PTMIB: Profiling Top Most Influential Blogger using Content Based Data Mining Approach.

Vasanthakumar G U, Priyanka R, Vanitha Raj K C, Bhavani S, Asha Rani B R, P Deepa Shenoy, Venugopal K R

Department of Computer Science and Engineering,
University Visvesvaraya College of Engineering, Bangalore, India.
E-mail: vasanthakumar.gu.in@ieee.org

Abstract—**Online Social Network (OSN) provides fastest way to communicate and spread information, influencing users in the network. Blog sites allow the users to reflect and share opinions on various topics of discussion in the form of blogs/online journals and letting readers to comment on their blogs/posts. In this work, a novel method to profile Top Most Influential Blogger (TMIB) is proposed based on content analysis. Contents of blog documents of bloggers under consideration in the blog network are compared and analyzed. Term Frequency and Inverse Document Frequency (TF-IDF) of two blog documents are obtained at a given point of time to get the Cosine Similarity score between those documents. The Influence Scores (IS) of bloggers under conflict are computed. The simulation results demonstrates that the proposed Profiling Top Most Influential Blogger (PTMIB) algorithm is adequately accurate in determining the top most influential blogger at any instant of time considered.**

Keywords: **Blog Document; Content Analysis; Cosine Similarity; Influential Blogger; Online Social Network; Profiling.**

I. Introduction

People from all over the world are catching up with the Online Social Networking sites 24/7 for various purposes. Blog sites are the impressive ones in social networking which are used by numerous users not just for communicating purposes but also to propogandize. Unlike other online social networking sites, blog sites are primarily not used either to share photos, videos or short messages, instead these sites allow users to create a detailed content on any topic of their interest. Each blog is a document containing data either with or without images along with a heading created by its user. Documents of one user can be accessed by other users in the network. The users of blog sites are called as Bloggers and can perform activities like comment, trackback, scrap and bookmark etc on others documents. Blogger.com, myspace.com etc are some of the examples of blog sites.

In recent days, generally people depend on data available in the internet to gather information or to clarify any confusions relating to any matter. Blog documents are one such source where data will be available created by users themself or any information gathered from others. Each blog document contains a heading or the topic which will be described further. The description can be the opinion/knowledge/experience of that user on that topic or it can be the collected data from blog document(s) of other users. Similar content in the documents of two users depict that those two users are of same thinking and interest. Based on this logic, users form groups and share all documents amongst themselves and publicly, where all users in that network can perform activities on it.

The activities of users on the document(s) of other users contribute in increasing the influential power of that document(s) and such activities by other users may occur depending on the content of that document. Hence, content of a document plays a vital role in increasing its own influential power. The influential power of all documents of a user will be the influential power of that user in the network. Finding such Influential Blogger by performing content analysis of their documents with respect to that of others in the network is a challenging task. OSN data is analyzed through dynamic mining [1] [2] for various purposes.

Motivation: Influential Bloggers play prominently as role models for other follower bloggers in the network. Some information can be spread rapidly through these Influential Bloggers. These Influential Bloggers are used positively for many advantages in the field of advertising, e-commerce, e-business etc., but they are as well being used negatively in many situations for propogandizing antisocial activities, which motivated us to carry out this work.

Contribution: In this paper, we have presented an algorithm PTMIB to profile the top most influential blogger based on the contents of blog documents. The contents of the blog network documents are analyzed by using tf-idf and obtaining the cosine similarity score, during conflicts in determining the Influential Blogger.

The rest of the paper is organized as below: Section-II gives a survey of literature. Background work is briefly discussed in Section-III. The problem is defined in Section-IV where as in Section-V the proposed system is discussed in detail. Profiling Top Most Influential Blogger (PTMIB) algorithm is presented in Section-VI. Simulation and Performance Analysis are discussed in Section-VII presenting Conclusions in Section-VIII.

978-1-5090-1281-7/16/$31.00 ©2016 IEEE

II. Literature Survey

Leonidas Akritidis et al. [3] have proposed solution for the problem of identifying the bloggers who are both influential and productive in a community blog site. As part of solution, they proposed two matrices considering the number of incoming links. Bloggers were characterized as recently influential or recently productive, or both, or none. Experiments conducted on En-gadget data set reveal significant temporal patterns in the behavior of blogging and its activity.

In paper [4], authors have proposed and developed an effective model MASS to mine and identify top-k influential bloggers based on varied factors like their domains of interest, comments and authority in page link network. The effectiveness of MASS is demonstrated empirically and proved that it can be used in multiple application scenarios.

The paper [5] presented an important groundwork to differentiate influential and popular bloggers. They considered weighting readers as a measure to identify influential bloggers. They also developed Quantifying Influence Model (QIM) by an interdisciplinary approach. The developed method promises in laying foundation stone of a qualitative approach while determining influential bloggers.

Based on implicit links between bloggers, Chang Sun et al. [6] have proposed SimRank algorithm to recommend influential bloggers in the network. Measuring text similarity of blog posts, they formed the linked graph of bloggers and also adopted PageRank algorithm to rank the importance of bloggers. Experimental results indicated the effectiveness of algorithm in recommending bloggers.

Mohammad Alodadi et al. [7] have proposed a model that creates the similarity metric over the term frequency-inverse document frequency (TF-IDF). There was open requirement from ICHI 2015 a healthcare domain concerning about health inquires in social media. The challenge they faced was to reduce the repetition of post for patient support forums. The experimental results of the model using cosine similarity and TF-IDF were proved improvised over existing models. Emily Hill et al. [8] have conducted an overall qualitative study of stemmers on software domain, specifically Java source code. For this purpose they used MAP and Rank measure to check the retrieval effectiveness and impact of stemming by conducting query-by-query quantitative study.

Mohamed H Haggag [9] proposed a model to extract keywords based on their relatedness weight among the entire text terms. Here semantic similarity is used to find the strength of terms relationship. Based on meaning, documents were assigned weights and most correlated terms were provisioned in both frequency and weight likeness. Keywords were recursively evaluated according to their cohesion to each other and to the document context. Key concepts used in model were Semantic Relatedness, Word pair similarity, Similarity Score Normalization, Average similarity and Keywords list stability. Experimental results proved that the model achieved enhanced precision and recall extraction values over other approaches in evaluating semantic relationships between individual words and in overall similarity.

Authors of [10] have presented a survey paper describing few stemming algorithms used for information retrieval. Each algorithm is explained with its advantages and drawbacks and also the detailed assessment of current status of stemming process with its historical evolution is given. S Megala et al. [11] proposed a TWIG stemming algorithm which produced meaningful stems. This algorithm reduced the error rate and this measure is the one used to validate its performance. The proportion of unmeaningful words to meaningful words provides its error rate. Experiments were conducted on Alan Beale's Core Vocabulary Dictionary data having 21,877 words. Experimental results proved that error rate for TWIG = 1.5percent which is a bare minimum one when compared with other stemming algorithms like STANS = 7.6percent and Porters Stemming Algorithm = 39.9percent.

The authors of [12] contributed a hybrid method combining the strength of dense distTributed representations as opposed to sparse term matching with that of tf-idf based methods to automatically reduce the impact of less informative terms. A novel system for analyzing temporal changes in the activities and interests of bloggers through a 3D visualization of phrase dependency structures in sentences is proposed by the authors of [13].

The relationship between bloggers and posts is modeled as a bipartite graph in [14] and a weighted link-based rank approach is proposed. The authors of [15] performed behavioral analysis of bloggers and topics to discover closer-knit groups in the social network, leveraging the semantically enriched links. Authors of [16] proposed Parato/NBD model for forecasting the bloggers behaviour using number of posts and their activeness. An evaluation model that describes the effects of explosive communication by microblogging that have a substantial impact on personal and public fields is proposed in [17].

III. Background Work

Seung-Hwan Lim et al. [18] proposed an approach and proved that the blogger with highest User Content Power (UCP) will be the influential blogger in the network. UCP is computed by summing-up the Document Content Power (DCP) of all documents of a blogger. Whereas DCP of a document is calculated by the number of times activities performed on that document of a blogger by other bloggers in the network. A document of a blogger may have highest activities either from same blogger or set of bloggers, which adds-up increasing the DCP of that document, which in turn increases UCP of that blogger, leading to wrongly identifying such blogger as Influential Blogger (IB) in the network.

As a solution to the said issue, PIB algorithm [19] illustrates that a blogger who influences highest number of unique bloggers in the network as the IB. The authors of the algorithm have considered that if any activity, either trackback or scrap or bookmark appears for a document of one blogger (first blogger) from a document of other blogger, and then any further activities between the documents of those two bloggers will not be considered to calculate the influential power of first blogger. This results in obtaining the unique bloggers who have got influenced by that first blogger. The comment activity is omitted while calculating the influential power of bloggers, since comment does not necessarily prove that the document has influenced the other blogger. They illustrated that the blogger who has highest Influential Blog Power (IBP) i.e., who has influenced highest number of unique bloggers as IB.

IV. Problem Definition

There exists a conflict that, at any given point of time, there can be more than one blogger with the same IBP value. To solve the conflict between such bloggers, we in this work propose a new method to determine and profile the Top Most Influential Blogger amongst them.

In order to prove that a document has influenced other bloggers in the network, there needs to be some activity between that document and a document of other blogger. By considering the activities performed on the blog documents, we can just infer that there exists some influence between them, but to find out to what level or extent the document has influenced the other blogger is the challenge which we have made an attempt to address in this work.

V. Proposed System

A. Proposed Method

In order to find the level or percentage of influence that a particular document has made on others, we analyze the contents of blog documents in its network, performed out of trackback and scrap activities. Contents of both the documents under consideration are analyzed and checked for content similarity between them using TF-IDF and Cosine Similarity approach.

TF-IDF: Term Frequency (TF) measure refers to the number of times a particular word under consideration appears in a given document. As documents may vary in size, it is therefore necessary to normalize the documents. The simplest way is to divide the term frequency by the total number of terms in the document.

For example, if a term appears 20 times in a document and if the total number of terms are 100, then the normalized term frequency considering all the terms to be of equal importance i.e 20/100=0.2. But in reality, few terms appearing frequently have little power to determine the relevance. Such terms need to be weighed down and few other terms which appear less frequently having more power to determine the relevance need to be weighed up.

Logarithms are used to solve this with Inverse Document Frequency (IDF) measure. IDF of a term is computed by adding one to the logarithm of total number of documents divided by the number of documents with that particular term. For example if there are 5 documents and the term appears in 3 documents, then its IDF is 1+log(5/3)=1.510825623.

Cosine Similarity: To find the similarity between two documents in their vector space, we first consider the magnitude of the vector differences between the two documents. The drawback of this measure is that, the two documents with similar content can have significant vector differences just because one document is much bigger than the other. Therefore to compensate the effect of document length, we find the similarity between two documents by computing the cosine similarity of their vector representations.

The Cosine Similarity score of two documents is given by the formula:

$$\text{Cosine Similarity(d1,d2)} = \frac{\text{Dot Product(d1,d2)}}{||d1|| * ||d2||} \tag{1}$$

Where

$$\text{Dot Product (d1,d2)} = d1[0] * d2[0] + d1[1] * d2[1] + ... + d1[n] * d2[n]$$

$$||d1|| = \text{Square Root}(d1[0]^2 + d1[1]^2 + ... + d1[n]^2)$$

$$||d2|| = \text{Square Root}(d2[0]^2 + d2[1]^2 + ... + d2[n]^2)$$

B. System Architecture

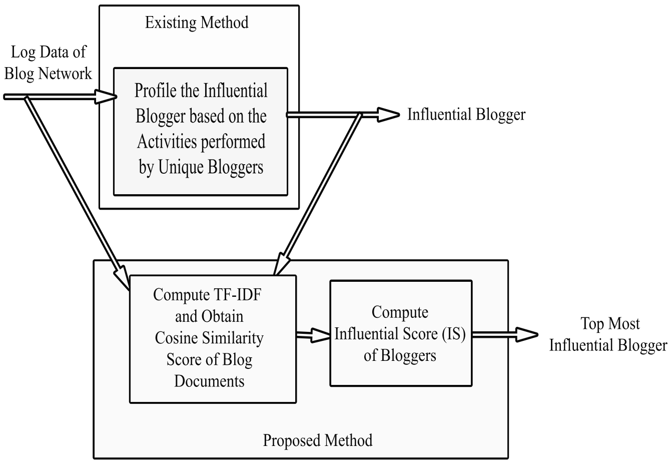
Existing Method
Log Data of Blog Network
Profile the Influential Blogger based on the Activities performed by Unique Bloggers
Influential Blogger
Compute TF-IDF and Obtain Cosine Similarity Score of Blog Documents
Compute Influential Score (IS) of Bloggers
Top Most Influential Blogger
Proposed Method

Fig. 1: System Architecture

As shown in Figure-1, the log data of blog network is analyzed to profile the influential blogger based on activities performed by unique bloggers [19]. If there more than one influential blogger identified, then those under conflict are considered for further analysis in our proposed method and by computing the Influence Scores (IS) of those bloggers, the top most influential blogger is determined.

The Influence Score of a blogger is given by the formula:

$$IS = \frac{\sum \text{Cosine Similarity Scores}}{\text{Number of Bloggers Influenced}} \tag{2}$$

C. Scenario Illustration

Table-1 shows a general scenario of PTMIB approach illustration. It is observed that the blogger B1 has 3 documents, whereas blogger B2 has 2 documents. Blogger B1s document B1D1 has influenced 1 unique blogger, B1D2 and B1D3 have influenced 2 and 1 unique bloggers respectively.

Consider the document B1D1 of blogger B1 along with its child document B5D3 as per blog documents network shown in Figure-2 and find the keywords in each document by stemming the words. Then the stemmed keywords of these two documents are processed through TF-IDF approach to obtain the cosine similarity score of 0.40 between them.

Then the same procedure is repeated with B1D2 of blogger B1 which has B4D2 and B3D1 as its child documents to get 0.88 and 0.89 as similarity scores respectively. Continuing the same process, the similarity score between B1D1 and B1D3 is 0.51 and by summing-up the similarity scores of all the documents and by dividing it with the number of documents, compute the Influence Score (IS) of Blogger B1 to be 0.67.

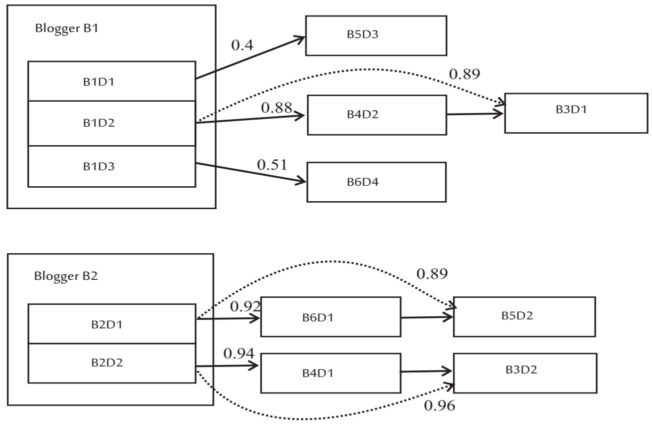

Fig. 2: Network of Blog Documents

Likewise, blogger B2s B2D1 and B2D2 documents have influenced 2 unique bloggers individually. After getting the stemmed words of root documents and its respective child documents, obtain the similarity score of B2D1 with B6D1 and B5D2 as 0.92 and 0.89 respectively. The same step is repeated with B2D2 to get the similarity scores 0.94 and 0.96 for B4D1 and B3D2 respectively. Then the Influence Score calculated for B2 is 0.93.

As already mentioned, blogger with highest Influence Score will be the TMIB of the network, 0.93 is greater than 0.67 and hence blogger B2 will be the TMIB of the network at that given point of time and thus resolving the conflict by accurately identifying the TMIB in the network.

VI. Algorithm

The steps followed in profiling the top most influential blogger is clearly shown in Algorithm-1.

Algorithm 1 *Profiling Top Most Influential Blogger (PTMIB) Algorithm*

```
while True do
    for "Every Blogger under conflict of Influentiality" do
        for "Every Document of the Blogger, considering it as
          Root document" do
            Perform Stemming
            Compute TF-IDF
            Obtain Activity-based Blog Document Network
            for "Every other document in Activity-based Blog
              Document Network, except that of Root" do
                Perform Stemming
                Compute TF-IDF
                Compute Cosine Similarity Score with respect to
                that of Root
            end for
        end for
        Compute the Influence Score (IS) of the Blogger
    end for
    Compute Max(IS)
    Profile the Top Most Influential Blogger of the network
      having Max(IS)
end while
```

VII. Simulation and Performance Analysis

Using Java and XML, we developed a social blogging site and allowed the public to register and use it for 6 months. Around 300 users registered themselves as bloggers and about 50 bloggers posted blog documents. Roughly around 75 blog documents and nearly 1000 activities got generated exponentially over a short span of time and were logged. Using MySQL, the logged data is stored for further analysis in our proposed PTMIB algorithm.

The collected data is analyzed by extracting the number of activities on each and every document of bloggers and computed influential blogger IB according to UCP method [18]. From the same dataset, we extracted the number of unique bloggers who were influenced from the documents posted by other bloggers in the network and then computed the Influential Blogger (IB) of the network according to [19]. Later during some period of our analysis, we found that there exists more than one blogger as IB according to both [18] and [19], leading to conflict in profiling the Influential Blogger of the network.

TABLE I: PTMIB Approach Illustration

Bloggers who Influenced same number of Unique bloggers	**Number of Documents**	**Document ID (Root)**	**Number of Bloggers Influenced**	**Blog Documents Network (Child)**	**Similarity Score between Child and Root Documents**	**Influence Score (IS) of the Blogger**
B1	3	B1D1 B1D2 B1D3	 2 1	B5D3 B4D2, B3D1 B6D4	0.40 0.88, 0.89 0.51	0.67
B2	2	B2D1 B2D2	2 2	B6D1, B5D2 B4D1, B3D2	0.92, 0.89 0.94, 0.96	0.93

We extended our analysis further and computed tf-idf and obtained cosine similarities of blog documents network of those bloggers under conflict. We then computed Influence Scores (IS) of those bloggers under conflict and finally profiled the Top Most Influential Blogger (TMIB) of the network having Max(IS) value as per our PTMIB algorithm. The results of our simulation are as illustrated.

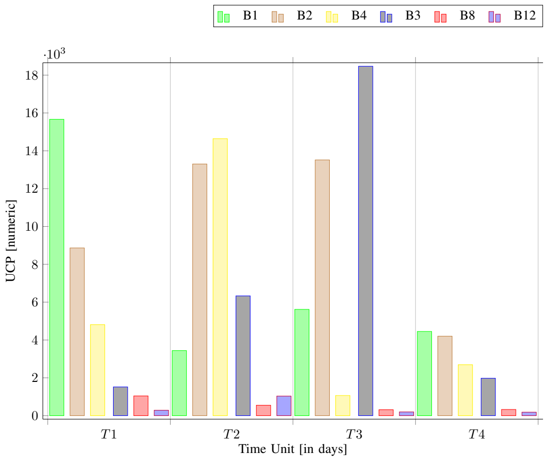

Fig. 3: User Content Power of six Top Bloggers over Time

Figure-3 shows UCP of each blogger in different time units and is observed that the blogger with highest UCP as influential, varying over different time units as according to [18].

Figure-4 depicts the IBP with respect to Time Unit. As according to the PIB algorithm [19], at T1, the graphs shows that blogger B8 has Max(IBP) value amongst all others, hence being the Influential Blogger during that time unit. Whereas in T2, B8 and B12 both have same Max(IBP) value, indicating conflict to accurately identify the influential blogger during that time unit. To resolve these kinds of conflicts, our PTMIB algorithm is proposed to determine and profile the top most influential blogger.

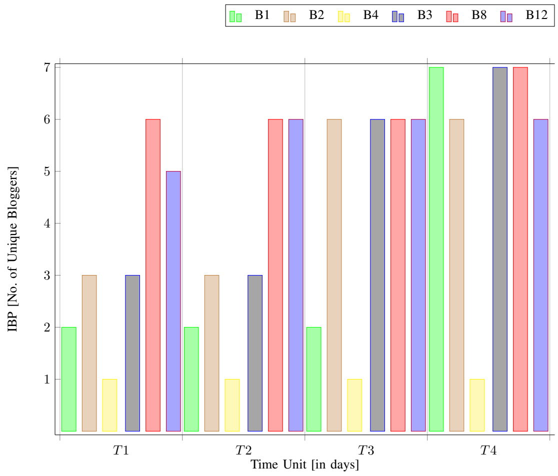

Fig. 4: Influential Blog Power of six Top Bloggers over Time

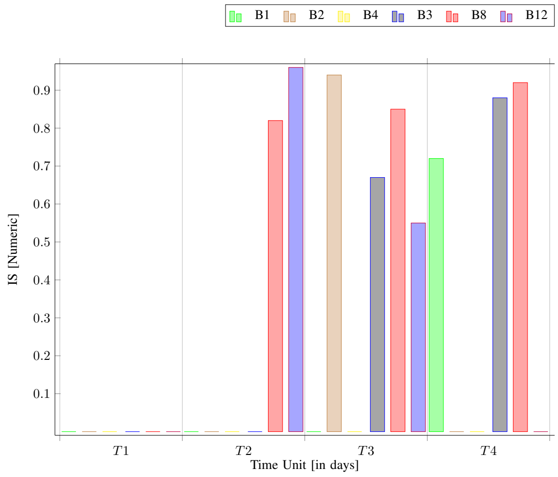

Fig. 5: Influence Scores of Bloggers having same Max(IBP) over Time

Figure-5 shows the simulation results of proposed algorithm. As depicted in the figure, during T2, B12 has IS=0.96 and hence resolving the conflicts efficiently. Similar kind of situation occurs during time units T3 and T4, where four bloggers and three bloggers respectively have same Max(IBP) values. To resolve the conflicts, we followed the same procedure to find IS of those bloggers and accurately profile the Top Influential Blogger during time units where conflicts arise.

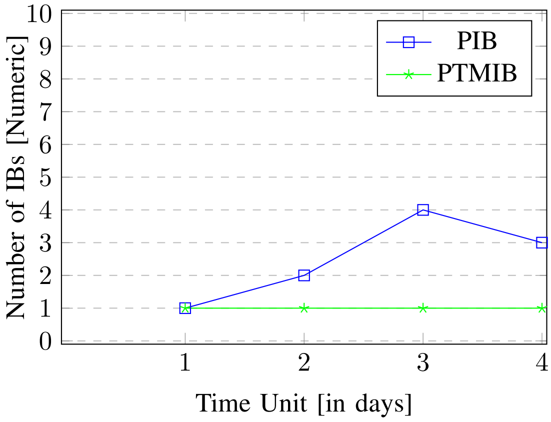

Fig. 6: Number of Influential Bloggers over Time

Comparing the proposed approach with the previous PIB [19], Figure-6 depicts the number of Influential Bloggers (IBs) in each time unit. Conflicts arise while identifying the top most influential blogger with the previous approach, but is over come by using proposed PTMIB approach giving rise to only one Top Most Influential Blogger in the network during any time unit of analysis.

VIII. Conclusions

The paper presents PTMIB algorithm for profiling the top most influential blogger. Methods adopted in the literature for profiling the influential blogger were not only based on number of activities performed on the blog documents [18], but also on the number of unique bloggers influenced by the blog documents of other bloggers [19] [20]. In this work, we have analyzed the contents of blog documents during conflicts in determining the influential blogger. It is illustrated and evaluated that there exist only one top most influential blogger at any instance of our analysis. The results after simulation denostrate the adequecy and accuracy of our proposed PTMIB algorithm.

The PTMIB algorithm helps in determining the information diffusion of criminal activities and also identifies the head of their group when applied on the blog network of criminals. This approach can be used for targeted marketing in product based business enterprises. Open avenues for future work are in analysing the content of documents semantically.

References

[1] P Deepa Shenoy, Srinivasa K G, Venugopal K R and Lalit M Patnaik, "Evolutionary approach for mining association rules on dynamic databases," *Advances in Knowledge Discovery and Data Mining*, pp. 325–336, April 2003.

[2] P Deepa Shenoy, Srinivasa K G, Venugopal K R and Lalit M Patnaik, "Dynamic association rule mining using genetic algorithms," *Intelligent Data Analysis*, vol. 9, no. 5, pp. 439–453, September 2005.

[3] Leonidas Akritidis, Dimitrios Katsaros and Panayiotis Bozanis, "Identifying the productive and influential bloggers in a community," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 41, no. 5, pp. 759–764, 2011.

[4] Yichuan Cai and Yi Chen, "Mass: a multi-facet domain-specific influential blogger mining system," *2010 IEEE 26th International Conference on Data Engineering (ICDE)*, pp. 1109–1112, 2010.

[5] Eunyoung Moon and Sangki Han, "A qualitative method to find influencers using similarity-based approach in the blogosphere," *International Journal of Social Computing and Cyber-Physical Systems*, vol. 1, no. 1, pp. 56–78, 2011.

[6] Chang Sun, Bing-quan Liu, Cheng-jie Sun, De-Yuan Zhang and Xiaolong Wang, "Simrank: A link analysis based blogger recommendation algorithm using text similarity," *2010 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 6, pp. 3368–3373, 2010.

[7] Mohammad Alodadi and Vandana P Janeja, "Similarity in patient support forums using tf-idf and cosine similarity metrics," *2015 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 521–522, 2015.

[8] Emily Hill, Shivani Rao and Avinash Kak, "On the use of stemming for concern location and bug localization in java," *2012 IEEE 12th International Working Conference on Source Code Analysis and Manipulation (SCAM)*, pp. 184–193, 2012.

[9] Mohamed H Haggag, "Keyword extraction using semantic analysis," *International Journal of Computer Applications*, vol. 61, no. 1, 2013.

[10] Cristian Moral, Angélica de Antonio, Ricardo Imbert, and Jaime Ramírez, "A survey of stemming algorithms in information retrieval." *Information Research: An International Electronic Journal*, vol. 19, no. 1, p. 1, 2014.

[11] S Megala, A Kavitha and A Marimuthu, "Improvised stemming algorithm–twig," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 7, pp. 168–171, 2013.

[12] De Boom, Cedric, Steven Van Canneyt, Steven Bohez, Thomas Demeester and Bart Dhoedt, "Learning Semantic Similarity for Very Short Texts," *IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 1229-1234, November 2015.

[13] Itoh, Masahiko, Naoki Yoshinaga, Masashi Toyoda and Masaru Kitsuregawa, "Analysis and visualization of temporal changes in bloggers' activities and interests," *Visualization Symposium (PacificVis)*, pp. 57-64, February 2012.

[14] Lu and Fuxi Zhu, "Discovering the important bloggers in blogspace," *IEEE International Conference on Artificial Intelligence and Education (ICAIE)*, pp. 151-154, October 2010.

[15] Macskassy and Sofus A, "Leveraging Contextual Information to Explore Posting and Linking Behaviors of Bloggers," *IEEE International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 64-71, August 2010.

[16] Rui, Cai, Qi Jia-yin and Wang Mian, "Forecasting bloggers' online behavior based on improved Pareto/NBD model," *IEEE International Conference on Management Science and Engineering (ICMSE)*, pp. 84-90, July 2013.

[17] Zhang, Yuan and Yuqian Bai, "Research on the Influence of Micro-bloggers, Take Sina Celebrity Micro-blog as an Example," *IEEE Eighth International Conference on Semantics, Knowledge and Grids (SKG)*, pp. 189-192, October 2012.

[18] Seung-Hwan Lim, Sang-Wook Kim, Sunju Park and Joon Ho Lee, "Determining content power users in a blog network: an approach and its applications," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 41, no. 5, pp. 853–862, 2011.

[19] Vasanthakumar G U, Bagul Prajakta, P Deepa Shenoy, Venugopal K R and Lalit M Patnaik, "PIB: Profiling Influential Blogger in Online Social Networks, A Knowledge Driven Data Mining Approach," *Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015), Procedia Computer Science, Elsevier B.V.*, vol. 54, pp. 362–370, August 2015.

[20] Vasanthakumar G U, P Deepa Shenoy and Venugopal K R, "PTIB: Profiling Top Influential Blogger in Online Social Networks," *International Journal of Information Processing (IJIP-2016), IK International Publishing*, vol. 10, no. 1, pp. 77–91, June 2016.