# Detecting Dynamic Association among Twitter Topics

Shuangyong Song, Qiudan Li, Hongyun Bao
State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{shuangyong.song, qiudan.li, hongyun.bao}@ia.ac.cn

## ABSTRACT
Over the last few years, Twitter is increasingly becoming an important source of up-to-date topics about what is happening in the world. In this paper, we propose a dynamic topic association detection model to discover relations between Twitter topics, by which users can gain insights into richer information about topics of interest. The proposed model utilizes a time constrained method to extract event-based spatio-temporal topic association, and constructs a dynamic temporal map to represent the obtained result. Experimental results show the improvement of the proposed model compared to static spatio-temporal method and co-occurrence method.

**Categories and Subject Descriptors:** H.3.3 [**Information storage and retrieval**]: Information search and retrieval – *Information filtering.*

**General Terms:** Algorithms, Design, Experimentation.

**Keywords:** Twitter, Burst detection, Topic association, Dynamic temporal map.

## 1. INTRODUCTION
Twitter, as one of the most popular micro-blogging services, has already attracted 200 million registered users up to April 29, 2011. Fast diffusion of information makes Twitter a convenient platform for users to generate and seek new trends about topics of interest. Generally, a topic may have different related topics at different time. Those dynamic associations among topics are generated from real-world events [1]. For example, 'Michael Jackson' always came together with 'vocal concert' around Jun 23, 2009, but co-occurred with 'funeral' around Sep 3, 2009. By detecting those event-based associations among topics, a user can better understand the details of topics he is interested in, such as their development tendency and associated events.

Existing study has focused on detecting topic association from text contents by co-occurrence method [6]. This approach is beneficial when the average length of texts is not short. However, it is hard to discover the co-occurrence relationship between topics in tweet-like short texts. In [5], we proposed a spatio-temporal model for topic association detection in Twitter, and this model can effectively discover potential correlations among Twitter topics. In this paper, we extend our work in [5] by proposing a dynamic topic association detection model, for discovering related topics with query topic in its different bursty periods.

Well-noticed events about a topic can result in message activity bursts for it, and most of those messages in an activity burst describe a common event [7]. Therefore, we can discover query topic's dynamic association with other topics by detecting their co-bursting relationship when events occur. In addition, related topics tend to present similar temporal dynamics and location statistics in a specific period of time [5], which is useful for detecting topic associations in Twitter. The proposed model first utilizes a burst detection algorithm to extract burst periods of query topic, then employs a co-bursting judgment method to find the potential related topics with it, finally calculates the spatio-temporal similarity between topics to provide user a related topic list. In the following of this article, we will illustrate the details of problem definition and our proposed model, along with the experimental results.


*WWW 2012 Companion,* April 16-20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

## 2. PROBLEM DEFINITION AND METHOD DESCRIPTION

### 2.1 Problem Definition
In this paper, we use the definition of *topic* as "any subject of interest to a user" [3], examples include Lakers, Roberto Baggio, Christmas, etc. A set of posts related a topic $k$ are defined as $P_k = \{p_{k1}, p_{k2}, \ldots p_{kn}\}$, where $n$ means the number of posts. Then we define the gaps in time between posts as $G_k = \{g_{k1}, g_{k2}, \ldots g_{k(n-1)}\}$, where $g_{ki}$ means the time interval between $p_{ki}$ and $p_{k(i+1)}$. Bursts of a topic are described as 'grow in intensity for a period of time, and then fade away' [2]. We define bursty periods of topic $k$ as $B_k = \{b_{k1}, b_{k2}, \ldots b_{km}\}$, where $b_{kj}$ means the $j^{th}$ detected burst period of topic $k$. We further represent a topic $k$ as the following spatio-temporal state series: $Topic_k = \{k_1, k_2, \ldots, k_i, \ldots, k_I, k_{(I+1)}, k_{(I+2)}, \ldots, k_{(I+S)}, \ldots, k_{(I+S)}\}$, where $k_i$ ($i \in [1, I]$) means the frequency of topic $k$ in the $i^{th}$ time interval, $k_{(I+s)}$ ($s \in [1, S]$) means the numbers of users who have posted tweets about topic $k$ in $s^{th}$ region. Then we can calculate spatio-temporal relationships between topics within a fixed period [5]. Given a query topic and queried time window, our goal is to detect the event-based spatio-temporal related topics with query topic in its different bursty periods.

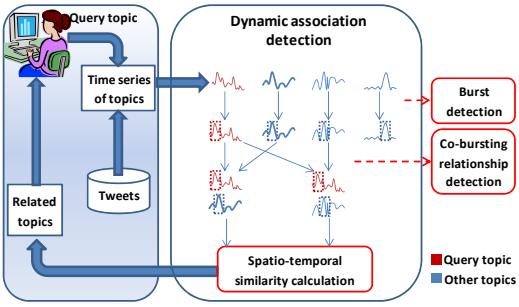


**Fig. 1. The framework of the proposed model**

### 2.2 Method Description
The framework of the proposed model is shown in Fig. 1. We first detect query topic's burst periods, then discover its co-bursting relationship with other topics, which can help find the potential related topics. Finally, we rank topics' associate degree with a spatio-temporal similarity computing method.

**Burst Detection** Burst detection is used to discover topics' event-based bursty periods. For an arbitrary topic $k$, we adopt the

burst detection technique proposed in [2] to obtain $B_k$ from $G_k$. With the number of bursty intensity states $z$ determined, we can calculate bursty intensity states of all gaps in $G_k$, which are between 0 and $z$. $C_j(i)$ is defined to be the minimum cost of $g_{ki}$ ending with state $j$, and $\underset{j}{\operatorname{argmin}} C_j(i)$ is chosen to be the state of $g_{ki}$. The formula of $C_j(i)$ is defined as $C_j(i) = -\ln f_j(g_{ki}) + \min_l(C_l(i-1) + \tau(l,j)), (0 \le l \le z)$, with initial conditions $C_0(0) = 0$ *and* $C_j(0) = \infty$ for $j > 0$. In the above formula, $f_j(g_{k^*})$ is a function representing the distribution rule of $G_k$, and $\tau(l, j)$ is a function of cost incurred by moving from state $l$ to state $j$. In addition, *burst of intensity* $v$ is defined to be a maximal interval over which states of index $v$ or higher persist. We define the sequence of those intervals as $B_t$. Finally, we delete the intervals less than one day, which are unlikely to be real bursty periods, and define $B_t'$ to save the left intervals.

**Co-bursting Relationship Detection** Topics show bursty states when events occurred and we can detect the event-based associations between topics by discovering the co-bursting relationship of them [1], where the co-bursting relationship is defined as 'a burst of the query topic share a common time window with a burst of another topic'. If $\exists b_{A^*} : b_{A^*} \wedge b_{Qi} \neq \varnothing$, where $b_{A^*}$ is an arbitrary bursty period of topic A, and $b_{Qi}$ is the $i^{th}$ bursty period of query topic, we define this situation as topic A has a co-bursting relationship with query topic in $b_{Qi}$. In this paper, given a query topic, we detect its co-bursting topics in each bursty period of it, which we will take as candidate related topics.

**Spatio-temporal Similarity Calculation** After representing topics as the spatio-temporal state series, we utilize Euclidean distance to measure the spatial similarity $Sim_S$ and temporal similarity $Sim_T$ respectively, then use a parameter $\lambda$ to adjust the significance of temporal similarity and spatial similarity as $Sim = \lambda Sim_T + (1-\lambda) Sim_S$. Finally, we calculate the integrated spatio-temporal similarities between the query topic and its potential related topics in its different bursty periods, and used the similarities as weights to get the rank list.

# 3. EXPERIMENTS

## 3.1 Dataset and Parameter Settings

We use Twitter API to gather tweets data about hot topics in Twitter. We first download hot topics appeared from Dec 4, 2011 to Dec 10, 2011, then track the tweets about those topics from Dec 11, 2011 to Dec28, 2011. The dataset finally consists of 1778 topics with 5843527 tweets. The parameter $k$ in burst detection was set to be 8 with the method described in [2], and we set $v$ to be 1, taking into account our data size. Finally, $\lambda$ is chosen to be 0.59 after the cyclic iterative method [5].

## 3.2 Results and Discussions

We measure the effects of our proposed model using top-k precision — the percentage of accurate results in the top-k results [4]. Accordingly, two additional topic association detection methods were conducted as baselines:

**Static spatio-temporal method:** like in [5], calculating the spatio-temporal similarity between topics without the co-bursting relationship detection step.

**Co-occurrence method:** using the co-occurrence method given in [6] to detect associations between topics.

For our experimental evaluation, we have 3 PhD candidates manually label the related topics. Manually examining all the topics in our data set is prohibitively expensive, so we choose 30 topics randomly from our experimental dataset. In each bursty period of the 30 topics, most relevant 5 topics are tagged.

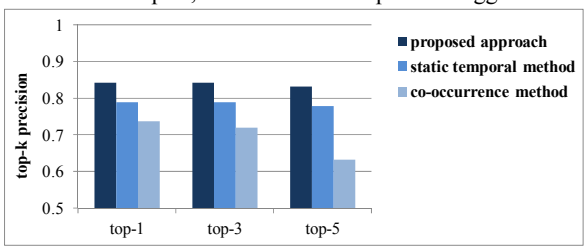


**Fig. 2. Evaluation results of comparing three methods**

Fig. 2 demonstrates the evaluation results, from which we can see our method outperforms the Co-occurrence method and Static spatio-temporal method. The improvement is due to the integration of co-bursting relationship detection. In addition, the two spatio-temporal similarity based method perform better than co-occurrence method, this result proves the effect of spatio-temporal similarity based method on topic association detection in Twitter-like short texts. We further construct a dynamic temporal map to represent the obtained result, which is shown in fig. 3. We can see a series of topics related to the American country music singer Toby Keith, such as his song 'Made in America', and some other singers who have various relationships with Toby Keith. From this dynamic temporal map, we can conveniently observe the related topic with the query in its different bursty time periods.

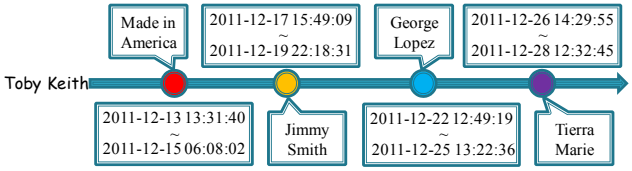


**Fig. 3. An example of dynamic temporal map**

# 4. CONCLUSIONS AND FUTURE WORK

In this paper, we provide a dynamic topic association detection model by discovering topics' spatio-temporal relationship in their bursty periods. Preliminary experiments support the effectiveness of the proposed model. In the future, we would like to add some new features to detect relationship among topics. This research is supported by the NNSFC project 61172106 and the BJNSF project 4112062.

# 5. REFERENCES

[1] Sarmay, A., Jainy, A. and Yu, C. 2011. Dynamic relationship and event discovery. In WSDM'11, pp 207–216.

[2] Kleinberg, J. 2002. Bursty and hierarchical structure in streams. In SIGKDD'02, pp 373–397.

[3] Sehgal, A.K. and Srinivasan, P. 2007. Profiling topics on the web. In WWW workshop' 07, pp 1–8.

[4] Liu, J., Dong, X. and Halevy, A. Y. 2006. Answering structured queries on unstructured data. In WebDB'06.

[5] Song, S., Li, Q. and Zheng, N. 2010. A spatio-temporal framework for related topic search in micro-blogging. In AMT' 10, pp 63–73.

[6] Terachi, M., Saga, R. and Tsuji, H.2006. Trends recognition in journal papers by text mining.In IEEE/SMC'06,pp 4784–4789.

[7] Grinev, M., Grineva, M., Boldakov, A., Novak, L., Syssoev, A., Lizorkin, D. 2009. Sifting Micro-blogging Stream for Events of User Interest. In SIGIR' 09, pp 838.