Event detection using blog tags

Chen, Wenda

2008

Chen, W. (2008, March). Event detection using blog tags. Presented at Discover URECA @ NTU poster exhibition and competition, Nanyang Technological University, Singapore.

https://hdl.handle.net/10356/79531

NANYANG TECHNOLOGICAL UNIVERSITY

URECA *Undergraduate Research Experience on CAmpus*

Category: 4
Project ID:SCE07052

Event Detection Using Blog Tags

Background

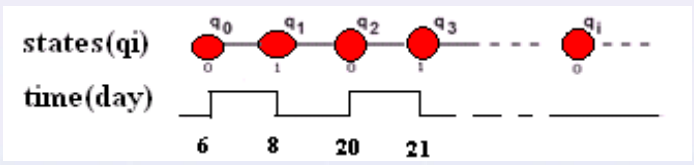
Bloggers often add tags (or keywords) chosen from an uncontrolled vocabulary to blog posts to enable browsing and searching on posts. Many of the tags are related to the events described in the blog posts. The purpose of this project is to detect events based on the tags and also the blog posts content. The evolvement of the tags used to describe the same event at different time points reflects the change of views among bloggers.

Event Definition

An event is represented by a subset of blog posts describing what happened in a particular place and location. With the tags attached to blog posts, we aim to firstly obtain those tags related to the same event and then select blog posts according to their tags and publication time.

Event Detection Algorithm

Step 1: obtain all the occurrences of a given tag in ascending order of publish time. Calculate the time gaps *x* to form a sequence *q*:



Step 2: apply 2-state automaton algorithm[1]. State for each occurrence is determined by a genetic algorithm to find the state sequence *q* that can maximize the function

$$\Pr[q \mid x] = \frac{\Pr[q]fq(x)}{\sum_{q'} \Pr[q']fq'(x)} = \frac{1}{Z}(p)^{b}(1-p)^{n-b}\prod_{t=1}^{n} f_{t}(x)$$

where *p* is consecutive state transition probability and $f_{t}(x)$ is the probability of time gap *x* in state *t*

Step 3: cluster the time series a of the tags in set B into events E by maximizing

$$P(E_K \mid D) = \frac{P(D \mid E_K)P(E_K)}{P(D)} \qquad P(E_K) = \frac{\bigcap_{j=0}^{|B|} \mid e_j = 1 a_j}{\bigcup_{j=0}^{|B|} \mid e_j = 1 a_j}$$

$$P(D \mid E_K) = \prod_{j=0}^{|B|} \left(\frac{|D_j|}{|M|}\right)^{e_j}\left(1-\frac{|D_j|}{|M|}\right)^{1-e_j} \qquad M = \bigcup_{\substack{j=0 \\ e_j=1}}^{|B|} D_j$$

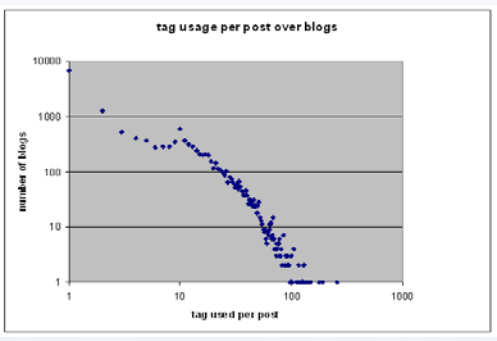
which considers the co-occurrence of the tags belonging to the event Ek (indicated by e_k =1) and the percentage of posts D carrying the tag over the period of the time when the event is bursty [2].

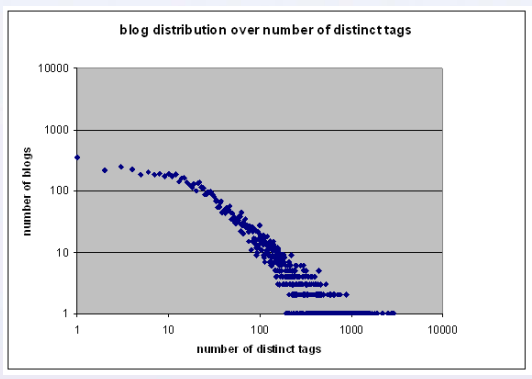
Future Work

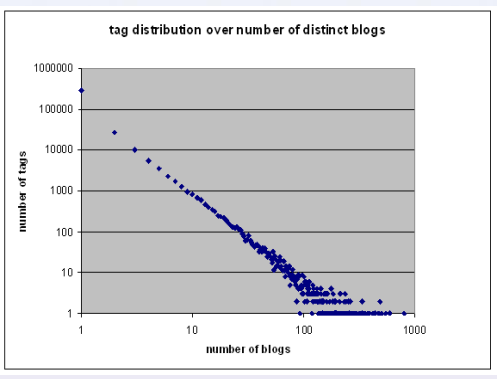
In our future work, we plan to study event detection based on the named entities in blog posts together with the tags and also event evaluation methods.

Dataset Statistics

Blogs with tags	8593
Blog posts	3316871
tags	2572401
Distinct tags	657036







Experiment Output

Tag	Event	Occurrence period and state state 0/1: non-bursty/bursty
NASCAR Busch	National Association for Stock Car Auto Racing	22/02/2007 to 01/07/2007: state 0; 01/07/2007 to 03/07/2007: state 1; 03/07/2007 to 07/09/2007: state 0; 07/09/2007 to 15/09/2007: state 1
Blessed Mother	Catholic Festival	17/02/2006 to 25/06/2007: state 0; 25/06/2007 to 03/07/2007: state 1; 03/07/2007 to 25/08/2007: state 0; 25/08/2007 to 02/09/2007: state 1; 02/09/2007 to 21/09/2007: state 0; 21/09/2007 to 23/09/2007: state 1
All Saints	Fornham Saint Martin Fornham All Saints	09/11/2006 to 11/01/2007: state 0; 23/01/2007 to 11/09/2007: state 1
newcastle	Joey Barton joined club Newcastle	01/10/2006 to 19/03/2007: state 0; 16/06/2007 to 25/08/2007: state 1

References:
[1] Jon Kleinberg. Bursty and hierarchical structure in streams. In Proc. SIGKDD 2002.
[2] Fung et al. Parameter free bursty events detection in text streams. In Proc. VLDB 2005.

School of Computer Engineering
Project Title: Dynamic Interactive Tag Clouds
Student: Chen Wenda
Supervisor: Asst Prof Sun Aixin
Collaborator: A/P Lim Ee Peng