DOI: 10.21917/ijsc.2017.0195

PTMIBSS: PROFILING TOP MOST INFLUENTIAL BLOGGER USING SYNONYM SUBSTITUTION APPROACH

Vasanthakumar G.U[1], Priyanka R, Vanitha Raj K.C, Bhavani S, Asha Rani B.R, P. Deepa Shenoy and Venugopal K. R.

Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, India

E-mail: [1]vasanthakumar.gu.in@ieee.org

Abstract

Users of Online Social Network (OSN) communicate with each other, exchange information and spread rapidly influencing others in the network for taking various decisions. Blog sites allow their users to create and publish thoughts on various topics of their interest in the form of blogs/blog documents, catching the attention and letting readers to perform various activities on them. Based on the content of the blog documents posted by the user, they become popular. In this work, a novel method to profile Top Most Influential Blogger (TMIB) is proposed based on content analysis. Content of blog documents of bloggers under consideration in the blog network are compared and analyzed. Term Frequency and Inverse Document Frequency (TF-IDF) of blog documents under consideration are obtained and their Cosine Similarity score is computed. Synonyms are substituted against those unmatched keywords if the Cosine Similarity score so computed is below the threshold and an improved Cosine Similarity score of those documents under consideration is obtained. Computing the Influence Score after Synonym substitution (ISaS) of those bloggers under conflict, the top most influential blogger is profiled. The simulation results demonstrate that the proposed Profiling Top Most Influential Blogger using Synonym Substitution (PTMIBSS) algorithm is adequately accurate in determining the top most influential blogger at any instant of time considered.

Keywords:

Blog Document; Content Analysis; Cosine Similarity Score; Influential Blogger; Profiling.

1. INTRODUCTION

Internet based application Web 2.0 provides services to online social networking sites. Most of the social networking sites provide means for its users to communicate or integrate with other users situated far apart physically. Presently, social networking sites facilitating online social networks have incorporated smart communication tools such as blogging, mobile connectivity, location sharing, photo/audio/video sharing and even audio/video calling. There are around 300 different sites which provide several similar and a few different types of services/features, in which users create their profile, specific to the service and use it. These user profiles are organized and maintained by social networking site organizations. Online community services are group centered, unlike user centric online social networks.

Today undoubtedly, online social networks have tremendously occupied the lives of people, especially younger generation, in all the matters. India has got world’s largest social media users and most of them in the mid age group, who highly depend on social networking sites for decision making on product purchasing, hotel booking, travel, movies and so on [1]. People today are so interdependent in social networks that they believe developing a vital network with others as an asset, supplementing people to fulfill their needs this way is adversely affecting the human day-to-day lives, relationships and even families.

In recent days online social network has gained huge attention from businesses point of view, where industries consider knowing their audience as the key to their success. Since to remain competitive and noticed in the market is a need for industries and it has made them to engage in social networking which involves all age group people from various sectors. Many companies provide various tools for their customers to engage and discuss about their experience on their products, creating interactive community and get socialized with visitors as well as with their customers. OSN data is analyzed through dynamic mining [2] [3] for various purposes.

Blog/Blogging is one of those online social networking sites available, with comparatively varied features. Blog sites are maintained and managed by organizations, where they follow few rules and regulations and impose the same to their users. In these sites, users have to create their profile by providing all those necessary details required and can use the features within the norms of the site. Each user may create blog documents and publish, where others in the network can access it after becoming friends. Each document contains a title and if attractive, catches the attention of others, while based on the content inside, it becomes popular. The users of blog sites are called Bloggers and perform activities like comment, trackback, scrap and bookmark, on others documents. Each blogger may create any number of documents as per their interest within the norms of the organization. However there are chances that the title and the content inside the document may not have any relationship. As time varies, users create documents on various topics upon their changing interest.

Based on the content of documents, the like-minded people get closer and form groups in the network. Such groups create documents and share it publicly to attain more popularity. Users usually create blog documents based on their interested topic (cricket, movie etc.), their knowledge (professionals), their experience and some social activities. But there is no end to the topics they involve and some topics evolve as the conversation increases among users on any specific document. Some users, if impressed from the content of document(s) of other users, they reproduce that content in their own documents and further share with their other groups and friends. 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 collection of formal and informal communication of text data that arrives over time is the blog data. Blogs are increasingly becoming a major source of information. The eminent feature of blogs is the accuracy and speed of extracting the recent information and discussion on a variety of topics.

The activities of users on the document(s) of others contribute in increasing the influential power of that document(s) and they occur depending on its content. Hence, content of a document plays a vital role in increasing its own influential power. The influential power of all documents of a user is the influential power of that blogger in the network. Finding such Influential Blogger by performing content analysis on their documents with respect to that of others in the network is a challenging task.

A user who attains more popularity becomes more influential in the network and attaining popularity directly depends on the content of documents created by that user. Achieving popularity in the network would be helpful to grab more attention for the products of company from users in a business prospective but the same opportunity can be encashed by some anti-social elements to propagandize their purposes as well. The reproduction of data in the network depicts the information diffusion in the network. The pattern of information diffusion identifies the virtual group that would have created within the network for serving either good or bad. Identifying the most influential user in the network based on information diffusion and its amount, is a challenging task and an active research topic.

In order to determine the amount of information diffusion in the network accurately, the appropriate approach would be through content analysis. Content analysis is performed in various ways; i) Formative Content Analysis and ii) Evaluative Analysis. The said approaches are used when the hidden agenda in the content is to be extracted [4]. The content is of different types like only text, text with image, image with hidden agenda (content), same image in different angle etc. Based on the type of content to be analyzed, the approach varies.

Extracting the information diffusion pattern in the network guides to analyze how and between whom the information is being spread, which in turn provides the group of users involved in the information diffusion process. Users in the network copy the document content of others and use it in their document(s), if the content is of utmost interest to them. Either full, partial, same or similar kind of information spreads from root document to other child document nodes in the network. The amount of information being diffused from root document to that of its child document nodes in the network decides the popularity of that root document. There are various methods available to find the content similarity between the documents and based on the content similarity measure/score, percentage of influence is computed.

Motivation: Influential Bloggers play predominantly as role models for other follower bloggers in the network. Information rapidly diffuse through these Influential Bloggers over the entire network. These Influential Bloggers many a times are used positively for advertising, e-commerce, e-business etc., but they are as well being used negatively in situations like propagandizing antisocial activities, which motivated us to carry out this work.

Contribution: In this work, we evaluated and presented an algorithm PTMIBSS, to profile the top most influential blogger based on the content of blog documents involved in the Information diffusion process. The contents of blog network documents are analyzed using tf-idf and their cosine similarity scores are obtained during conflicts in determining the Influential Blogger. Many a times, content in the document is so huge with sparse data, that quantifying the content with statistical measure becomes mandatory and that it will be either same or similar semantically, which requires additional approach to measure the similarity between the documents appropriately. Therefore, if the cosine similarity score is below the threshold, then the synonyms of those unmatched keywords are substituted to improve the score. Computing the Influence Score after Synonym substitution (*ISaS*) of those bloggers under conflict, the top most influential blogger is profiled.

Organization of remaining sections of the paper is as below: Section 2 gives a gist of literature. The background work is highlighted briefly in section 3. The problem is defined in section 4 whereas the proposed system is discussed in detail in section 5. Profiling Top Most Influential Blogger using Synonym Substitution (PTMIBSS) algorithm is presented in section 6. Simulation and Performance Analysis are discussed in section 7 presenting Conclusions in section 8.

2. LITERATURE SURVEY

It is important to identify both productive and influential blogger in the network. This seems to be challenging since a productive blogger need not be influential and vice versa. The authors of [5] have provided a mechanism to identify blogger who is both productive and influential. Two matrices are considered and based on the number of incoming links, bloggers are categorized into three kinds; either recently influential / productive / none or both. Bloggers activities, behavioral patterns and temporal patterns were identified easily using their proposed method and is proved from the experimental results conducted on En-gadget data set.

MASS [6], an effective model in identifying top-k influential bloggers by mining the network considers comments, authority in page link and interested domains of bloggers to evaluate them against influential in the network. Empirical demonstration of MASS is conducted and the results proved that it can be applied on multiple scenarios.

Eunyoung et al. [7] have provided an important variation to determine or differentiate popular bloggers from influential bloggers. Based on weightage of readers, influential bloggers are identified in this method. Using an interdisciplinary procedure, they have developed Quantifying Influence Model (QIM), which directly depends on number of readers and the experimental results proved that it is qualitative primary approach in determining influential bloggers in the network. Based on text similarity measured, linked graph of bloggers are identified in the network. Then PageRank algorithm is adopted to rank the bloggers. SimRank algorithm proposed by the authors [8] uses this data and recommends the influential bloggers in the network. Results of experiments prove that the algorithm can be used in blogger recommendation applications.

In healthcare community site, a forum for patients is provided to express their problems/experience/concerns to support their patients. But the forum contained lot of repeated data in the form of posts. ICHI 2015 gave an open requirement to reduce such repeated posts in the forum. To address the issue, the authors of [9] proposed a model which measures the similarity metric using TF-IDF. Their proposed model composed of TF-IDF and cosine similarity is effective than existing methods and is proved from the experimental results. To identify Stemmers, Emily Hill et al. [10] conducted qualitative study on software domain and specifically on Java source code. They conducted query-by-query study using Rank measure and MAP to retrieve the impact of stemming and its effectiveness.

Keywords were extracted effectively using a model proposed by Mohammed Haggag et al. [11] based on relativity weight measured with entire text terms. Apart from relatedness among them, semantic similarity is measured for finding the strength of terms relationships. Terms meanings are considered for assigning the weights and highly correlated words are considered both in weight likeness and frequency. Recursive elevations of keywords are performed accordingly to their cohesion with each other and with that of document context. The proposed approach is based on the five key concepts- keywords list; stability, word pair similarity, semantic relatedness, similarity score normalization and average similarity. Proposed approach achieved enhanced recall and precision extraction values compared to other approaches and is proved experimentally.

Few stemming algorithms along with their advantages and disadvantages are described in the survey conducted by Moral et al. [12]. With historical evaluation, assessment details of current status of stemming process are also explained. Stemming algorithms used for information retrieval process, where in meaningful stems are obtained using the TWIG stemming algorithm [13]. This algorithm reduced (proportion of un-meaningful to meaningful words) error rate of stemming process and the performance of this algorithm is proved with the same measure. Experiments conducted on Alan Beales Core Vocabulary Dictionary data proved that TWIG algorithm outperforms STANS and Porters stemming algorithm.

To automatically reduce the effect and impact of terms which are less informative, Boom et al. [14] proposed dense distributed representations and TF-IDF method's sparse term matching combined. To analyze and track the changing interests and changing activities of bloggers in the network using phrase dependency structures in sentences, Itoh et al. [15] proposed a 3D visualization technique.

Bipartite graph is developed by authors of [16], which depicts that the relationship exists between bloggers and the posts. Providing each link with weights based on relativity, a link-based rank approach to identify the influential bloggers is proposed. By semantically analyzing documents, bloggers are linked and their behaviors are analyzed. Topics are discovered to form closer nit group of bloggers in the network by the authors of [17]. Based on the bloggers activeness in the network and the number of posts, behavior of bloggers are forecasted using Parato/NBD model [18]. The effects and impacts of explosive communication in micro-blogging are discussed by the authors of [19].

Detailed characteristics of online social network with a model providing facility to its users to view real-time updates in social media and statistical data collected on daily basis related to their respective profiles is presented in [20]. It even provides data from the user profiles of Twitter, YouTube, SlideShare and Facebook through its APIs with key features. Asynchronous discussions happen in social networks, which leads to the need for the study of online social networks to analyze the communication structure and transcripts. Erlin et al. [21] developed Sollers model for social network and its content analysis quantitatively with nine sub and three main category variables. For the purpose analysis, they used network indicators to get access to the data and in-turn graph theory, adjacency matrix and network analysis techniques to extract interaction pattern in online asynchronous discussions. Experimental results proved that proposed model through content analysis identifies central actor and the users who are not involved in the discussion.

Boudiba et al. [22] presented a new approach using folksonomies to profile users in online social networks. A tripartite hyper graph is created using the keywords or tags via content indexing. This approach uses normalized measure [23] to weight vertices of graphs to provide degree of preference and identifies user's interests. A visualization toolkit to identify influence, analyze the propagation and prediction in multiple views of location distribution and time is developed by the authors of [24]. Apart from identifying the features of individuals, this approach also determines the features of information content. The results of visualization toolkit illustrate the influence of individuals, its propagation and also provides multiple ways of online social information.

A framework is presented to visualize and analyse social network gathered using query-dependent exploration [25]. The relationship between bloggers is identified using link analysis. Similar blog messages are explored directly and semantically using content analysis and topic underneath is identified. To explore at different levels of abstraction, users are provided with various interactive information visualization techniques. Social cognitive theory [26] is applied to analyze the conversation pattern, nature and propagation of information in online social network conducting experiments using Twitter data on cancer topic.

A study conducted on Content Analysis Mechanisms [27], describe the features of available content analysis models for cognitive aspects of students learning in discussion forums. Electronic databases like ScienceDirect, EdITLib, ProQuest and SpringerLink are used for article search during experimentation. Most researches cannot execute their proposed approaches since there is scarcity of data, especially in Twitter due to the restriction of 140 characters. To overcome this disadvantage, Adham et al. [28] described a novel approach to explore the complete knowledge in tweets using content analysis primarily. From much richer sources, more sophisticate data gets bootstrapped. Using Dirichlet process, they demonstrated the complex topic structure and evaluated the effectiveness of their proposed model on autism-related tweets.

Yung-Chung et al. [29] used content analysis mechanism to analyze the recent research topics from Taiwin in Information Technology field. By considering paper-tile and content, they have categorized papers into 5 catalogs. Using this method they found the decreasing and increasing researches of IT-related issues. Nitin Agarwal et al. [30] presented an approach using innovative ways for collective wisdom and to employ contextual information to aggregate bloggers which is a challenge, existing in aggregating similar bloggers. Considering category similarity and profile similarity as parameters, a Collective Wisdom based search approach was used to identify similar bloggers.

As an innovative approach, Faiza et al. [31] intended to identify the gender of blogger in online social network using texts written by them in their documents. For this purpose, they classified specific words into classes based on specific features. Based on these characteristics, they assigned a score to each blog and then found the bloggers gender. Experiments were conducted on Corpus dataset and results proved 82 percent effective than referenced collection. Based on two regressive techniques, Modified General Regression Neural Network and Extreme Learning Machine, three models were proposed by Bi Chen et al. [32] combining content, social dimensions and temporal dimensions for modeling the behaviors' of bloggers, conducting experiments on DailyKos blog dataset.

3. BACKGROUND WORK

Seung-Hwan Lim et al. [33] proposed an approach to determine the Influential Blogger in the network based on the highest *UCP* value. According to the authors, *UCP* is the summation of *DCP* of all the documents of a blogger and that the *DCP* of a document is the total number of activities performed by other bloggers. 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 [34] illustrates that a blogger who influences highest number of unique bloggers in the network as the IB. The authors consider that an activity, either *trackback*/*scrap*/*bookmark* performed for a document of one blogger (first blogger) from a document of other blogger is enough to declare that the other blogger is influenced by the document of the first blogger. Any further activities between the documents of those two bloggers have no significance in calculating the influential power of the first blogger. This results in obtaining the unique bloggers influenced by that first blogger. The comment activity is omitted while calculating the influential power of bloggers, since it does not necessarily prove that the document has influenced the other blogger. Thus, the authors illustrated that the blogger having highest Influential Blog Power (*IBP*) i.e., who has influenced highest number of unique bloggers as IB of the network.

The PTMIB algorithm [35] illustrates the adequacy of profiling the top most influential blogger during any conflict of influentiality based on the contents of blog network documents using tf-idf and cosine similarity.

4. PROBLEM DEFINITION

At any given point of time, there can be more than one blogger with the same Influential Blog Power (*IBP*) [34] leading to conflict in determining the top most influential blogger of the network. To solve such conflicts between the top influential bloggers, we in this work propose a novel method to determine and profile the Top Most Influential Blogger using Synonym Substitution approach amongst them.

In order to establish that a blogger's document has influenced other bloggers in the network, there needs to be some activity between that document and documents of other bloggers. By considering the activities performed on the blog documents, it can be inferred that there exists some influence between them, but to find out the extent to which the document has influenced other bloggers in the network is a challenge, which we have made an attempt to address in this work by analyzing their content.

5. PROPOSED SYSTEM

5.1 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. *Trackback* is an activity in the blog site where a blogger copies the content of a document of other blogger into his own document and gives back a link to the original document as to where he has copied it from. *Scrap* is another type of activity performed in the blog site by way of which a blogger copies the content of a document of other blogger into his own document and does not give back any link to the original document as to where he has copied it from. In both type of the activities, the blogger who copies the content can append his document with additional content other that the copied ones. Thus the contents of both the documents under consideration are analyzed and checked for content similarity between them using TF-IDF and Cosine Similarity approach.

5.1.1 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.

Considering that in a document, a term appears 20 times and that the total number of terms contained are 100, then the normalized term frequency, taking into account all the terms to be of equal importance is 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.

5.1.2 Cosine Similarity Score (CSS):

To find the similarity between two documents in their vector space, we first consider the magnitude of the vector differences between the two documents. To compensate the effect of document length, we find the similarity between two documents by computing the cosine similarity of their vector representations.

The *CSS* between two documents d1 and d2 is given by the formula:

$$CSS(d1,d2)=\frac{\text{DotProduct}(d1,d2)}{\|d1\|*\|d2\|} \quad (1)$$

where,

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

$$\|d1\| = \sqrt{\left(d1[0]^2 + d1[1]^2 + \ldots + d1[n]^2\right)}$$

$$\|d2\| = \sqrt{\left(d2[0]^2 + d2[1]^2 + \ldots + d2[n]^2\right)}$$

5.1.3 Synonym Substitution:

Substitution of synonym is performed when the *CSS* is below some threshold. Two blog documents [say Root(*r*) and Child(*c*)] are considered and the Cosine Similarity Score before Synonym Substitution (*CSSbS*) of the Child with respect to that of Root document is obtained as per Eq.(2).

$$CSS(r,c) = \frac{\text{DotProduct}(r,c)}{\|r\| * \|c\|} \tag{2}$$

where,

$$\text{DotProduct}(r,c) = r[0]*c[0] + r[1]*c[1] + \ldots + r[n]*c[n]$$

$$\|r\| = \sqrt{\left(r[0]^2 + r[1]^2 + \ldots + r[n]^2\right)}$$

$$\|c\| = \sqrt{\left(c[0]^2 + c[1]^2 + \ldots + c[n]^2\right)}$$

If their *CSSbS* obtained is less than the threshold fixed, then the process continues by copying the content of the Child(*c*) document to a Temporary(*t*) document. Every keyword in the Root is matched with that of the Temporary Child document and if the match is not found in the Temporary Child document, synonyms for that particular keyword is fetched from the WordNet 3.0.1 dictionary database. If any synonym matches the keyword(s) in the Temporary Child document, then each occurrence of the synonym is replaced with the keyword in the Root and the changes are saved to the Temporary Child document. The Cosine Similarity Score after Synonym Substitution *CSSaS* is then computed between Temporary Child document(*t*) and the Root(*r*) according to the Eq.(3).

$$CSS(r,c) = \frac{\text{DotProduct}(r,t)}{\|r\| * \|t\|} \tag{3}$$

where,

$$\text{DotProduct}(r,t) = r[0]*t[0] + r[1]*t[1] + \ldots + r[n]*t[n]$$

$$\|r\| = \sqrt{\left(r[0]^2 + r[1]^2 + \ldots + r[n]^2\right)}$$

$$\|t\| = \sqrt{\left(t[0]^2 + t[1]^2 + \ldots + t[n]^2\right)}$$

5.1.4 WordNet Dictionary Database:

It is a large lexical database in which nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms called synsets, each expressing a distinct concept. Important relationship between the words in the WordNet is synonymy, which are clustered as unordered sets (synsets). Every synset of WordNet's 117,000 synsets are interlinked by "conceptual relations" called "Word forms" with several distinct meanings. WordNet distinguishes among Types (common nouns) and Instances (specific persons, countries and geographic entities). Instances are always leaf (terminal) nodes in their hierarchies. Verb synsets are arranged into hierarchies, whereas verbs towards the bottom of the trees (troponyms) express increasingly specific manner characterizing an event. Adjectives are organized in terms of antonyms whereas Pairs of "direct" antonyms reflect the strong semantic contract of their members.

5.2 SYSTEM ARCHITECTURE

As shown in Fig.1, the log data of blog network is analyzed to profile the influential blogger based on the activities performed by unique bloggers according to the existing system [34]. If there are more than one influential blogger identified, then those under conflict are considered for further analysis in our proposed method and by computing the Influence Score (*IS*) of those bloggers, the top most influential blogger is profiled.

The Influence Score of a blogger is the summation of *CSS* to that of the Number of Bloggers Influenced (*N*) and is given by the equation:

$$IS = \frac{\sum CSS}{N}. \tag{4}$$

The System Architecture diagram in Fig.1 shows a detailed structure of the proposed system. The input to the proposed system is both *N* bloggers under conflict from the existing system and also the log data of the blog network which contains the activities based relationships between all the bloggers in the Blog Network. In the proposed system, firstly, activity based blog document network for each blogger under conflict is obtained.

Later, considering each document of the blogger under conflict as the Root document, the Cosine Similarity Scores between it and its Child documents are obtained.

In order to compute the Cosine Similarity Scores, the Root and Child documents are given as input to the HTML Parser, Word Extractor and Stopword Eliminator for the purpose of preprocessing and obtaining a list of keywords from both the documents. The HTML Parser removes the HTML tags and hyperlinks, the Word Extractor removes all the special characters and white spaces and the Stopword Eliminator removes all the stopwords from the documents.

Further, TF-IDF is computed which helps in computing the Cosine Similarity Score. Synonym substitution is performed based on the threshold condition. Synonym substitution uses the wordNet database to retrieve synonyms for the unmatched words. After Synonym substitution, the TF-IDF and Cosine Similarity Score is again computed. Further, the Influence Score is computed making use of the values of obtained Cosine Similarity Scores. Finally, the blogger with the highest Influence Score is profiled as the Top most influential blogger of the network.

5.3 SCENARIO ILLUSTRATION

The illustration of general scenario of PTMIBSS approach is as shown in Table.1. Two bloggers B1 and B2 are considered to have influenced the same number of bloggers, thus having the same Max(*IBP*) value. It is observed that the blogger B1 has 3 documents, whereas blogger B2 has 2 documents posted. Blogger B1's document B1D1 has influenced 1 unique blogger, B1D2 and B1D3 have influenced 2 and 1 unique bloggers respectively, whereas Blogger B2's documents B2D1 and B2D2 both have influenced 2 unique bloggers respectively.

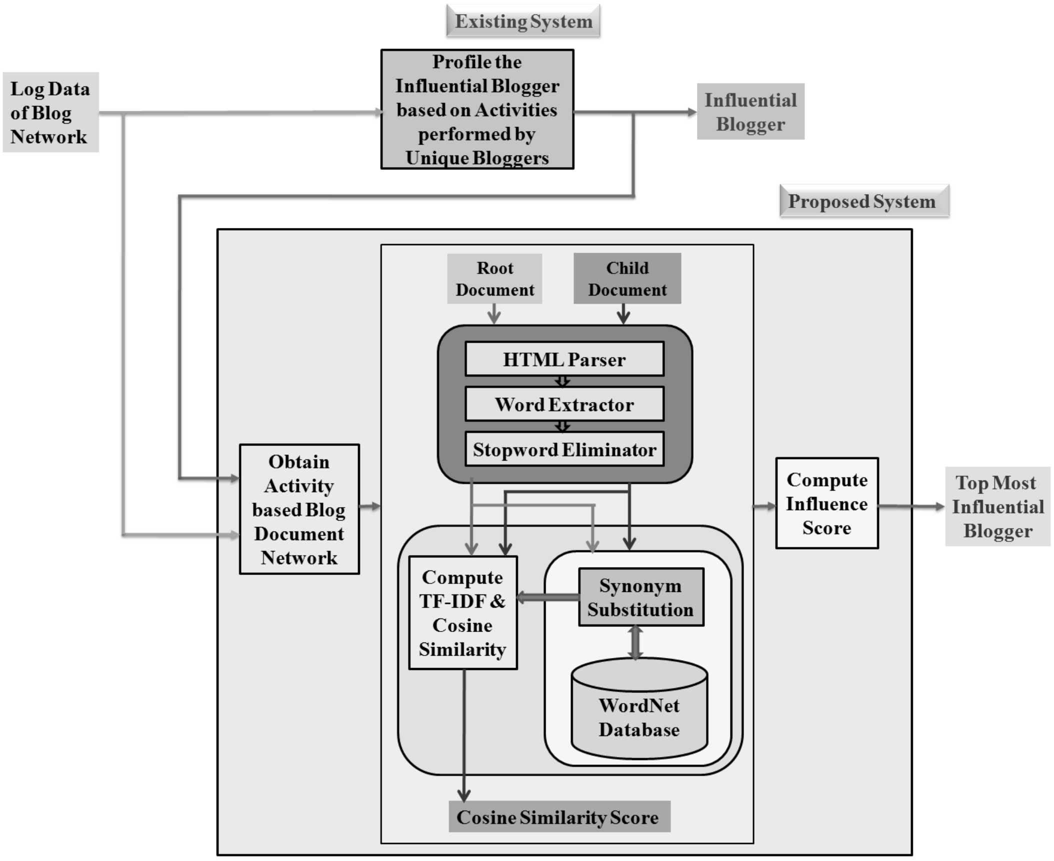


Fig.1. System Architecture

Table.1. PTMIBSS Approach Illustration

Bloggers Influenced same number of Unique bloggers	Number of Documents (*D*)	Document ID (*r*)	Number of Bloggers Influenced (*N*)	Blog Documents Network (*c*)	Cosine Similarity Score before Synonym Substitution (*CSSbS*)	Cosine Similarity Score after Synonym Substitution (*CSSaS*)	Influence Score before Synonym Substitution (*ISbS*)	Influence Score after Synonym Substitution (*ISaS*)
B1	3	B1D1 B1D2 B1D3	1 2 1	B5D3 B4D2, B3D1 B6D4	0.94 0.73, 0.72 0.92	0.94 0.85, 0.78 0.92	0.82	0.87
B2	2	B2D1 B2D2	2 2	B6D1, B5D2 B4D1, B3D2	0.80, 0.86 0.92, 0.82	0.80, 0.86 0.92, 0.82	0.85	0.85

Consider the document B1D1 as Root document of blogger B1 along with its child document B5D3 as per blog documents network shown in Fig.2 and find the keywords in each document by eliminating the Stopwords and Stemming the words.

Then the stemmed keywords of these two documents are processed through TF-IDF approach to obtain the *CSSbS* of 0.94 between them. The same procedure is repeated with B1D2 of blogger B1 which has B4D2 and B3D1 as its child documents to get 0.73 and 0.72 as *CSSbS* respectively. Continuing the same process, the *CSSbS* between B1D3 and B6D4 of 0.92 is obtained.

We have set the Threshold (T) to be 0.75 for better results and as the *CSSbS* of B4D2=0.73 and B3D1=0.72 with respect to the Root document B1D1 of Blogger B1 and is less than the Threshold, Synonym Substitution is performed and their *CSSaS* is computed to be 0.85 and 0.78 respectively. The *CSSaS* of those documents having their *CSSbS* above the threshold remain unchanged. The *CSSbS* and *CSSaS* of all the documents of the considered bloggers are as shown in Table.1.

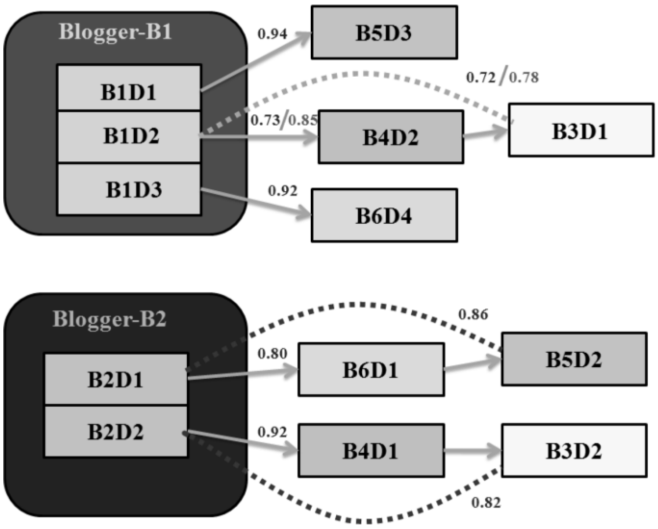


Fig.2. Activity based Blog Documents Network

The Influence Score before Synonym Substitution (ISbS) and the Influence Score after Synonym Substitution (ISaS) of the Blogger is computed according to the Eq.(5) and Eq.(6) respectively.

$$ISbS = \frac{\sum CSSbS}{N} \tag{5}$$

$$ISaS = \frac{\sum CSSaS}{N} \tag{6}$$

Observing *ISbS*, we can see that the values obtained for B1 is 0.82 and B2 is 0.85, which says that the Blogger B2 is the top most influential blogger of the network. But to make the determination of the top most influential blogger of the network to be more precise and accurate, when synonym substitution is performed for those documents, whose *CSSbS* is below Threshold, that is, for the documents B4D2 and B3D1 as per the considered scenario, their value increases to 0.85 and 0.78 respectively. Thus, the *ISaS* calculated for Blogger B1 and Blogger B2 are 0.87 and 0.85 respectively. Therefore, now the Blogger B1 is determined to be the Top Most Influential Blogger in the network accurately.

6. ALGORITHM

The various steps followed while profiling the top most influential blogger using synonym substitution approach is clearly shown in Algorithm 1. The input to PTMIBSS Algorithm is generally a set of *N* bloggers having same Max(*IBP*) value. The documents, number and type of activities performed on them, of each of those bloggers under consideration are utilized for computation in the algorithm and is designed to break the conflict between those bloggers as explained in the Algorithm 1.

Considering one blogger at a time, taking every document of that blogger as Root, stopwords are removed and stemming of each document is performed. TF-IDF is calculated for the Root. The Blog Document Network for each of the Root document is obtained from the activity-based log record. From the obtained Blog Document Network, TF-IDF and Cosine Similarity score of each of those Child document on the obtained network is calculated with respect to that of the Root.

Algorithm 1: ***Profiling Top Most Influential Blogger using Synonym Substitution (PTMIBSS) Algorithm***

```
while True do
  for "Every Blogger under conflict of Influentiality" do
    for "Every Document of Blogger, considering it as $r$" do
      Eliminate Stopwords and Perform Stemming
      Compute TF-IDF
      Obtain Activity-based Blog Document Network
      for "Every $c$ in Activity-based Blog Document Network, except that of $r$" do
        Eliminate Stopwords and Perform Stemming
        Compute TF-IDF
        Compute $CSSbS$ with respect to that of $r$
        if "$CSSbS < T$" then
          copy $c$ to $t$
          for "Every keyword in the $r$" do
            if "No match in $c$" then
              Fetch synonyms from the WordNet
              if "Any synonym matched in $c$" then
                Replace each occurrence of the synonym with the - keyword in $r$ and save it to $t$
              end if
            end if
          end for
          Compute TF-IDF of $t$
          Compute $CSSaS(r)$ with respect to that of $r$
        end if
      end for
    end for
    Compute Influence Scores $ISbS$ and $ISaS$ of the Blogger
  end for
  Compute Max($ISaS$)
  Profile the Top Most Influential Blogger of the network having Max($ISaS$)
end while
```

Each time the Cosine Similarity score is computed, it is compared against the Threshold and whenever it is below the set Threshold, synonym substitution process is revoked. During the synonym substitution process, for each unique keyword say '*w*' in the Root is compared with every other word on the Child document. When the match is found, it is simply copied to the Temporary document, where as if the match is not found, synonyms of that particular keyword is fetched from the WordNet database and kept in the buffer. Each synonym in the buffer is searched for in the Child document and if a synonym matches a word(s) in the Child document, then those words in the Child document are copied to the temporary document by replacing them with the keyword in the Root. This process is repeated for all the keywords in the Root. Later, TF-IDF of the temporary document is computed along with its Cosine Similarity score with respect to that of the Root.

Once the Cosine Similarity scores of all the Child documents in the Blog Document Network are obtained, the Influence Score of the blogger is determined. Likewise, the Influence Scores for all those bloggers under conflict are determined and the blogger with the Maximum Influence Score value is profiled to be the top most influential blogger of the blog network.

7. 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 205 bloggers posted blog documents. Roughly around 460 blog documents and nearly 1050 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 PTMIBSS algorithm.

The collected data is analyzed by extracting the number of activities on each and every document of bloggers and IB is computed according to UCP method [33]. From the same data set, we extracted the number of unique bloggers who were influenced from the documents posted by other bloggers in the network and then computed the IB of the network according to [34] as well. Later during some period of our analysis, we found that there exists more than one blogger as IB according to both [33] and [34], leading to conflict in profiling the Influential Blogger of the network.

We extended our analysis further and computed TF-IDF and obtained *CSSbS* of blog documents network of those bloggers under conflict. We then adopted Synonym Substitution approach and recomputed the cosine similarity scores for those documents below threshold *T* to obtain *CSSaS*. We later computed the Influence Scores after Synonym Substitution (ISaS) of those bloggers under conflict and finally profiled the Top Most Influential Blogger of the network having Max(ISaS) value as per our PTMIBSS algorithm.

7.1 THRESHOLD SETTING

In order to set the upper threshold *T* appropriately, we performed an experiment for determining the most accurate *T* value using the metrics such as *Precision*, *Recall*, *Accuracy* and *F1* measures. We considered 100 documents for this purpose, containing two activity-based blog networks, with 10 and 20 documents respectively and remaining 70 documents not belonging to either of the two activity-based blog networks from the data set that is logged from our blog site.

We observed that the root documents of the two activity based blog networks, R1 and R2 have influenced 9 and 19 child documents respectively. The Cosine Similarity Scores of all the documents with respect to both R1 and R2 were computed by considering *T* values from 0.5 to 0.99.

The Algorithm determines if a document is influenced by the root document, based on the threshold *T* value. If the influenced document is in the activity-based network of that root, then the result is considered to be True Positive (TP), and if not in the network, then it is considered as False Positive (FP). Similarly, if a document is identified as not influenced by the root and belongs to the network, then it is False Negative (FN), where as if it does not belong to the network, then it is True Negative (TN).

Precision is the positive predictive value, which is the number of influences that have been correctly identified and is given by,

$$Precision = \frac{TP}{FP + TP} \tag{7}$$

Recall (also known as sensitivity) gives the percentage of the actual influence identified and is given by,

$$Recall = \frac{TP}{FN + TP} \tag{8}$$

Accuracy gives the percentage by which the algorithm identifies the actual influence and is computed using the equation,

$$Accuracy = \frac{TP + TN}{\left(FN + TP\right) + \left(TN + FP\right)} \tag{9}$$

F1 score is another measure to test the accuracy which is given by,

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}. \tag{10}$$

We obtained the count of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) with respect to both R1 and R2. Using these values, we calculated *Precision*, *Recall*, *Accuracy* and *F1* scores for each value of *T*. The obtained values of the experiment are as shown in Table.2.

Table.2. Simulation Values for Threshold Setting

Threshold	**0.5**		**0.55**		**0.6**		**0.65**		**0.7**		**0.75**		**0.8**		**0.85**		**0.9**		**0.95**	
Root Document	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
Precision	0.375	0.5	0.375	0.5	0.375	0.5	0.375	0.5	0.5	0.66	0.6	0.75	0.5	0.66	1.0	1.0	1.0	1.0	1.0	1.0
Recall	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.25	0.5	0.75	0.75	0.25	0.5	0.25	0.25	0.25	0.25	0.25	0.25
Accuracy	0.867	0.88	0.867	0.88	0.867	0.88	0.867	0.88	0.714	0.8	0.889	0.895	0.714	0.8	0.727	0.727	0.727	0.727	0.727	0.727
F1 score	0.5	0.6	0.5	0.6	0.5	0.6	0.5	0.6	0.333	0.571	0.667	0.75	0.333	0.57	0.4	0.4	0.4	0.4	0.4	0.4

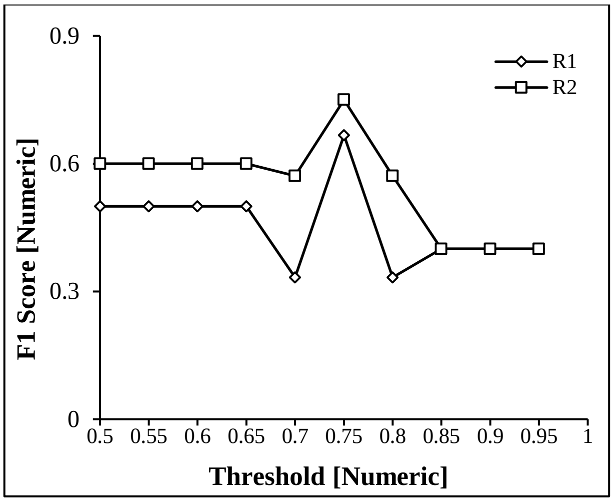


Fig.3. *F1* Scores over Threshold values

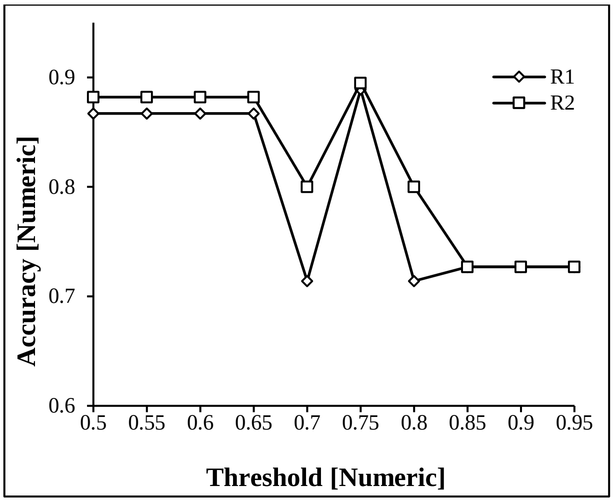


Fig.4. *Accuracy* measure over Threshold values

The values obtained from the experiment are plotted and are as shown in Fig.3 and Fig.4. From the graphs shown, we observe that both *F1* score and accuracy measure are maximum at threshold value *T*=0.75 and thus we set the threshold (*T*) to be 0.75 appropriately for Synonym Substitution in our algorithm.

7.2 RESULT ANALYSIS

The results of our algorithm simulation are as illustrated and shown in graphs with detailed discussion. The Fig.5 shows an intermediate output of proposed algorithm after Synonym Substitution Approach, where in the Output Document clearly highlights the word sources. The words in green are from the Root Document and those in red are in the Child Document which are added additionally by the blogger of the Child Document. The words in blue are the words in Root Document which were modified (presented in synonym form) in the Child Document. These words are highlighted in the Root as well as in the Child Document with blue colour.

The word “artistic” in Root Document has been modified as “aesthetic” in Child Document as shown in the figure. The word “aesthetic” is a synonym of the word “artistic”. Hence the word “aesthetic” is replaced by the word “artistic” of the Root Document in the Output Document. These words are also highlighted in the Root and Child Documents. That is, by highlighting “artistic” word in Root Document and also the word “aesthetic” in Child Document with blue colour. The word replaced in Output Document is also highlighted using blue colour to make it clearly understandable.

The graphs in Fig.6 shows the distribution of word sources of the Output Document obtained by performing content analysis on the documents shown in Fig.5. The first graph (Before Synonym Substitution) shows 84% of the words in the document are copied from the Root Document and 16% of the words in the document are additionally added. The second graph (After Synonym Substitution) is obtained after performing Synonym Substitution operation in which 5% of the words in the document are replaced by synonyms in the Child Document from the Root Document which were previously considered as additional words in Child Document.

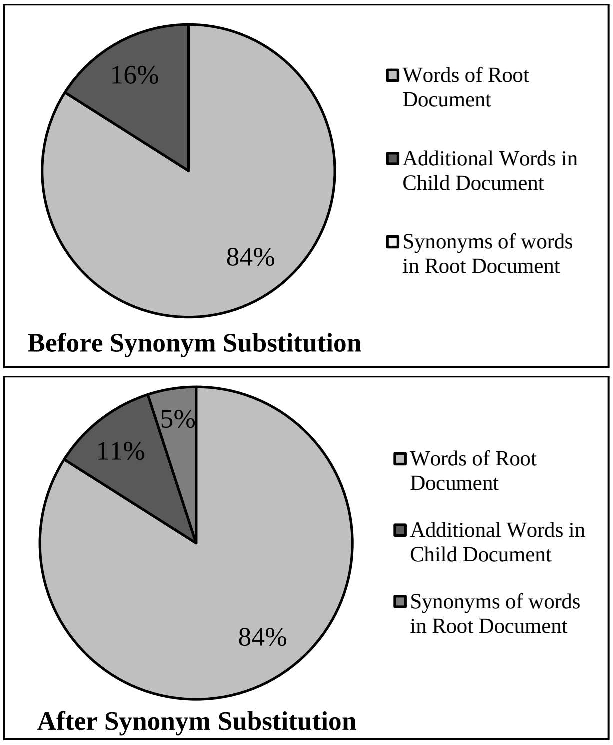


Fig.6. Distribution of Word Sources of the Output Document

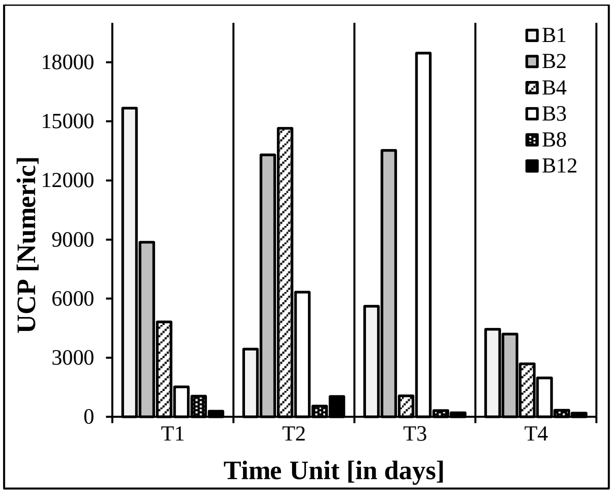
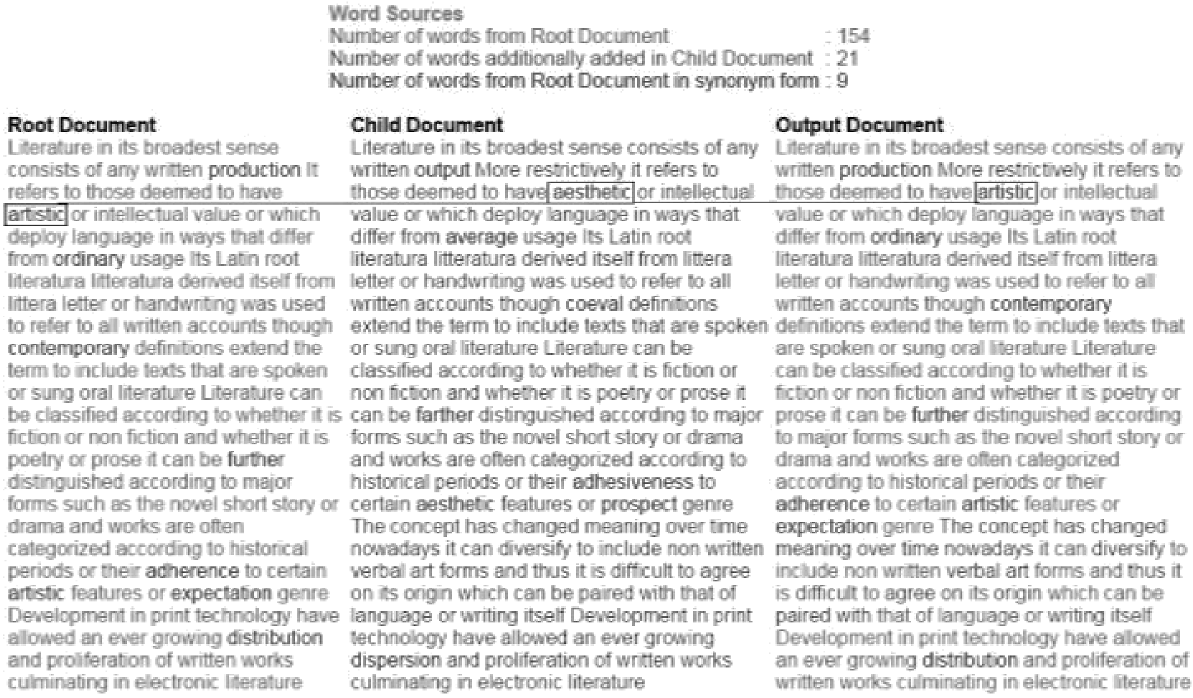


Fig.7. User Content Power of six Top Bloggers over Time



Word Sources

Number of words from Root Document : 154
Number of words additionally added in Child Document : 21
Number of words from Root Document in synonym form : 9

Root Document

Literature in its broadest sense consists of any written **production** It refers to those deemed to have **artistic** or intellectual value or which deploy language in ways that differ from **ordinary** usage Its Latin root literatura litteratura derived itself from littera letter or handwriting was used to refer to all written accounts though **contemporary** definitions extend the term to include texts that are spoken or sung oral literature Literature can be classified according to whether it is fiction or non fiction and whether it is poetry or prose it can be **further** distinguished according to major forms such as the novel short story or drama and works are often categorized according to historical periods or their **adherence** to certain **artistic** features or **expectation** genre Development in print technology have allowed an ever growing **distribution** and proliferation of written works culminating in electronic literature

Child Document

Literature in its broadest sense consists of any written **output** More restrictively it refers to those deemed to have **aesthetic** or intellectual value or which deploy language in ways that differ from **average** usage Its Latin root literatura litteratura derived itself from littera letter or handwriting was used to refer to all written accounts though **coeval** definitions extend the term to include texts that are spoken or sung oral literature Literature can be classified according to whether it is fiction or non fiction and whether it is poetry or prose it can be **farther** distinguished according to major forms such as the novel short story or drama and works are often categorized according to historical periods or their **adhesiveness** to certain **aesthetic** features or **prospect** genre The concept has changed meaning over time nowadays it can diversify to include non written verbal art forms and thus it is difficult to agree on its origin which can be paired with that of language or writing itself Development in print technology have allowed an ever growing **dispersion** and proliferation of written works culminating in electronic literature

Output Document

Literature in its broadest sense consists of any written **production** More restrictively it refers to those deemed to have **artistic** or intellectual value or which deploy language in ways that differ from **ordinary** usage Its Latin root literatura litteratura derived itself from littera letter or handwriting was used to refer to all written accounts though **contemporary** definitions extend the term to include texts that are spoken or sung oral literature Literature can be classified according to whether it is fiction or non fiction and whether it is poetry or prose it can be **further** distinguished according to major forms such as the novel short story or drama and works are often categorized according to historical periods or their **adherence** to certain **artistic** features or **expectation** genre The concept has changed meaning over time nowadays it can diversify to include non written verbal art forms and thus it is difficult to agree on its origin which can be paired with that of language or writing itself Development in print technology have allowed an ever growing **distribution** and proliferation of written works culminating in electronic literature

Fig.5. Intermediate output of proposed algorithm after Synonym Substitution Approach

The Fig.7 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 [33].

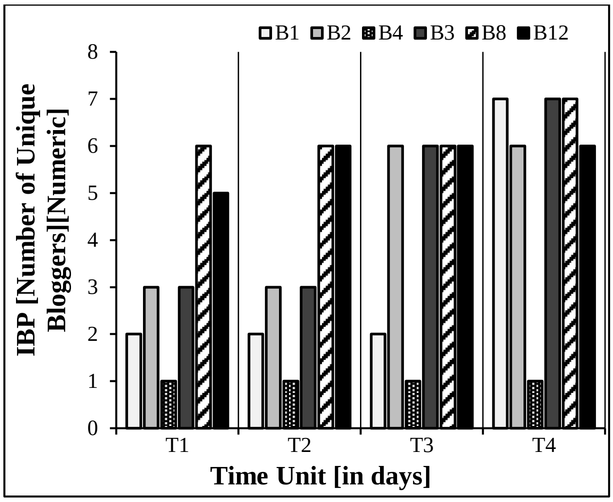


Fig.8. Influential Blog Power of six Top Bloggers over Time

The Fig.8 depicts the *IBP* with respect to Time Unit. As according to the PIB algorithm [34], 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 PTMIBSS algorithm is proposed to determine and profile the top most influential blogger.

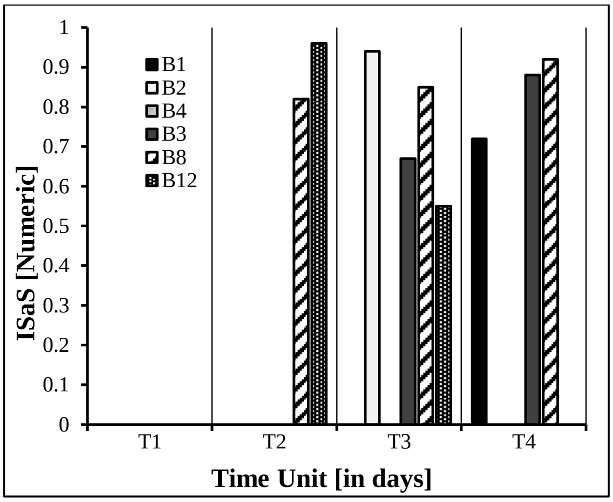


Fig.9. Influence Score after Synonym Substitution of Bloggers having same Max(IBP) over Time

The Fig.9 shows the simulation results of our proposed algorithm. As depicted in the figure, during T2, B12 has *ISaS*=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 *ISaS* of those bloggers and accurately profile the Top Influential Blogger during time units where conflicts arise.

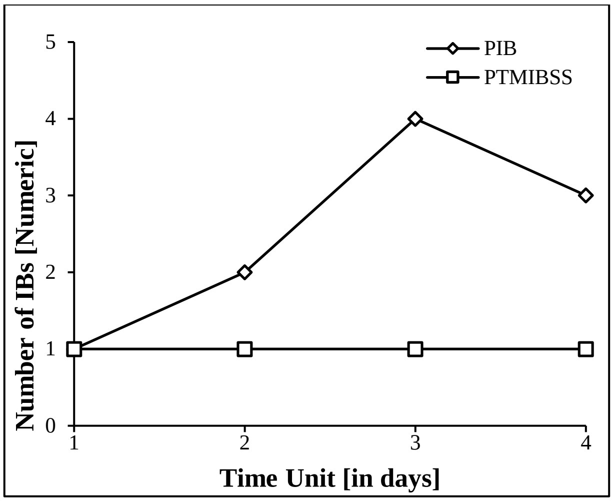


Fig.10. Number of Influential Bloggers over Time

Comparing the proposed approach with the previous PIB [34], Fig. 10 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 overcome by using proposed PTMIBSS approach giving rise to only one Top Most Influential Blogger in the network during any time unit of analysis.

The Fig.11 shows the Influence Scores of Top Most Influential Blogger Before and After Synonym Substitution over Time. As observed in the graph, the Influence Scores of top most influential bloggers in both the time units T2 and T3 remained to be the same Before and After Synonym Substitution as the *CSSbS* of their documents were above the threshold value. Where as in time unit T4, we notice that the top most influential blogger having *ISaS*=0.95 had previously *ISbS*=0.72 and this is because of the reason that the *CSSbS* of few of thier documents were below the threshold value leading to recomputation of their Cosine Similarity scores *CSSaS* after Synonym Substitution.

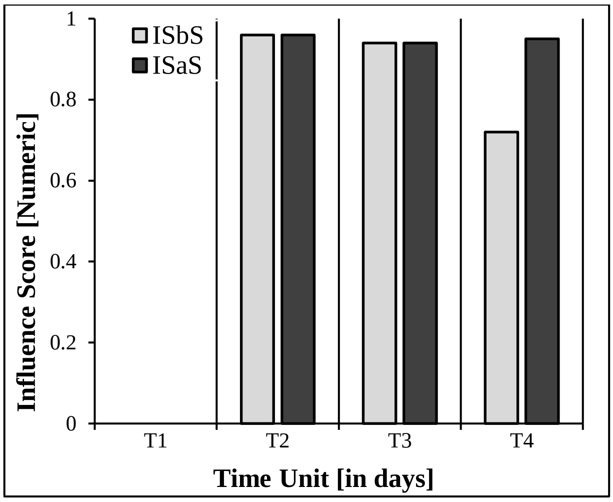


Fig.11. Influence Scores of Top Most Influential Blogger before and after Synonym Substitution over Time

This is because of the fact that the content of Child documents in the Activity based Blog Documents Network of that blogger had more of similar content having the same semantic meaning, leading to the increase in *ISaS* of that blogger after recomputation of their cosine similarity *CSSaS* after Synonym Substitution Approach.

8. CONCLUSIONS

The paper presents PTMIBSS algorithm for profiling the top most influential blogger. Methods adopted in the literature for determining and profiling the influential blogger were not only based on number of activities performed on the blog documents [33], but also based on the number of unique bloggers influenced by the blog documents of other bloggers [34] [36]. 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 of simulation demonstrate the adequacy and accuracy of our proposed PTMIBSS algorithm.

The PTMIBSS algorithm when applied on the blog network of criminals, helps in determining the information diffusion of criminal activities, as well as in identifying the head of their group. The approach can be used for targeted marketing and advertising, in product based business enterprises. Open avenues for future work are in analyzing the content of blog documents semantically across all the documents of all the bloggers, irrespective of their network.

REFERENCES

[1] Cristina Castronovo and Lei Huang, “Social Media in an Alternative Marketing Communication Model”, *Journal of Marketing Development and Competitiveness*, Vol. 6, No. 1, pp. 117-136, 2012.

[2] P. Deepa Shenoy, K.G. Srinivasa, K.R. Venugopal and Lalit M. Patnaik, “Evolutionary Approach for Mining Association Rules on Dynamic Databases”, *Proceedings of the 7th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp. 325-336, 2003.

[3] P. Deepa Shenoy, K.G. Srinivasa, K.R. Venugopal and Lalit M. Patnaik, “Dynamic Association Rule Mining using Genetic Algorithms”, *Intelligent Data Analysis*, Vol. 9, No. 5, pp. 439-453, 2005.

[4] Colleen Jones, “Clout: The Role of Content in Persuasive Experience”, *Proceedings of the First International Conference of Design, User Experience and Usability: Theory, Methods, Tools and Practice*, Vol. 6770, pp. 582-587, 2011.

[5] 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.

[6] Yichuan Cai and Yi Chen, “Mass: A Multi-Facet Domain-Specific Influential Blogger Mining System”, *Proceedings of 26th IEEE International Conference on Data Engineering*, pp. 1109-1112, 2010.

[7] 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.

[8] 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", *Proceedings of International Conference on Machine Learning and Cybernetics*, pp. 3368-3373, 2010.

[9] Mohammad Alodadi and Vandana P Janeja, "Similarity in Patient Support Forums using TF-IDF and Cosine Similarity Metrics", *Proceedings of International Conference on Healthcare Informatics*, pp. 521-522, 2015.

[10] Emily Hill, Shivani Rao and Avinash Kak, "On the use of Stemming for Concern Location and Bug Localization in Java", *Proceedings of IEEE 12th International Working Conference on Source Code Analysis and Manipulation*, pp. 184-193, 2012.

[11] Mohamed H Haggag, "Keyword Extraction using Semantic Analysis", *International Journal of Computer Applications*, Vol. 61, No. 1, pp. 1-6, 2013.

[12] Cristian Moral, Angelica de Antonio, Ricardo Imbert, and Jaime Ramirez, "A Survey of Stemming Algorithms in Information Retrieval", *Information Research*, Vol. 19, No. 1, 2014.

[13] 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.

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

[15] Masahiko Itoh, Naoki Yoshinaga, Masashi Toyoda and Masaru Kitsuregawa, "Analysis and Visualization of Temporal Changes in Bloggers' Activities and Interests", *Proceedings of IEEE Pacific Visualization Symposium*, pp. 57-64, 2012.

[16] Lu and Fuxi Zhu, "Discovering the Important Bloggers in Blogspace", *Proceedings of IEEE International Conference on Artificial Intelligence and Education*, pp. 151-154, 2010.

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

[18] Rui, Cai, Qi Jia-yin and Wang Mian, "Forecasting Bloggers' Online Behavior based on Improved Pareto/NBD Model", *Proceedings of IEEE International Conference on Management Science and Engineering*, pp. 84-90, 2013.

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

[20] Riccardo Cognini, Damiano Falcioni and Alberto Polzonetti, "Social Networks: Analysis for Integrated Social Profiles", *Internet Technologies and Applications*, pp. 68-72, 2015.

[21] B. Erlin, Norazah Yusof and Azizah Abdul Rahman, "Analyzing Online Asynchronous Discussion using Content and Social Network Analysis", *Proceedings of IEEE Ninth International Conference on Intelligent Systems Design and Applications*, pp. 872-877, 2009.

[22] Boudiba Tahar-Rafik and Ahmed-Ouamer Rachid, "Towards a New Approach for generating user Profile from Folksonomies", *Proceedings of IEEE 4th International Symposium on ISKO-Maghreb: Concepts and Tools for knowledge Management*, pp. 1-6, 2014.

[23] Yi Cai and Qing Li, "Personalized Search by Tag-based User Profile and Resource Profile in Collaborative Tagging Systems", *Proceedings of 19th ACM International Conference on Information and Knowledge Management*, pp. 969-978, 2010.

[24] Bo Wang, Yingjun Sun, Cheng Tang and Yang Liu, "A Visualization Toolkit for Online Social Network Propagation and Influence Analysis with Content Features", *Proceedings of IEEE International Conference on Orange Technologies*, pp. 129-132, 2014.

[25] Christopher C. Yang and Tobun D. Ng, "Terrorism and Crime related Weblog Social Network: Link, Content Analysis and Information Visualization", *Intelligence and Security Informatics*, pp. 55-58, 2007.

[26] Hong-Jun Yoon and Georgia Tourassi, "Analysis of Online Social Networks to Understand Information Sharing Behaviors through Social Cognitive Theory", *Proceedings of Annual Oak Ridge National Laboratory Biomedical Science and Engineering Center Conference*, pp. 1-4, 2014.

[27] Noor Izzati Ariff and Zaidatun Tasir, "Meta-analysis of Content Analysis Models for Analysing Online Problem Solving Discussion", *Proceedings of IEEE Conference on e-Learning, e-Management and e-Services*, pp. 148-152, 2015.

[28] Adham Beykikhoshk, Ognjen Arandjelovic, Dinh Phung and Svetha Venkatesh, "Overcoming Data Scarcity of Twitter: Using Tweets as Bootstrap with Application to Autism-related Topic Content Analysis", *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 1354-1361, 2015.

[29] Yung-Chung Tsao, Kevin Chihcheng Hsu and Yin-Te Tsai, "Using Content Analysis to Analyze the Trend of Information Technology Toward the Academic Researchers at the Design Departments of Universities in Taiwan", *Proceedings of IEEE 2nd International Conference on Consumer Electronics, Communications and Networks*, pp. 3691-3694, 2012.

[30] Nitin Agarwal, Huan Liu, Shankara Subramanya, John J. Salerno and S. Yu Philip, "Connecting Sparsely Distributed similar Bloggers", *Proceedings of 9th IEEE International Conference on Data Mining*, pp. 11-20, 2009.

[31] Faiza Belbachir, Khadidja Henni and Lynda Zaoui, "Automatic Detection of Gender on the Blogs", *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications*, pp. 1-4, 2013.

[32] Bi Chen, Qiankun Zhao, Bingjun Sun and Prasenjit Mitra, "Predicting Blogging Behavior using Temporal and Social Networks", *Proceedings of Seventh IEEE International Conference on Data Mining*, pp. 439-444, 2007.

[33] 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.

[34] G.U. Vasanthakumar, Bagul Prajakta, P. Deepa Shenoy, K.R. Venugopal and Lalit M. Patnaik, "PIB: Profiling Influential Blogger in Online Social Networks, A Knowledge Driven Data Mining Approach", *Proceedings of Eleventh International Multi-Conference on Information Processing*, Vol. 54, pp. 362-370, 2015.

[35] G.U. Vasanthakumar, R. Priyanka, K.C. Vanitha Raj, S. Bhavani, B.R. Asha Rani, P. Deepa Shenoy and K.R. Venugopal, "PTMIB: Profiling Top Most Influential Blogger using Content Based Data Mining Approach", *Proceedings of IEEE International Conference on Data Science and Engineering*, 2016.

[36] G.U. Vasanthakumar, P. Deepa Shenoy and K.R. Venugopal, "PTIB: Profiling Top Influential Blogger in Online Social Networks", *International Journal of Information Processing*, Vol. 10, No. 1, pp. 77-91, 2016.