International Journal of Business and Technology

Volume 4
Issue 2 *Spring 2016*

Article 6

May 2016

Exploring the role of sentiments in identification of active and influential bloggers

Mohammad Alghobiri
Ibn Rushd College, maalghobiri@kku.edu.sa

Umer Ishfaq
COMSATS Institute of Information Technology, umer.bravo@gmail.com

Hikmat Ullah Khan
COMSATS Institute of Information Technology, hikmatullah@comsats.edu.pk

Tahir Afzal Malik
Ibn Rushd College, tahir.malik@ibnrushd.edu.sa

IJBTE ISSN (2223 – 8387)
Volume 4, Issue #2, Spring 2016

Exploring the role of sentiments in identification of active and influential bloggers

Mohammad Alghobiri, Umer Ishfaq, Hikmat Ullah Khan, Tahir Afzal Malik

Abstract. The social Web provides opportunities for the public to have social interactions and online discussions. A large number of online users using the social web sites create a high volume of data. This leads to the emergence of Big Data, which focuses on computational analysis of data to reveal patterns, and associations relating to human interactions. Such analyses have vast applications in various fields such as understanding human behaviors, studying culture influence, and promoting online marketing. The blogs are one of the social web channels that offer a way to discuss various topics. Finding the top bloggers has been a major research problem in the research domain of the social web and big data. Various models and metrics have been proposed to find important blog users in the blogosphere community. In this work, first find the sentiment of blog posts, then we find the active and influential bloggers. Then, we compute various measures to explore the correlation between the sentiment and active as well as bloggers who have impact on other bloggers in online communities. Data computed using the real world blog data reveal that the sentiment is an important factor and should be considered as a feature for finding top bloggers. Sentiment analysis helps to understand how it affects human behaviors.

Keywords: Blogger, Sentiment, Social web, Big Data.

M. Alghobiri[1], U. Ishfaq[2], H. Khan[3], T. Malik[4]

[1,2]Department of Management Information Systems, Ibn Rushd College Management Sciences, Abha, Kingdom of Saudi Arabia
[3,4]Department of Computer Science, COMSATS Institute of Information Technology, Attock, Pakistan

1. Introduction

The rise of Web 2.0 has turned online information consumers into information producers. Its interactive and dynamic features allow the development of numerous innovative web services. The Web 2.0 enables the masses to interact in social as well as collaborative environment also known as the social networks. Blogs are one of widely used form of social networks, where users share their views and experiences in the form of text, images or video forms [1]. Users get the advice of others before making decisions, for example, choosing a place for shopping or buying products of a particular brand. They listen and trust their opinions. In this way, they can be influenced by others whom they are connected with and such users are termed as influential bloggers. Identification of such users has been a real challenge [2].

Blogging has become a popular way of online interaction. People use it for voicing their opinions, reporting news or mobilizing political campaigns [3]. The bloggers usually create their interest groups and play a unique and significant role in social communities. The democratic nature of the blogosphere has made it a lucrative platform for social activities and influence propagation [4]. Finding top bloggers has direct applications in various fields, including online marketing and e-commerce. By gaining the trust of influential bloggers, companies can turn them into their allies and can save huge advertising sum [5]. Similarly, it helps in learning the market trends and to improve insights for better custom care. Such top bloggers who have affected others can further assist in reshaping the businesses and re-branding their products [6]. In this work, we explore the role of sentiments and check how the sentiments of blog content correlate with the overall activities of the top bloggers. We use the blog post content, compute their sentiment scores using SentiStrength, which is the widely used technique to compute the sentiment of the content. We then find the correlation of the sentiment with the characteristics of the top bloggers who are active as well as influential. In addition to some standard features, we propose features related to sentiment. The blog of Engadget has been used in this regard. In the following sections, first, we present the relevant work in section 2. Next, the proposed framework will be outlined where the research methodology is discussed in details in section 3. The evaluation results are discussed in section 4. Finally, conclusion and direction for further work will be presented in section 5.

2. Related Work

As blogging has gained considerable momentum in this era of time, people are highly influenced by their fellow bloggers' opinions. Researchers have highlighted various issues arising due to the plethora of information piling up on the web. Social media is believed to be behind this revolution. Such a development emphasizes the need of developing intelligent tools for mining valuable information [7]. Therefore, social media contains huge amounts of information about people, their customs and traditions. This valuable information can be exploited for better understanding individuals and their communities [8]. Studies show that influence has multiple dimensions and types. To differentiate one type from another and evaluate its importance over the other is real challenge [9]. PageRank is a well-known algorithm for ranking pages on the web. Zhou At el., [10] used PageRank by taking into account nodes on the social media network for finding the opinions of leaders. The authors in [11] identified the influential bloggers

https://knowledgecenter.ubt-uni.net/ijbte/vol4/iss2/6
DOI: 10.33107/ijbte.2016.4.2.06

by ranking the blogs. Various blogs on the blogosphere are compared and their relative importance is quantitatively measured. The work proposed in [12] developed a data mining based model for identifying the leading bloggers. It calculated the potential influence of influentials to maximize the spread of information over social media. Most of the people on social media act as information consumers. Research presented in [2] proposed a model which identifies the influential bloggers on the basis of information they produce rather than their popularity within the social community.
A new ranking metric has been proposed in [13], which critically analyze well-known blogger ranking algorithms, discussed their shortcomings. We found a standard model [14] which is based on four basic blogger's characteristics. It is novel in this regard that it identified influential bloggers. It compared them with active ones and concluded that both are important. After that it was compared with Pagerank as well and found that the model is better instead of Pagerank to find top bloggers. This work is an extension effort by the same authors who actually initiated this domain of finding top influential bloggers by introducing their model known as iIndex [15].
The most recent and comprehensive work that proposes a new metric, named as MIIB, finds the top influential in a blogosphere community using modular approach. It proposes three modules and then uses standard evaluation measures as well to show the significance of the proposed metric as well as its modules [16]. The current work is continuation of earlier works [17] [18] on modeling in the dynamic research domain of social web and semantic web.

3. Proposed Methodology

In this work, we propose the use sentiment feature. We use the existing features which are used to compute a blogger's activity as well as their effect within the community. The activity of bloggers is calculated by counting their number of blog posts. It is an effective method which represents the productivity of bloggers [4]. The recognition represents the position or status of a blogger in the social community. The number of urls linking to the blog post (inlinks) and comments on a blog post is taken as the recognition of the blogger. A large number of comments on a blog post show the interest of other bloggers as they write their comments [15]. It is a good feature to calculate the impact of a blogger on other bloggers as already considered by researchers in their works [4] [19] [15] [14].
Here, let us propose that adding a new feature of sentiment score, can increase the accuracy of the model as discussed in detail in Section. 4. We examine how the sentiment score of a blogger correlates with their overall activity as well as recognition. To the best of our knowledge, the sentiment feature has not been used so far. Table. 1 shows the features included in the proposed model.

Table 1. List of Features used in the paper

Sr	Feature Name
1	The number of blog posts initiated by a logger
2	The number of URLs linking to a blog posts (inlinks)
3	The number of comments received in a blog post
4	The sentiment score of a blogger (computed using SentiStrength)

3.1 Data Source

We perform our experiments on data of Engadget[1], a technology blog with real time features of a blog site. The characteristics of the dataset are shown in table 1.

Table 2. Dataset Characteristics

Characteristics	Engadget
The number of Bloggers	93
The number of Posts	63,358
The number of Inlinks	319,880
The number of Comments	3,672,819

3.2 SentiStrength

We compute the sentiment expressed by a blogger in his/her blog posts. We use a standard method for quantitative analysis of blog posts known as SentiStrength[2]. It is an opinion mining technique which identifies the sentiment associated with a word and assigns positive and negative scores. For that it uses a dictionary which categorizes the words into positive and negative just like a human being. The snapshot of the online tool of the SentiStrength is given in Figure 1.

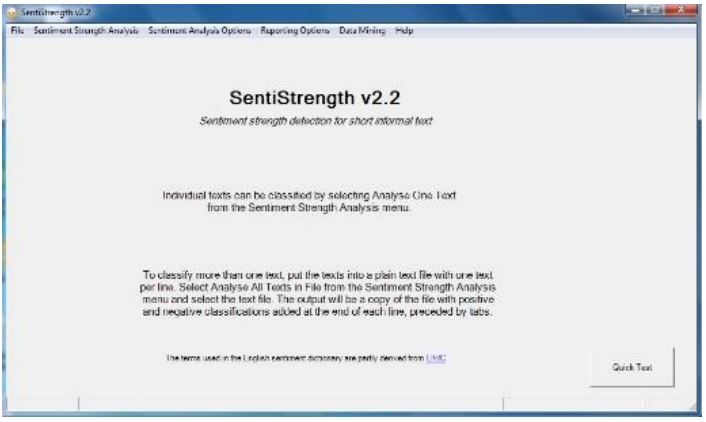

Figure. 1 Snapshot of SentiStrength v2.2 (Source: http://sentistrength.wlv.ac.uk)

3.3 Performance Evaluation Measures

Here, we discuss the both correlation measures that have been used for finding the correlation.

3.3.1 Spearman Correlation

A monotonic function is a function between ordered sets of two quantities. An increase in one quantity causes the increase or decrease in the other. Spearman correlation[3] uses monotonic function to measure the association or strength of relationship between the two variables. It is computed using the following equation:

[1] http://www.engadget.com
[2] http://sentistrength.wlv.ac.uk
[3] http://www.biostathandbook.com/spearman.html

https://knowledgecenter.ubt-uni.net/ijbte/vol4/iss2/6
DOI: 10.33107/ijbte.2016.4.2.06

$$\rho = 1 - \frac{6\Sigma d_i{}^2}{n(n^2-1)} \quad (1)$$

Where $d_i = x_i - y_i$ is the rank difference between the two ordered sets x and y which are obtained by converting the variables X and Y into ranks. n represents the sample size of the data.

3.3.2 Kendall Correlation

It is a non-parametric test to measure the dependence of two quantities upon one another. The following equation is used to compute the dependence between the two quantities:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \quad (2)$$

Where n_c and n_d are number of concordant and number discordant pairs respectively.

4. Results Discussion

In this section, we examine how the sentiment score correlates with activity (number of posts) and recognition (inlinks and comments) of a blogger. We perform this experiment on data of Engadget blog. Table. 3 presents the results of our findings. The last three columns show the sentiment score, activity and recognition score of the top-10 bloggers of Engadget dataset. As discussed earlier, the blog posts of a blogger represent his/her activity. The number of inlinks and comments on a post together constitute the recognition of a blogger. Sentiment score, on the other hand, is calculated using SentiStrength, which calculates the sentiment score of content, the blog posts in this case, using a mining based scaling system. A post has words with positive, negative, or neutral sentiment. These words are analyzed on a scaling system and assigned positive and negative values. The system assigns values from a dictionary which interprets the positivity or negativity with human level accuracy [20].

Table 3. Blogger ranking based on all the features

Rank	blog posts	Inlinks	Comments	Activity	Recognition	Sentiment
1	D. Murph	D. Murph	D. Murph	D. Murph	L. June	D. Murph
2	P. Rojas	T. Ricker	R. Block	R. Block	D. Murph	R. Block
3	R. Block	P. Miller	L. June	P. Miller	T. Ricker	P. Miller
4	P. Miller	N. Patel	J. Topolsky	P. Rojas	P. Miller	P. Rojas
5	D. Melanson	R. Block	P. Miller	D. Melanson	N. Patel	D. Melanson
6	T. Ricker	D. Melanson	N. Patel	T. Ricker	J. Topolsky	T. Ricker
7	N. Patel	J. Topolsky	T. Ricker	N. Patel	R. Block	N. Patel
8	E. Blass	C. Ziegler	D. Melanson	J. Topolsky	D. Melanson	E. Blass
9	J. Topolsky	R. Miller	C. Ziegler	C. Ziegler	C. Ziegler	J. Topolsky
10	C. Ziegler	V. Savov	P. Rojas	E. Blass	R. Miller	C. Ziegler

Table 4. displays the correlation analysis of the proposed features with the two performance evaluation measures of Spearman and Kendall correlation. The table shows the results of the features. The comparison indicates higher values using Spearman equation. However, Kendall correlation has smaller values for the comparing features. This is because Spearman's correlation checks the difference between the ranking orders of the two quantities whereas Kendall correlation analyzes all the comparing quantities and tries to quantify the difference between the percentage of the concordant pairs and the discordant pairs.

Table 4. The Correlation Analysis of the proposed features

	Spearman Correlation	**Kendall Correlation**
Activity vs. Recognition	0.536	0.328
Activity vs. Sentiment	0.993	0.911
Recognition vs. Sentiment	0.485	0.401

Conclusion

In this paper, in addition to activity and influence, we also focus on the sentiment analysis of blog posts. We use the blog post content, compute their sentiment scores using standard methods. Then we correlate them with the characteristics of the top bloggers who are active as well as have influence on other bloggers within the social network. Several features have been used and analysis has been carried out on data of Engadget blog. The correlation between the sentiment and activity is very high which show that the sentiment is an important factor for active bloggers but it is relatively low for recognition part which suggest that sentiment does not have strong correlation with sentiment. Sentiment is an important characteristic in social web and blog posts are not an exception. It helps in analysis of overall community and human behaviors. For future work, we intend to include naming disambiguation techniques proposed in [21] to check whether naming ambiguity plays role to find the actual user or not. Other potential future work can be to find the top bloggers in a blogging community using social network analysis metrics which are graph based measures.

References

1 Kale , A. Karandikar, P. Kolari, A. Java, T. Finin and A. Joshi, "Modeling trust and influence in the blogosphere using link polarity," in International Conference on Weblogs and Social Media (ICWSM), Boulder, 2007.

2 D. M. Romero, W. Galuba, S. Asur and B. A. Huberman, "Influence and passivity in social media," in 20th international conference companion on World wide web, Hyderabad, 2011.

3 Johnston, B. Friedman and S. Peach, "Standpoint in Political Blogs: Voice, Authority, and Issues," Women's Studies: An inter-disciplinary journal, vol. 40, no. 3, pp. 269-298, 2011.

4 N. Agarwal, D. Mahata and H. Liu, "Time- and Event-Driven Modeling of Blogger Influence," in Encyclopedia of Social Network Analysis and Mining (ESNAM), New York, Springer, 2014, pp. 2154-2165.
5 Sun and V. T. Ng, "Identifying influential users by their postings in social networks," in 3rd international workshop on Modeling social media, Milwaukee, 2012.
6 G. Mishne and M. d. Rijke, "Deriving wishlists from blogs show us your blog, and we'll tell you what books to buy," in 15h International conference on World Wide Web, Edinburgh, 2006.
7 N. Agarwal and H. Liu, "Blogosphere: research issues, tools, and applications," ACM SIGKDD Explorations Newsletter, vol. 10, no. 1, 2008.
8 S. Kumar , N. Agarwal , M. Lim and H. Liu, "Mapping socio-cultural dynamics in Indonesian blogosphere," in Third International Conference on Computational Cultural Dynamics, Washington DC, 2009.
9 J. Tang, J. Sun, C. Wang and Z. Yang, "Social influence analysis in large-scale networks," in 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, 2009.
10 H. Zhou, D. Zeng and C. Zhang, "Finding leaders from opinion networks," in IEEE International Conference on Intelligence and Security Informatics (ISI), Dallas, 2009.
11 X. Song , Y. Chi, K. Hino and B. Tseng, "Identifying opinion leaders in the blogosphere," in 6th ACM conference on Conference on information and knowledge management, Lisbon, 2007.
12 Y. Singer, "How to win friends and influence people, truthfully: influence maximization mechanisms for social networks," in Fifth ACM international conference on Web search and data mining (WSDM), Seattle, 2012.
13 J. Bross, K. Richly, M. Kohnen and C. Meinel, "Identifying the top-dogs of the blogosphere," Social Network Analysis and Mining, vol. 2, no. 2, pp. 53-67, 2012.
14 N. Agarwal, H. Liu, L. Tang and P. Yu, "Identifying the influential bloggers in a community," in in: Proceedings of the International Conference on Web Search and Web Data Mining, New York, 2008.
15 N. Agarwal, H. Liu, L. Tang and P. S. Yu, "Modeling Blogger Influence in a community," Social Network Analysis and Mining, vol. 2, no. 2, pp. 139-162, 2012.
16 H. U. Khan, A. Daud and T. A. Malik, "MIIB: A Metric to Identify Top Influential Bloggers in a Community," PLoS One, vol. 10, no. 9, pp. 1-15, 2015.
17 H. U. Khan and T. A. Malik, "Finding Resources from Middle of RDF Graph and at Sub-Query Level in Suffix Array Based RDF Indexing Using RDQL Queries," International Journal of Comuter Theory and Engineering, vol. 4, no. 3, pp. 369-372, 2012.

18 T. A. Malik, H. U. Khan and S. Sadiq, "Dynamic Time Table Generation Conforming Constraints a Novel Approach," in International Conference on Computing and Information Technology (ICCIT), Al-Madinah Al-Munawwarah, 2012.
19 L. Akritidis, D. Katsaros and P. Bozanis, "Identifying the Productive and Influential Bloggers in a Community," IEEETransaction on System, Man, Cybernetics, Part C, vol. 41, no. 5, pp. 759 - 764, 2011.
20 M. Thelwall, K. Buckley, G. Paltoglou, D. Cai and A. Kappas, "Sentiment strength detection in short informal text," Journal of the American Society for Information Science and Technology, vol. 61 , no. 12, p. 2544–2558, 2010.
21 M. Shoaib, A. Daud and M. S. H. Khayal, "Improving Similarity Measures for Publications with Special Focus on Author Name Disambiguation," Arabian Journal for Science and Engineering, vol. 40, no. 6, pp. 1591-1605, 2015.