Convergence of Influential Bloggers for Topic Discovery in the Blogosphere

Shamanth Kumar, Reza Zafarani, Mohammad Ali Abbasi, Geoffrey Barbier, and Huan Liu

Computer Science and Engineering
Arizona State University
Tempe, AZ 85281
{shamanth.kumar,reza,ali2,geoffrey.barbier,huan.liu}@asu.edu

Abstract. In this paper, we propose a novel approach to automatically detect "hot" or important topics of discussion in the blogosphere. The proposed approach is based on analyzing the activity of influential bloggers to determine specific points in time when there is a convergence amongst the influential bloggers in terms of their topic of discussion. The tool BlogTrackers, is used to identify influential bloggers and the Normalized Google Distance is used to define the similarity amongst the topics of discussion of influential bloggers. The key advantage of the proposed approach is its ability to automatically detect events which are important in the blogger community.

1 Introduction

For thousands of years human beings have been interested in understanding and measuring their surroundings. The Greek mathematician Thales used a method known as triangulation to measure the height of objects. A general concept of triangulation involves using two or more perspectives to accurately locate something else. The approach of using multiple perspectives to gain a more accurate measure or understanding of the details of an object, triangulation, is applied in other domains as well. Surveyors plan roads, geologist study earthquakes, and even biologist use radio triangulation to study wildlife behavior. Similar to how others use triangulation in their fields of work, we believe it is possible to detect significant events and produce a more accurate sentiment of the blogosphere by examining how much independent influential bloggers have in common with each other during a finite time period. We consider a significant event, a news topic, opinion, or other trigger that motivates the influential bloggers to write about some subject during the same period of time. The subject is significant in that it changes the focus to the subject from the previous and usually disparate topics that the independent influential bloggers were focussed on before the significant event.

2 Motivation

Social Media has played an increasingly significant role in our everyday life. We use Facebook to connect to our friends, YouTube to share videos, etc. The blogosphere is one of the most popular facets of Social Media. It has shown consistent exponential growth over the past few years. State of the Blogosphere[1], a study of bloggers performed annually by Technorati, presents some interesting results on the importance of the blogosphere. In the latest study it was observed that 76% of the bloggers who were surveyed, blogged to express their opinions. Additionally, the blogosphere captures more detailed aspects of public opinion than polling and surveys over a wider geographical area without traditional limits of time, space, and survey administrators. Opinions are associated with topics, which generally correspond to a real world event like the launch of a new mobile phone or the elections in a country. Topic discovery is therefore, an essential first step in the analysis of the blogosphere. An effective way to detect topics significant from the blogging community's perspective is to look at the discussions amongst influential bloggers in a community. The influential bloggers can be considered to be the first to start conversation on a key topic and hence provide a way of detecting what could be termed as the "hot" topics amongst the bloggers.

As the Internet continues to play a major role in communication among people from various communities, organizations, and nations, there is a growing interest in analyzing world wide web content to better understand the culture, sentiment, and social relationships amongst the people that use the web and provide a variety of social information. Computational methods can aid researchers in mining the vast amount of social computing data that is available today and help pave the way for a deeper understanding of how human actions in the "cyber-world" correlate with the "real-world."

3 Related Work

Some of the leading blog indexing services have blog search and topic tracking features. BlogPulse[2] supports conversation tracking, which looks at blog-roll data to track replies to a blog post, which when put together form a conversation. Technorati[3] supports advanced blog search capabilities including searching for other blogs that link to a specific blog post. Technorati also provides a list of key topics, videos, and people being discussed in the current blog posts. BlogScope [3] is an analysis and visualization tool for the blogosphere, which currently indexes 133 million blogs. This web based tool is capable of generating popularity trends for specific topics, identifying information bursts, and performing geographical search. Our tool, BlogTrackers [1], combines unique blog and blogger

[1] http://technorati.com/blogging/article/day-1-who-are-the-bloggers1/
[2] http://www.blogpulse.com
[3] http://technorati.com

analysis capabilities by identifying topics of discussion, influential bloggers in the community, and unique topic tracking features.

Influential bloggers can be identified by analyzing the link structure of the blogs as suggested in PageRank [9] and HITS [7] which can identify important nodes in a community. Authors in [11] propose the InfluenceRank algorithm which combines link-structure information and the novelty of the blog content to detect opinion leaders in a community.

Topic discovery and tracking has been a popular research area in information retrieval. In [2, 10], the authors discuss the use of vector spaces in detecting topics. Tools such as [6, 5] focus on the temporal analysis of topics to provide context to identified topics.

4 Methodology

An important issue in the analysis of the blogosphere is the detection of topics. In this section, we propose a new approach to detect "hot" topics by using the information from the blog posts of the influential bloggers. An influential blogger can be defined as a blogger whose opinions can influence the opinions of a significant number of other individuals in the community. This makes influential bloggers an important part of the blogosphere who can be targeted for advertising or for gauging the political sentiments of the community. Influential bloggers have their niche topics on which they are able to exert influence on others and we assume that they do not talk about similar topics unless a real world event of high significance such as elections or a new product launch occurs and motivates them to start talking about these topics. By focusing our attention on the topics discussed by the influential bloggers we hope to have a more efficient approach in detecting "hot" topics in the community.

Each influential blogger in the community can be represented as a vector of high frequency terms extracted from his blog posts for a specific period of time. Therefore, every blogger I_i can be defined as a vector of high frequency terms t_j as,

$$I_i = \{t_j | 0 \leq j \leq n\} \tag{1}$$

This vector of high frequency terms can be used to identify the convergence of the influential bloggers by measuring the distance between the average term frequency vector and that of a particular blogger. Due to the inability to use systems like WordNet to identify the semantic relationship between Indonesian terms, we used the Normalized Google Distance(NGD) [4] to measure semantic relationship between the terms of two vectors. The NGD distance measure calculates the normed semantic distance between two terms using the number of Google search results for the terms jointly and independently as a measure of their semantic relationship. The Normalized Google Distance for two terms x and y can be defined as,

$$NGD(x, y) = \frac{max\{logf(x), logf(y)\} - logf(x, y)}{logM - min\{logf(x), logf(y)\}} \tag{2}$$

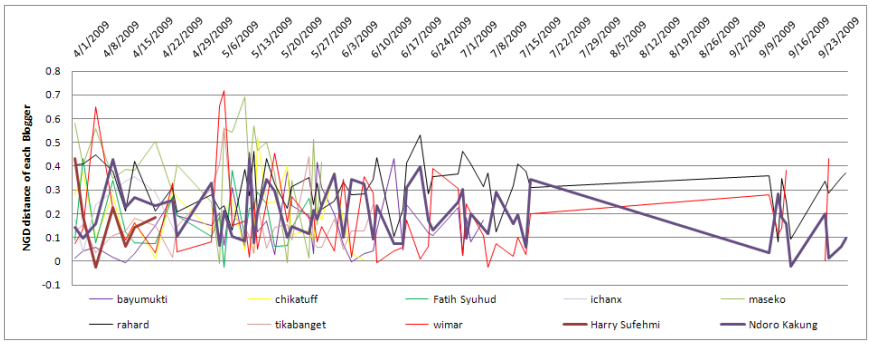

Fig. 1. NGD distance between the top 10 influential bloggers and the average terms during the period April, 2009 to August, 2009

where f(x) and f(y) are the number of search results returned by Google for the search query x and y respectively, f(x,y) denotes the number of search results returned by Google for the search query containing both x and y, and M denotes the number of pages searched by Google, which is estimated to be around 10^{10}. If two words a and b have a distance smaller than the distance between a and c, then the two words a and b are said to be semantically closer to each other than a and c.

The similarity of influential bloggers, in terms of their topics of discussion can be defined as follows,

$$S_i = P_i \times B_i, \forall i \tag{3}$$

where P_i denotes the number of blogposts published by all the influential bloggers within a given time interval i and B_i denotes the number of influential bloggers during the time interval i. S gives us an estimate of the amount of activity, which if significantly greater than the neighboring time periods indicates a key or a significant event being discussed by the influential bloggers. P is used to scale the value of B to remove irrelevant points from consideration. The number of relevant influential bloggers during any given period B can be defined as

$$B_i = \Sigma b_i, b_i < T \tag{4}$$

where T is a suitable threshold which can be varied to increase or decrease the number of relevant influential bloggers B. If both P and B values are high then this is an indication that there is a significant proportion of blog posts from a significant proportion of influential bloggers, the topics from which together converge towards the average topic of discussion amongst all the influential bloggers.

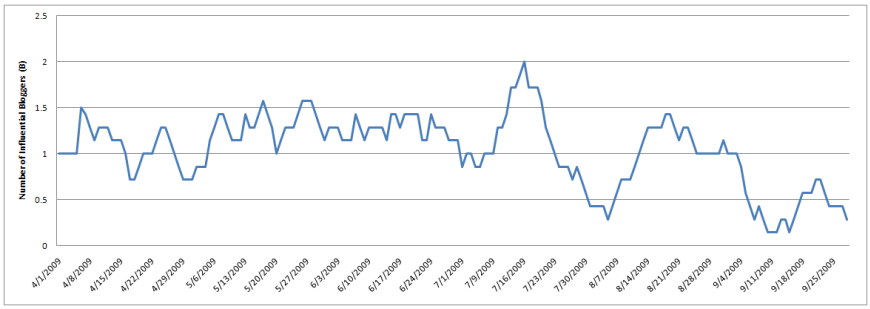

Fig. 2. Number of influential bloggers (B) whose deviation from average below the threshold T

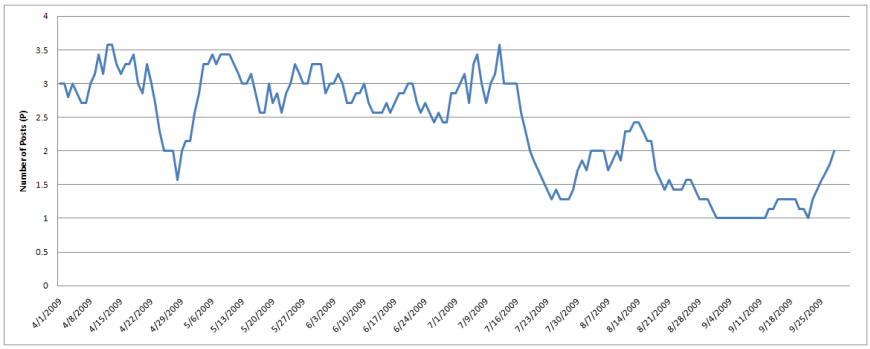

Fig. 3. Total number of blog posts (P) during each time interval

5 Case Study

In order to explore the convergence of topics of influential bloggers, we identified the top 10 influential bloggers from a dataset comprising of 50 Indonesian blogs [8] crawled using BlogTrackers for the period April 2009 to September 2009. To represent the topics discussed by influential bloggers, we computed the top 10 keywords and their frequency, which formed the term-frequency vector for each blogger, from the blog posts published by them during all the time periods for the aforementioned period. In this case study we fixed the time interval to consecutive periods of 7 days each. To compute the average term frequency vector for all the influential bloggers presented in Figure 1 we combined the already identified top keywords for individual bloggers for each period and then chose the top keywords from this list as the representative or average keywords for the duration. Using NGD, described in Section 4 we identified the deviation of each blogger from the average term vector by computing the average of the pairwise

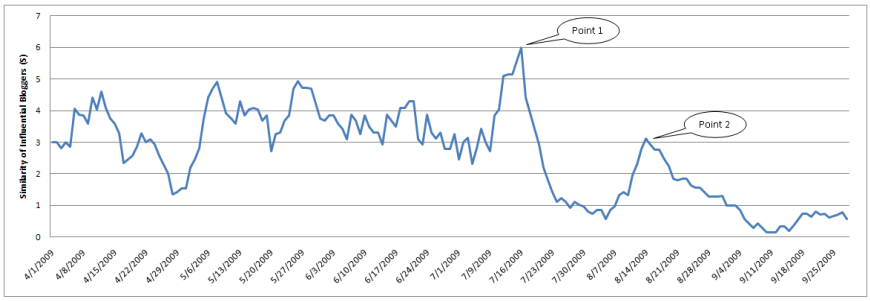

Fig. 4. Similarity (S) of influential bloggers for the period April, 2009 to September, 2009

NGD distance between the terms in the two given vectors. These values for each blogger is presented in the Figure 1.

For the purpose of this case study we fixed the value of the threshold T to 0.2. We computed the number of relevant influential bloggers for each time interval using the definition provided in Eq 4, which is presented in the Figure 2. The value of B as shown the figure should be as large as possible for the convergence of the influential bloggers to represent the overall sentiment of the influential blogger community. To compute the weights P for each time interval i we determined the average number of blog posts published during each time interval by all the influential bloggers and these values are shown in Figure 3.

We present the similarity of influential bloggers S computed using the definition from Eq 3 in Figure 4. We hypothesize that the peaks in the figure represent points of convergence of a majority of influential bloggers to the average topic of discussion amongst all the influential bloggers under consideration. The keywords corresponding to these points would most likely represent "hot" topics pertinent from the blogging community's perspective. After analyzing two of the most recent points as highlighted in Figure 4 we found that Point 1 represented keywords such as "president", "sby", "elections", and "political". All of these keywords correspond to the Indonesian presidential elections held on July 8th. It is clear that most of the influential bloggers considered the event to be significant enough to deviate from their normal activity and blog about this topic. On the other hand, when we analyzed Point 2 we found that the high frequency keywords during the period were "father" and "pin" which had no correlation with each other and which could not be associated with any specific event. This is in contrast to our earlier observations and highlights the need for further refinement in our approach to handle such points.

6 Conclusion and Future Work

In this paper we proposed a new approach to detect key topics of discussion in the blogosphere using the information from the activity of influential bloggers.

Our case study on the influential bloggers from the Indonesian blogosphere shows that it may be possible to detect "hot" topics in the blogosphere by observing the convergence of topics from the blog posts of influential bloggers. The preliminary results obtained from our study highlight the need for additional work to validate our methodology and approach. The detection of a suitable threshold which can be used to automatically extract the points of convergence of influential bloggers is another relevant area for future research.

7 Acknowledgements

This work was supported in part by the Office of Naval Research under the grant NR N00014-09-1-0165.

References

1. N. Agarwal, S. Kumar, H. Liu, and M. Woodward. Blogtrackers: A tool for sociologists to track and analyze blogosphere. In *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2009.
2. J. Allan. *Topic detection and tracking: event-based information organization.* Springer Publishers, 2002.
3. N. Bansal and N. Koudas. Blogscope: a system for online analysis of high volume text streams. In *33rd International Conference on Very Large Data Bases*, 2007.
4. R. L. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. volume 19, pages 370–383, Piscataway, NJ, USA, March 2007. IEEE Transactions on Knowledge and Data Engineering.
5. E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 482–490, New York, NY, USA, 2004. ACM.
6. S. Havre, B. Hetzler, and L. Nowell. Themeriver: visualizing theme changes over time. In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, volume 00, pages 115–123, Los Alamitos, CA, USA, 2000. IEEE Computer Society.
7. J. Kleinberg. Authoritative sources in a hyperlinked environment. In *9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.
8. S. Kumar, N. Agarwal, M. Lim, and H. Liu. Mapping socio-cultural dynamics in Indonesian blogosphere. In *3rd International Conference on Computational Cultural Dynamics*, 2009.
9. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
10. J. M. Schultz and M. Y. Liberman. Towards a "universal dictionary" for multi-language ir applications. Topic Detection and Tracking: Event-based Information Organization., 2002.
11. X. Song, Y. Chi, K. Hino, and B. Tseng. Identifying opinion leaders in the blogosphere. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 971–974, New York, NY, USA, 2007. ACM.