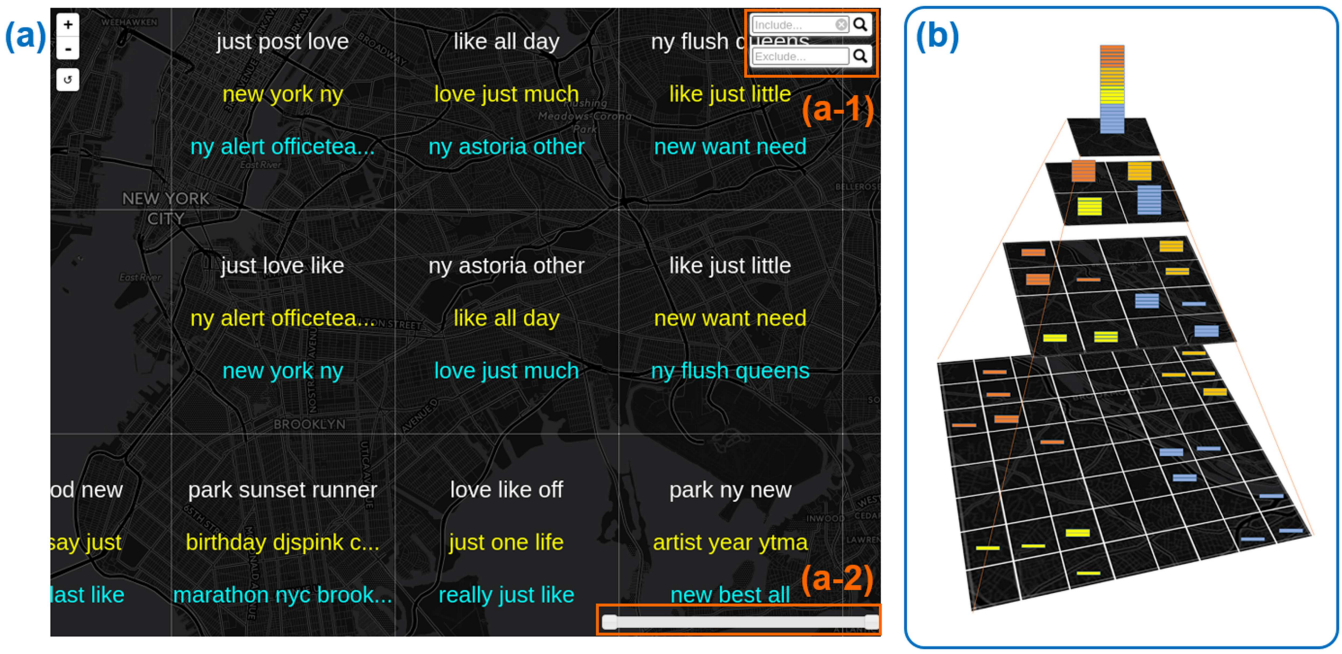

Figure 1: Our main user interface showing topic modeling results on a tile-based map view. In the tile-based map (a), the three topic clusters are shown, where each set of three keywords represents a single topic. Our system also supports keyword filtering that obtains the subset of documents including/excluding user-specified keywords (a-1) and the time frame filtering (a-2) on document data. In response to such a filtering interaction, our system re-computes the topic modeling results for the corresponding subset of documents per tile in parallel and updates its visualization as soon as the newly computed topic modeling results become available. As we zoom in, tiles are further split up at a fine-grained level (b), and the precomputed topic modeling results corresponding to these tiles are efficiently visualized accordingly.

Abstract

We present a visual analytics system that supports the geospatio-temporal analysis of social media data based on a large-scale distributed topic modeling technique. Through the analysis of social media data in a given time and region, we can identify critical events in real time. However, it takes significant time to perform such analyses against a large amount of social media data. As a way to handle this issue, we developed an efficient tile-based topic modeling approach, which divides textual data into multiple subsets with respect to different regions and time frames at different zoom levels and applies topic modeling to each subset.

Keywords: geospatial visualization, topic modeling, tile-based visualization, text analytics, social media.

*e-mail: mchoi@korea.ac.kr

1 Introduction

Geospatio-temporal information is crucial to extract meaningful information such as the important events occurring in a particular time and a region. Text data such as social media data are good sources to extract such information [1]. We can analyze and find interesting topics by analyzing their dominant keywords, which can serve as clues to, say, disease outbreak and other critical events.

We propose the visual analytics system that allows a user to analyze the social media data based on geospatio-temporal information in a map type of a view. In our system, a map is composed of multiple small geospatial tiles at different zoom levels, working in a similar manner to a google map. That is, we divide the entire social media data based on geospatial information into tiles and maintain them separately to analyze them on demand when a user aims to examine potentially interesting regions.

The topic modeling provides the capability of summarizing large-scale text data, and our contribution mainly lies in integrating it with tile-based geospatio-temporal visual analytics to analyze such large-scale data.

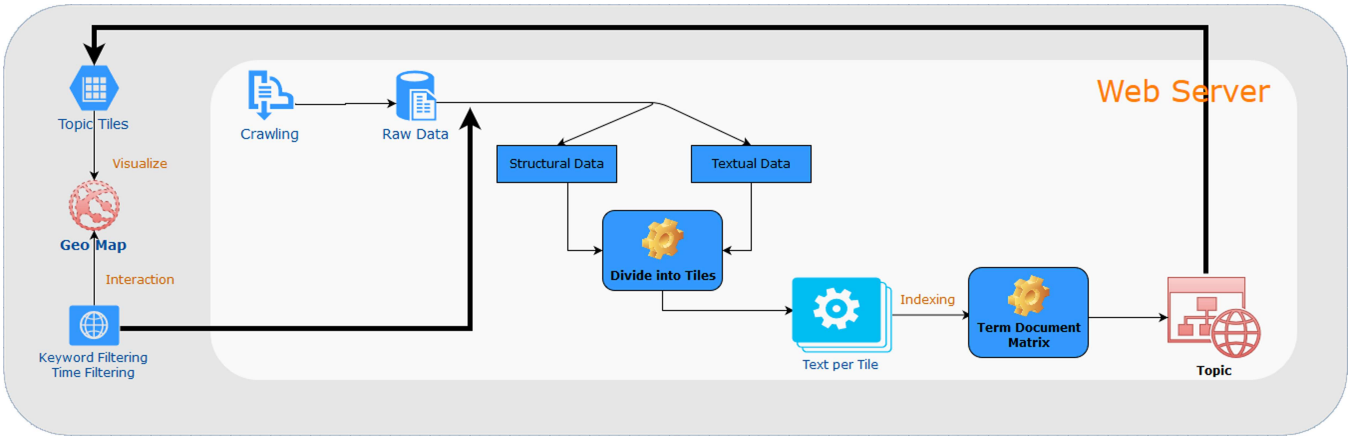

Figure 2: A workflow of our system. The crawled raw social media data are used for tile-based on-demand topic modeling. The most representative keywords of several topic clusters are transferred and visualized in the tile-based map view.

2 Geospatio-Temporal Visual Analytics via Tile-Based Topic Modeling

2.1 Overall Workflow

As shown in Fig. 1, the main idea of our system is to immediately deliver the topic modeling results for real-time interactive analysis. To achieve this, our system divides the document data into tile groups to reduce computing time. Furthermore, our web client requests the topic modeling output corresponding to a particular tile on demand. The web server sends the results once the topic modeling computation is done for each requested tile separately.

2.2 Tile-Based On-Demand Topic Modeling

Fig. 2 shows the overall workflow of our system. Initially, our system crawls the raw social media data and stores them in the storage. The raw data are composed of the textual data which include the contents and the structured data which include meta-data such as geospatio-temporal information. The textual data are used for the topic modeling while the structured data are used for filtering textual data to compute segmented tile groups. Afterwards, topic modeling based on distributed nonnegative matrix factorization [2] is applied separately to each of the tile group data. Once the topics are obtained, the most representative keywords of several topic clusters are visualized in the tile-based map view.

2.3 Visualization and Supported User Interactions

Users can determine the region of interest (ROI) by specifying a location in a web browser. The ROI determines the amount of data to be processed at a time. First, the user specifies the ROI depending on the currently shown region and the zoom level used in the web browser. A user can find the interesting topics in this area in a particular time using the slider interface.

The keyword filtering interaction allows users to focus on those text documents containing user-specified keywords while excluding another set of user-specified keywords. The server re-compute the topic modeling given such dynamically changing subsets of documents on the fly, generating a newly updated set of topic tiles.

3 Usage Scenarios

We apply our system to Twitter messages generated in New York City at the third of November, 2013 with "marathon" keyword included. Thus, our system extracts the topic modeling from such tweet messages. As shown in Fig. 3, we can see words such as "marathon", "run" and "runner" on the tiles of Brooklyn and Manhattan regions. Actually, the 2013 New York City Marathon was held around there at that time. Similar to this, we can extract other

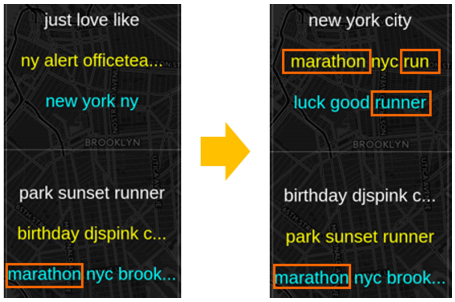

Figure 3: A usage scenario. For example, a user can include a "marathon" keyword to extract topic modeling, and related words are shown on other tiles after that.

interesting events such as hot issues about politics, entertainment, sports, disease and disaster, generated in a particular area and time using our analysis.

In this example shown in Fig. 1, the number of used text documents was 37,655. Each of the tiles has approximately 904 tweets on average and took 3,830 milliseconds for topic computation. Such an amount of computation time is still significant for supporting the real-time user interaction. Thus we will further improve this work in order to reduce the computation time.

4 Conclusion and Future Work

We presented a novel system that provides geospatio-temporal analysis of document data via topic modeling. Our system supports on-demand user interactions by showing topic modeling results focused on the region of interest. We plan to extend our system to support a distributed platform and enhance the efficiency and the scalability of topic modeling algorithm.

Acknowledgements

This work was supported in part by DARPA XDATA grant FA8750-12-2-0309. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] Daniel Cheng, Peter Schretlen and William Wright *Tile Based Visual Analytics for Twitter Big Data Exploratory Analysis.*, IEEE Big Data Conference, 2013.

[2] Jingu Kim and Haesun Park, *Fast nonnegative matrix factorization: An active-set-like method and comparisons*, SIAM Journal on Scientific Computing, 33(6), pp. 3261-3281, 2011.