μTOP: Spatio-Temporal Detection and Summarization of Locally Trending Topics in Microblog Posts

Paras Mehta
Freie Universität Berlin
Germany
paras.mehta@fu-berlin.de

Manuel Kotlarski
Freie Universität Berlin
Germany
kotlarski@inf.fu-berlin.de

Dimitrios Skoutas
IMIS, Athena R.C.
Greece
dskoutas@imis.athena-innovation.gr

Dimitris Sacharidis
Technische Universität Wien
Austria
dimitris@ec.tuwien.ac.at

Kostas Patroumpas
IMIS, Athena R.C.
Greece
kpatro@imis.athena-innovation.gr

Agnès Voisard
Freie Universität Berlin
Germany
agnes.voisard@fu-berlin.de

ABSTRACT

User-generated content in social media can offer valuable insights into local trends, events, and topics of interest. However, navigating through the vast amounts of posts either to retrieve certain pieces of information or to obtain an overview of the existing content, is often a challenging and overwhelming task. In this work, we present μTOP, a system for detecting and summarizing locally trending topics in microblog posts based on spatial, temporal and textual criteria. Using a sliding window model over an incoming stream of posts, μTOP detects locally trending topics, and associates each one with a spatio-temporal footprint. Then, for each spatial region and time period in which a certain topic is trending, the system generates a summary of the relevant posts, by selecting top-k posts based on the criteria of coverage and diversity. μTOP includes a Web-based user interface, providing a comprehensive way to visualize and explore the detected topics and their spatio-temporal summaries via a map and a timeline. The functionality of the system will be demonstrated using a continuously updated dataset containing more than 30 million geotagged tweets.

1. INTRODUCTION

Millions of posts are generated daily by users in social media, including text messages, photos, location check-ins, etc. These posts comprise textual content (typically, short text messages or tags), temporal information (the post's timestamp), and often spatial information (the post's geolocation). These spatial-temporal-textual objects are valuable pieces of information for revealing insights and trends regarding topics and events the users are interested in. However, given the sheer volume of this content, and its inherent redundancy and noise, retrieving relevant information or browsing and obtaining an overview of what is happening, is often a challenging and overwhelming task.

One solution to restrict the amount of incoming posts and focus on more relevant information is to filter out posts according to specified textual, spatial and/or temporal filters, as for example in publish/subscribe systems (e.g., [6]). However, given that social media content often involves new and emerging topics and events, the user may not know in advance what is interesting or relevant, and thus may not be able to specify a suitable geographic area, time period, or keywords for search.

To make it easier for users to get a quick grasp of the most important or interesting information, a common practice is to detect and present to the users a set of popular or trending topics (e.g., sets of hashtags in Twitter) that have high frequency (overall, or currently with respect to the past). However, the popularity of a topic is often not uniformly distributed across space and time; instead, a given topic may only be popular within specific geographic regions and over certain periods of time. In fact, recently there has been a lot of interest in finding local topics and events in Twitter (e.g., [1, 2, 3]). Nevertheless, even if a topic is detected as popular or trending, the posts belonging to it may still be in the order of hundreds or thousands. Hence, besides topic detection, generating topic summaries is also of high importance.

In this work, we present μTOP, a system for detecting and summarizing *locally trending topics* in streams of microblog posts. Each topic is represented by a set of one or more keywords (e.g., hashtags

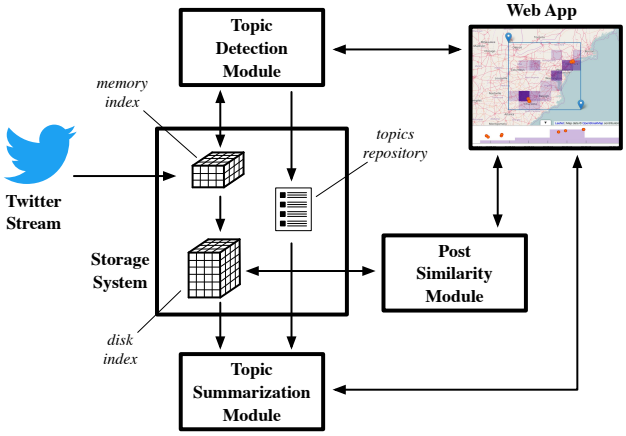

Figure 1: Architecture of μTOP.

in the case of Twitter), and is associated with a *spatio-temporal footprint*, i.e., a set of geographic regions and time periods over which this topic is identified to be popular. Thus, the spatio-temporal evolution of each detected topic is explicitly captured, and can be further explored. In fact, for each of these spatial regions and time intervals for which a topic is popular, μTOP can generate a summary of relevant tweets to describe the topic in more detail.

The discovery of locally trending topics is based on the approach presented in [5]. This method segments the space into a uniform grid and detects a set of trending topics in each cell by processing the incoming stream of posts applying a sliding window model. Thus, the topics are generated and monitored across space and time as new posts arrive and old ones expire, resulting in an evolving spatio-temporal footprint for each identified topic. Moreover, given a topic and its footprint, the system can generate a summary of relevant tweets. For this purpose, the relevant tweets are first retrieved using a spatial-temporal-textual filter, and then the top-k ones are selected according to the criteria of *coverage* and *diversity*, following the approach presented in [4].

Figure 1 presents an overview of the system architecture, which comprises the following main components. The *storage system*, detailed in Section 2, is responsible for ingesting the microblog posts (e.g., from Twitter's streaming API), and storing them in main memory and later on disk. In addition, this system maintains all topics and their spatio-temporal footprints. The core components of μTOP are the three data processing modules: *Topic Detection*, *Topic Summarization*, and *Post Similarity*, which are discussed in Section 3. Finally, the *Web App*, presented in Section 4, consists of the web-based user interface that allows users to issue queries, via invoking the appropriate modules, and visualize their results.

In the following sections, we describe in more detail the subsystems of μTOP, and present some usage examples in Section 5.

2. STORAGE SYSTEM

Each ingested post is represented as a spatial-temporal-textual object $D = \langle u, loc, t, \Psi \rangle$, where u is the identifier of the user making the post, $loc = (x, y)$ is the post's geolocation, t is the post's timestamp, and Ψ is a set of keywords representing the post's textual content.

To allow for efficient real-time detection of locally trending topics and the exploration (retrieval, summarization) of past topics and posts, we adopt a hybrid data indexing structure, involving both the main memory and the disk. This structure, depicted in Figure 2, indexes along all four attributes, latitude, longitude, time, and text. A 3-dimensional grid provides access along the first three attributes, while within each cell an inverted index provides efficient retrieval by keyword.

Each grid cell has size $g \times g \times \beta$, where g is a fixed arc range (for latitude and longitude) partitioning the world (or the spatial area of interest), and β is a fixed time interval. The inverted index of each cell associates each keyword with a list of posts in that cell that contain it. A slice of the grid in the temporal dimension containing posts that were published in an interval of β time units (e.g., one hour) is called a *pane*. The pane collecting the most recent posts is called the *head pane*.

The main memory index only stores the latest ω/β panes, and thus indexes posts that were published within a *sliding window* of ω time units (e.g., one day) in the past. This part of the grid is used by the topic detection module (Section 3.2). On the other hand, the disk-based index stores all panes except the head. This index is used by the topic summarization and the post similarity modules (Sections 3.3 and 3.4).

Besides this hybrid index structure, the storage system of μTOP includes a repository archiving all trending topics, along with their spatio-temporal footprints. The repository receives the continuous output of the topic detection module, and provides input to the topic summarization module when requested.

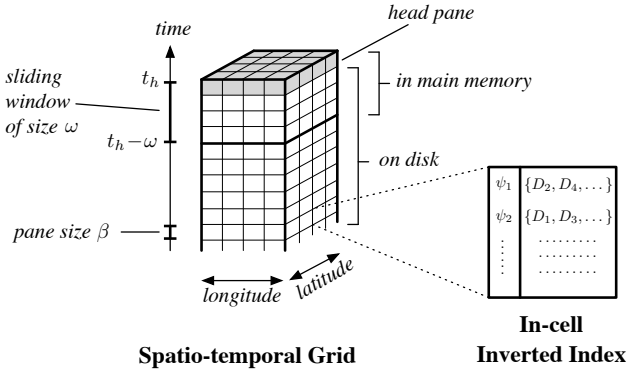

Figure 2: Overview of indexing scheme in μTOP.

3. SYSTEM MODULES

3.1 Preliminaries

First, we need to define textual, spatial and temporal distance functions between posts. Given two posts D_i and D_j, their *textual distance* δ_ψ is measured by the Jaccard similarity between their keyword sets:

$$\delta_\psi(D_i, D_j) = 1 - \frac{|D_i.\Psi \cap D_j.\Psi|}{|D_i.\Psi \cup D_j.\Psi|}.$$

The spatial and temporal distances are measured, respectively, by the Euclidean distance d of the posts' locations and the time difference of the posts' timestamps. To be able to aggregate distance scores across dimensions, we normalize spatial and temporal distances to values in the range $[0, 1]$ (notice that $\delta_\psi \in [0, 1]$). For that purpose, we assume that the posts under consideration are enclosed by a bounding box with diameter length γ and a time interval of length τ. Then, we define the (normalized) *spatial distance* δ_s and *temporal distance* δ_t as follows:

$$\delta_s(D_i, D_j) = \frac{d(D_i.loc\,, D_j.loc)}{\gamma}\ ,\ \delta_t(D_i, D_j) = \frac{|D_i.t - D_j.t|}{\tau}.$$

3.2 Topic Detection

In μTOP, *topic detection* is based on the work presented in [5]. We briefly describe the main aspects of the process below.

To process the incoming stream of posts, a lightweight, in-memory spatial index comprising a uniform spatial grid is used, as explained in Section 2. Upon arrival, each incoming post D is assigned to the corresponding grid cell c according to its geolocation $D.loc$. In each cell, the local stream of posts is processed to generate and maintain locally popular topics with respect to a sliding window W of range ω and sliding step β.

A topic C is characterized by a set of keywords (e.g., hashtags) $C.\Psi$ and is associated with the grid cell c and the time window W in which it is detected. The *popularity* $C.pop$ of a topic C within the cell c and time window W is determined by the number of users having posts in c and W that textually match this topic. We say that a post D matches a topic C if their textual similarity $\delta_\psi(D.\Psi, C.\Psi)$ is above a specified threshold $\theta_\psi \in [0, 1]$. The popularity score of a topic is normalized by the total number of users having posts within the cell c and window W. If an incoming

post does not match any of the existing topics in the current cell and time window, a new topic is created having as keywords those appearing in this post. Eventually, those topics with popularity higher than a specified threshold $\theta_u \in [0,1]$ are marked as *locally trending*, and are returned.

If the same topic is detected in multiple cells and/or time windows, these are merged to construct the topic's *spatio-temporal footprint* $C.\mathcal{F} = \{(c_i, W_i)\}$. Hence, this process not only detects locally popular topics but also explicitly associates each one with the exact geographic region(s) and time period(s) within which it was popular.

3.3 Topic Summarization

Once topics are detected, the next step is to get a summarized overview of each topic. A summary of a topic is already provided by the set of keywords defining it and its spatio-temporal footprint. However, a list of representative posts may also be needed in order to describe the topic in more detail.

For this purpose, μTOP can generate a summary, comprising k posts, for any part of the topic's spatio-temporal footprint. In other words, it can compute a set of k representative posts for any region and time window in which the given topic has been popular. The size of each summary, i.e., the value of the parameter k, can be specified by the user, and can be different for each summary.

The selection of the k representative posts to be included in the summary is based on the criteria of *coverage* and *diversity*. In particular, each summary is constructed by executing a *Coverage & Diversity Aware Top-k Spatial-Temporal-Keyword* (kCD-STK) query, following the approach presented in [4]. We outline the main aspects of this process next.

Formally, a kCD-STK query is defined by a tuple of the form $Q = \langle R, T, \Psi, k\rangle$, where R is a spatial region, T is a time interval, Ψ is a set of keywords, and k is the number of results to return. In our case, the filters R, T and Ψ are derived from the topic's keyword set and spatio-temporal footprint, while k is determined by the desired summary size. The distinguishing aspect of the kCD-STK query is that instead of selecting the top-k posts ranked by relevance, it selects a more representative set of k posts using the criteria of coverage and diversity, which are defined below.

Let $\mathcal{D}_F$ denote the set of all posts satisfying the spatial, temporal and textual filters R, T and Ψ in the query Q. The *coverage* of a post $D \in \mathcal{D}_F$ is defined as the ratio of relevant posts that are within spatial distance θ_s and temporal distance θ_t from D, i.e.:

$$cov(D, \mathcal{D}_F) = \frac{|\{D' \in \mathcal{D}_F : d_s(D, D') \leq \theta_s \wedge d_t(D, D') \leq \theta_t\}|}{|\mathcal{D}_F|}.$$

This is a measure of how representative this particular post is with respect to other relevant posts. Moreover, this is extended to measure the coverage of a set of selected posts $\mathcal{R} \subseteq \mathcal{D}_F$ of size k:

$$cov(\mathcal{R}, \mathcal{D}_F) = \frac{1}{k} \sum_{D \in \mathcal{R}} cov(D, \mathcal{D}_F).$$

Essentially, the criterion of coverage favors the selection of posts from locations that contain a large number of relevant posts.

On the other hand, to avoid a high degree of redundancy, the criterion of *diversity* is used to increase the dissimilarity among the selected posts. Specifically, the diversity of a pair of posts $D_i, D_j \in \mathcal{D}_F$ is defined as:

$$div(D_i, D_j) = \alpha \cdot d_s(D_i, D_j) + (1 - \alpha) \cdot d_t(D_i, D_j),$$

where $\alpha \in [0,1]$ is an adjustable weight parameter between the spatial and the temporal distances. Furthermore, the diversity of a set of posts $\mathcal{R} \subseteq \mathcal{D}_F$ of size k is calculated as:

$$div(\mathcal{R}) = \frac{1}{k \cdot (k-1)} \sum_{D_i, D_j \in \mathcal{R}, i \neq j} div(D_i, D_j).$$

Based on the above, the kCD-STK query returns a set of k posts $\mathcal{R}^*$ that maximizes a combined measure of coverage and diversity:

$$\mathcal{R}^* = \underset{\mathcal{R} \subseteq \mathcal{D}_F, |\mathcal{R}| = k}{\arg\max} \{(1 - \lambda) \cdot cov(\mathcal{R}, \mathcal{D}_F) + \lambda \cdot div(\mathcal{R})\},$$

where $\lambda \in [0,1]$ is a parameter determining the tradeoff between maximum coverage ($\lambda = 0$) and maximum diversity ($\lambda = 1$).

3.4 Retrieving Similar Posts

The above process provides a flexible and adjustable way to get a summary of representative and diverse posts for a topic across the whole extent of its spatio-temporal footprint. Then, the user can further drill down into the topic, by selecting any of the posts in the presented summary that seems interesting, and requesting other similar posts to it. That is, the posts contained in each summary can serve as *seeds* for further exploration of the topic's contents.

This is performed by executing a standard top-k spatial-temporal-keyword query $Q = \langle loc, t, \Psi, k\rangle$, where loc, t, and Ψ are, respectively, the location, the timestamp and the keyword set of the selected post D, and k is the number of similar posts to be retrieved. In this case, the query returns the top-k results ranked by *relevance* determined by an aggregate distance score δ combining the partial distance scores in the spatial, temporal and textual dimensions, i.e.:

$$\delta(D, D') = w_s \cdot \delta_s(D, D') + w_t \cdot \delta_t(D, D') + w_\psi \cdot \delta_\psi(D, D')$$

where $w_s \in [0,1]$, $w_t \in [0,1]$ and $w_\psi = 1 - w_s - w_t$ are weights determining the relative importance of each distance score.

4. USER INTERFACE

The user interface is shown in Figure 3. The map continuously depicts locally trending topics as discovered by the topic detection module. Topics are shown as stars, with brightness indicating popularity. Hovering over a star reveals the topic's spatial footprint, whereas clicking on it shows its keywords together with two options (Figure 4 left). The first option is to invoke the *post similarity* module to retrieve a ranked list of similar posts (in terms of spatial proximity, time closeness, and textual relevance). The resulting posts are displayed in a pop-up window on the right, and also as orange dots on the map and on the timeline located at the bottom.

The second option for a locally trending topic is to explore its spatio-temporal footprint by invoking the *topic summarization* module. The sidebar on the left displays a form detailing the spatial and temporal ranges for the summary, as well as the keywords and the number of returned results (default is ten). Naturally, the user can specify her own summarization request. The summarization results are listed in a pop-up window on the right, where the user can filter them by the top keywords shown at the top. The spatial and temporal distributions of the results are shown on the map and on a timeline at the bottom using orange bullets, respectively. The height of the purple bars in the timeline indicates the average coverage in the corresponding temporal range. Similarly, the purple rectangles on the map illustrate the average coverage in the corresponding regions. The darker the color, the higher the coverage in the area.

Further exploration of the topic summarization results is provided by two means. First, the timeline allows the user to filter the results by selecting a temporal sub-range. This issues a new topic summarization request and updates the results. Second, by clicking on a result on the map, besides showing its content and a link to

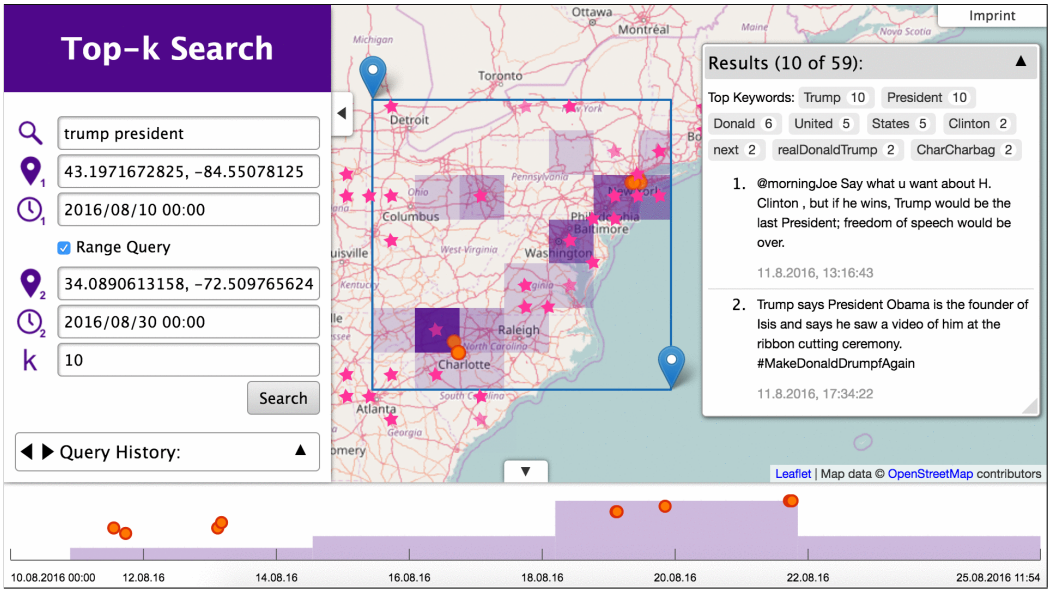

Figure 3: The user interface showing the results of a topic summarization request.

the post, μTOP displays two additional links (Figure 4 right). The one issues a *retrieve similar posts* request, while the other allows the user to further explore the highlighted spatio-temporal region issuing a new *topic summarization* request.

5. DEMONSTRATION

To demonstrate the efficiency and effectiveness of μTOP, tweets are continuously being collected from the public Twitter Streaming API[1]; the current dataset contains over 30 million geotagged tweets with worldwide coverage. The topics are monitored on a stream arriving at an average rate of approximately 500,000 tweets per day. A live demo[2] of μTOP is available online, accompanied by a video[3] explaining and demonstrating its functionality.

Next, we outline a typical usage scenario for demonstration. Initially, the user interface shows locally trending topics on a map, depicted by star icons. Clicking on a star icon reveals the topic's hashtags, for example "#trump #president", as shown in Figure 4. The *explore* link is then used to summarize the topic. It issues a topic summarization request that displays the resulting tweets in a list, on the map and on the timeline. Alternatively, the user may enter query parameters manually using the form in the sidebar on the left, for example to increase the spatial area and time interval.

At the top of the result list a set of keywords is shown that are popular among the result set. This reveals new keywords that are frequently used together with the query keywords *Trump* and *President*. For example *Clinton* is used in 20% of the results. We can click on it to view only those posts that contain this word.

[1]https://dev.twitter.com/streaming/public
[2]http://mtop.imp.fu-berlin.de
[3]https://youtu.be/OmXJUGndaQA

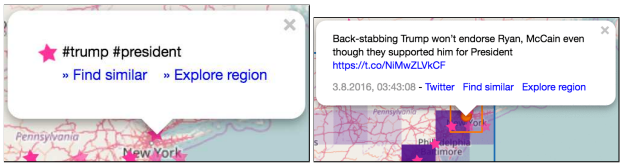

Figure 4: A locally trending topic, and a post summarizing it.

When a topic is summarized, the average coverage is shown as purple blocks and bars in addition to the results. This allows to easily identify spatial regions and time intervals where the topic is popular. For example, Figure 3 shows that the topic is popular around New York City and between the 18th and 22nd of August. This spatial region and time interval can be further explored by issuing another topic summarization request, for example by moving the blue markers on the map or selecting a temporal range on the timeline. We can return to the previous result set by clicking the back-arrow button in the *Query History*, shown in the sidebar.

Instead of summarizing a particular topic, we can also explore a topic by invoking a post similarity search without limiting the spatial and temporal range. By clicking the *Find similar* link, a list of posts similar in spatial, temporal, and textual content is compiled.

Acknowledgements

This work was partially supported by the EU Project City.Risks (H2020-FCT-2014-653747).

6. REFERENCES

[1] H. Abdelhaq, C. Sengstock, and M. Gertz. EvenTweet: Online localized event detection from twitter. *PVLDB*, 6(12):1326–1329, 2013.

[2] C. Budak, T. Georgiou, D. Agrawal, and A. El Abbadi. Geoscope: Online detection of geo-correlated information trends in social networks. *PVLDB*, 7(4):229–240, 2013.

[3] W. Feng, C. Zhang, W. Zhang, J. Han, J. Wang, C. Aggarwal, and J. Huang. STREAMCUBE: hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream. In *ICDE*, pages 1561–1572, 2015.

[4] P. Mehta, D. Skoutas, D. Sacharidis, and A. Voisard. Coverage and diversity aware top-k query for spatio-temporal posts. In *SIGSPATIAL*, page 19, 2016.

[5] K. Patroumpas and M. Loukadakis. Monitoring spatial coverage of trending topics in twitter. In *SSDBM*, pages 7:1–7:12, 2016.

[6] X. Wang, Y. Zhang, W. Zhang, X. Lin, and Z. Huang. SKYPE: top-k spatial-keyword publish/subscribe over sliding window. *PVLDB*, 9(7):588–599, 2016.