Cost-Effective Node Monitoring for Online Hot Event Detection in Sina Weibo Microblogging

Kai Chen
Shanghai Jiaotong University
kchen@sjtu.edu.cn

Yi Zhou*
Shanghai Jiaotong University
zy_21th@sjtu.edu.cn

Hongyuan Zha
Georgia Institute of Technology
zha@cc.gatech.edu

Jianhua He
Aston University
j.he7@aston.ac.uk

Pei Shen
Shanghai Jiaotong University
shenpei310@sjtu.edu.cn

Xiaokang Yang
Shanghai Jiaotong University
xkyang@sjtu.edu.cn

ABSTRACT

We propose a cost-effective hot event detection system over Sina Weibo platform, currently the dominant microblogging service provider in China. The problem of finding a proper subset of microbloggers under resource constraints is formulated as a mixed-integer problem for which heuristic algorithms are developed to compute approximate solution. Preliminary results show that by tracking about 500 out of 1.6 million candidate microbloggers and processing 15,000 microposts daily, 62% of the hot events can be detected five hours on average earlier than they are published by Weibo.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications; H.4 [**Information Systems Applications**]: Miscellaneous; I.2.6 [**Artificial Intelligence**]: Learning

Keywords

Microblog; event detection; subnet; greedy algorithm

1. INTRODUCTION

Microblogging has become a popular means of communication, information diffusion and marketing. Sina Weibo has more than 300 million registered users and about 100 million messages are posted on Sina Weibo daily [1]. With microblogging's increasing importance as sources of news updates and information dissemination, the problem of identifying and predicting hot events from microblogging services is receiving increasing research and development interests. Most of the reported works on microblog hot event detection have been focused on approach of retrieving and processing all the microposts generated in the interested periods [2, 3]. However, microblogging has distinct features in terms of the number of users and the volume and the size of microposts. retrieving and processing all the microposts requires substantial communication and computational capacity.

*Corresponding Author.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

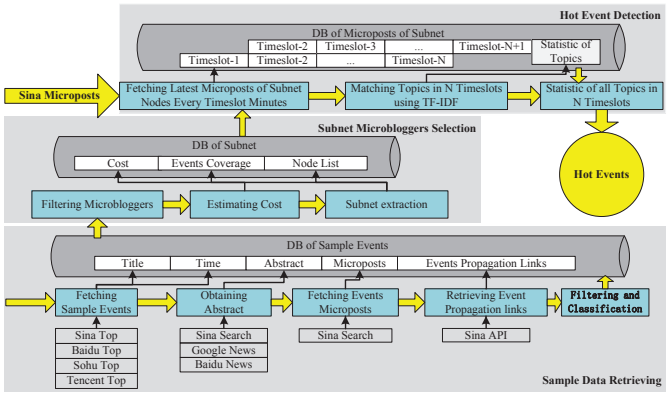

Figure 1: Overall hot events detection system.

In our ongoing research we are investigating cost-effective detection of hot events in Sina Weibo under resource constraints on computation and communication capacity. For the studied problem we strategically track only a small subset of the microbloggers under the resource constraints while maintaining good detection quality. A mixed-integer optimization problem is formulated for the selection of microbloggers to monitor and heuristic algorithms are developed to solve the optimization problem. Through extensive experiments with real-time test data retrieved from Sina Weibo it is shown that we can achieve cost effective online detection of hot events.

2. DETECTION SYSTEM AND EVALUATION

We developed a hot event detection system with an overall framework shown as Figure.1. The system consists of three main modules: 1) Sample data retrieving; 2) Subnet microbloggers selection; and 3) Hot event detection. These modules work together to reduce communication and computation loads for hot event detection, and improve hot event detection quality.

Selection of subnet microbloggers is a key problem for the detection system. Node selection problem has been addressed in [4][5], but only single cover is considered for event detection. We extended the method proposed in [5] for our research problem, but found that the extended method has

a poorer performance than our proposed one to be presented next.

Assume that we have a given set of nodes in the microblog network, which is denoted by $\mathcal{V}$ with $|\mathcal{V}| = N$. Suppose that for each node $v \in \mathcal{V}$ it generates m_v microposts on average daily. Let $\mathcal{E}$ denote the set of retrieved sample hot events. Define a binary variable $a_{v,e}$ to reflect the participation of event e ($e \in \mathcal{E}$) by a node v ($v \in \mathcal{V}$), with $a_{v,e} = 1$ if node v participated in the event e and 0 otherwise. Let's define *Degree of Covers (DoC)* as a measure of event coverage. For a general event e and a given subset of nodes $\hat{\mathcal{S}}$, the *DoC* of event e by $\hat{\mathcal{S}}$ is the total number of nodes in $\hat{\mathcal{S}}$ that participated in event e, denoted by $D_e(\hat{\mathcal{S}})$. The threshold on the *DoC* for a hot event e to be detected is set to X_e ($e \in \mathcal{E}$).

Based on the above settings, an optimization problem of selecting a subset $\mathcal{S}$ of nodes from $\mathcal{V}$ can be formulated with objective of maximizing covers for the hot event under the resource constraints. Let x_v be a binary variable such that $x_v = 1$ if and only if node $v \in \mathcal{V}$ is selected to follow. By setting the constraints on the *DoC* of each individual hot event with the selected nodes, we can obtain the following optimization problem:

$$\min \sum_{e \in |\mathcal{E}|} \xi_e \quad (1)$$

$$s.t. \quad \sum_{v \in \mathcal{V}} m_v x_v \leq M; \ \sum_{v \in \mathcal{V}} a_{v,e} x_v \geq X_e - \xi_e, \quad e \in \mathcal{E},$$

$$\sum_{v \in \mathcal{V}} x_v \leq K; \ \xi_e \geq 0, \quad e \in \mathcal{E}; \ x_v \in \{0,1\}, \quad v \in \mathcal{V}.$$

M is the constraint on the average number of microposts to receive and K on the number of nodes.

The optimization problem is a mixed integer programming, which is NP-hard in general. Therefore we proposed a heuristic algorithm to solve it, which is called *NewInd*. A key idea with *NewInd* algorithm is to find a node maximizing event coverage among the nodes with costs less than a continuously updated cost bound. Intuitively if there are a relatively large number of candidate nodes to select, there is a larger cost budget to find nodes with better coverages and we can allow the next node to be selected having relatively larger node cost. Based on this consideration we empirically set a cost bound associated with nodes to be selected, which is denoted by $M_b(N_l, M_l)$, if there are M_l budget available on the number of microposts and N_l nodes that can be added. For any node that has not been selected to monitor, if its associated cost is larger than $M_b(N_l, M_l)$, that node is not to be considered in the next round of node selection. $M_b(N_l, M_l)$ is calculated by:

$$M_b(N_l, M_l) = \frac{M_l}{N_l^{0.7}} + \frac{M_l}{N_l}. \quad (2)$$

A complete algorithm has been designed based on the above cost bound formula and evaluated. We have implemented the detection system in Linux using a generic Dell PC. The proposed algorithm was compared with two simple algorithms. One is called algorithm *FM*, which iteratively picks the next node with the most followers; while the other is called algorithm *ECM*, which iteratively picks the next node participating in the most of hot events. We considered three combined configurations on the system parameters (M, X_e, K): (6600, 22, 250), (10000, 35, 500) and (25000, 65, 1500). We collected 739 hot events in total from hot event ranking websites spanning from 21th September to 20th October, 2012. There are 16,047 hot propagation links, 2,605,032 retweets and 1,622,843 nodes. These events were used as sample dataset to select the subset of nodes.

Table.1 presents system performance with various node selection algorithms over the real-time test dataset. Column 'R-K' means the actual size of selected subset of nodes. Column 'N-BST' denotes the percentage of detected Top-10 events published by Baidu/Sohu/Tencent top-lists, and column 'N-Sina' means the percentage of detected Top-10 events published by Sina Top. Columns 'BST-Time' and 'S-Time' mean the average time that hot events are detected in advance of events published by Baidu/Sohu/Tencent top-lists and Sina top-lists, respectively.

Table 1: Algorithms comparison over test dataset

Alogrithm	R-K	N-BST	BST-Time	N-Sina	S-Time
FM	250	43.2%	13.4h	45.8%	3.2h
FM	500	28.4%	9h	27.1%	2h
FM	1500	40.9%	8.9h	29.9%	2.2h
ECM	250	55.7%	15.4h	52.3%	2.85h
ECM	500	52.2%	16.2h	55.1%	3h
ECM	1500	53.4%	15.5h	56.1%	3.6h
NewInd	225	56.8%	13h	62.7%	3.8h
NewInd	437	61.3%	16.2h	64.1%	6.7h
NewInd	1094	64.3%	18.5h	65.5%	7.3h

From these results we can observe that $NewInd$ performs much better than FM and ECM. By tracking about 500 out of 1.6 million candidate microbloggers and processing about 15,000 microposts daily, about 62% of the hot events could be detected five hours earlier than their publication time at Sina Weibo. The percentage of detected hot events with $NewInd$ increases with the cost budget in the range of 2000 to 15000 (microposts), but does not change much with further cost budget increase. This indicates that a large system cost budgets may not necessarily increase the percentage of detected hot events. Therefore we may just need to select a small number of nodes to monitor for hot event detection. In our future work we plan to test the system with data at larger scale and extend the system to Twitter microblogging service.

3. ACKNOWLEDGMENTS

The work is partially supported by National Natural Science Foundation of China (Grant No. 61129001,No.61201384) the National Grand Fundamental Research 973 Program of China (Grant No.2010CB731406), and Shanghai Science and Technology Committees of Scientific Research Project (Grant No. 11dz1505502).

4. REFERENCES

[1] www.weibo.com
[2] M. Mathioudakis, Twittermonitor: trend detection over the twitter stream. In ICMD'10.
[3] M. Cataldi et al. Emerging topic detection on Twitter based on temporal and social terms evaluation. In IWMDM'10.
[4] W. Chen et al. Efficient influence maximization in social networks. In ACM SIGKDD'09.
[5] J. Leskovec et al. Cost-effective outbreak detection in networks. In ACM SIGKDD'07.