IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 4, Ver. V (Jul – Aug. 2014), PP 73-80
www.iosrjournals.org

Mining Of Influential Users in a Blog Network

Azaim Khan
Lovely Professional University, School of technology and sciences,
NH-1, Punjab, India

Abstract: *Blogging sites are very popular in today's world; users interact with each other and create social relationships between them. Data mining methods can be used to extract the blogging sites. Users can be active or dormant in a blog environment, they could be directly influenced by the motivators or not. Direct DCP and Indirect DCP both can be calculated. Some influential user's motivates the other normal users on the basis of some features through which rating could be high and it could be increase the business in terms of monetary. Determining the Content Power Users (CPU) is very essential now a day. It represents the blogging atmosphere which is used to calculate the influential users.*

Keywords: *Blogging sites, Content Power User, Direct DCP, Document Content Power, Data mining, Indirect DCP.*

1. Introduction

Novel developments in internet self gesture and networking are quickly moving the document formation from offline to online approach. As a result, social sites are very popular among users through which they can write comments, can modify any information if required and also establish online relationship between various users. Blogosphere is a combination of all the different blogs and it is an e.g. of social site. Any user can create a blog related to any field or topic and they can issue their blogs also via internet those users who publish their own blogs are known as author of that blog. The service given to the user in which they make the good production as a result blog environment is maintained. All the users develop social relations with other users with the use of blog services. Blog represents the social network (for e.g. face book.com). Through this idea activity of other users on owners blog can also be recognized very easily whether that user is interested on that specific blog or not. Influential users can be recognized in terms of business due to which money affected. Establish method to identify large number of influential user who can make comments, shares, likes etc in blog. Influential users are those who encourage other normal users to do some activities on a particular blog. The main disadvantage of social sites is that company is concentrating on important relationship issue because of poor blogging [6].

1.1 Developing a blog environment

In this section blog keep track of user interest on which blog which makes it simple and suitable for user to visit those blogs which is known as bookmark, blogroll or neighbor. T in fig.1.1 (a) shows the trackback means writing a new text linked to someone else's text while putting a connection to the original document in one's personal blog. S in fig.1.1 (a) represents the scrap which means copies someone else's document in our own document.

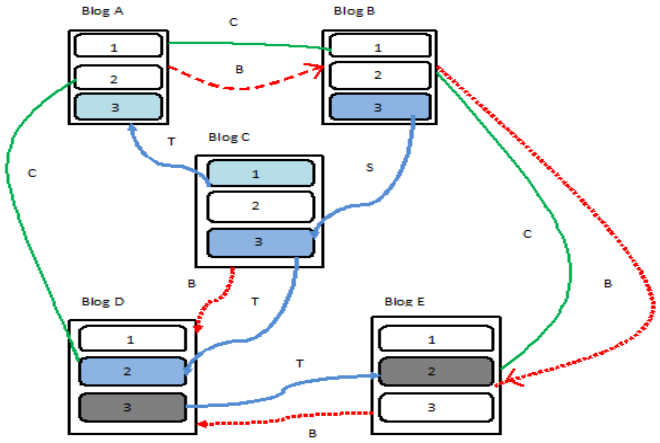


Fig.1.1 (a) Example of blog environment: - Blog user's actions.

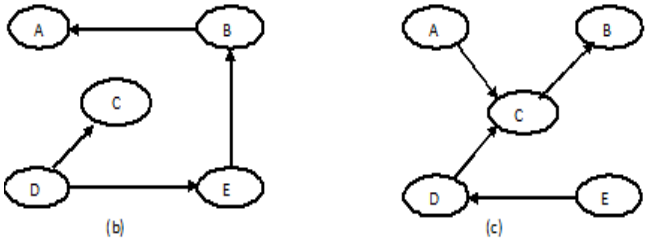


Fig.1.1 (b) Blog environment based on bookmarks. **(c)** Blog network based on trackback and scrap actions.

C shows the comments which are written by users and owner of the blog in fig.1.1 (a). In fig.1.1 (a) Blog A, Blog B, Blog C, Blog D, Blog E is the number of blogs and small rounded circle represents the document in that blog. B stands for bookmark which is used to define the two user's actions such as: first action represent the influence relationships between user and a blog. Second action represents an influence relationship between the owner's document and the user who has performed some action on that document of owner. C shows comments; user can write any comment on any blog if they are willing to do. S denotes scrap means copied someone else's document into owns document.

In fig.1.1 (a) arrow represents an action from one document to another such as trackback, scrap, comment and bookmark. Documents in the same shade means they are connected during actions. For example user A place comments on document 1 of blog B whereas user C set trackback to document 2 of blog D in document 3within user's blog and user B also put trackback to document 3 of user C.

In fig.1.1 (b) and (c) circle denotes blog and arrow denotes the influential relationship between blogs. Fig.1.1 (b), in which B →A means user A place blog B in his bookmark which represents user B has some influence over blog A. Similarly in fig.1.1 (c) A→C shows that user C set trackback link over document 3 of user A which displays the influence of user A over user C [6].

2. Problem

2.1 Problem formulation

Investigating the influential relations in a blog network environment some work has about the development of blog network. Based on the bookmarks or user actions, two different influence relationships can be defined they are as shown below: -

- Firstly there can be an influence relationship between a user and a blog through the use of bookmarks. While using bookmarks for constructing the blog network they may not pass to confine the dynamic and fast-changing influence interactions in a blog which helps to develop the relationships in a blog network.
- Secondly a relationship between several users on social networking site can exists a user and owner of a document through a way from which an action can be performed over the novel document of the owner. The actions like trackback or scrap which is used to spreading the impact of influence and can be measured for direct or indirect influence of a document in a blog network.

DCP defined as level of influence of a document can be considered by adding up the weighted occurrences of other user's actions which encourage by that particular document. It needs some parameters that can help for calculating the degree of a document. At the same time DCP calculation shall produce results that can be used to fetch the high average utility documents and users.

2.2 Objectives

Prediction of influential users and influential documents in a blog network has been done. Mining of powerful documents and users as the intention for this work has been proceed to calculate the following details out of the blogosphere.

1. To calculate the direct influence of the users from the document.
2. To calculate the indirect influence of the users from the document.
3. To calculate the users utility on a blog this is based on documents by giving some threshold value by the user.

3. Current Work

3.1 Scope of the study

The idea of Content Power User (CPU) introduces a blog environment which is the mixture of users which uses the blog and users activities on that specific blog.

Following are the some important points for creating a blog atmosphere by calculating the influential users on a given blog, which create social relationships between various users.

- It deals with the users various activities on a blog network. All the activities of normal user on other user document can be calculated which shows the user's get influence by the blog of other user. User can do some activities such as comments, like, rating, track back etc. all these features are used to identifying the influential users in a blogosphere.
- Introduces a new method to identifying the document of any blog. Through document user gets more influenced by it that's why document is so significant for a blog. Content of a document should always give an idea about that particular blog due to which users can attract towards it. Direct and indirect users behavior shows some impact on blog.
- User's content can be analyzing by simply writing important document on a blog. If blog is popular than it shows document which contain some influential topic through which any user can attract on that specific blog.
- Finally, creating the CPUs in a blog environment by selecting the number of users from the users influential document through which normal users get attracted on it [8].

3.2 Research methodology used

This research shows the content mining of a blogosphere which is a collection of several blogs and data can be gathered from many various sources. On applying techniques of data mining on huge data user can get outcome for blogs which represents that a particular blog is good or not. Influential users can get by observing the influential document of a blogs. Influential users always encourage the other users for a specific blog. Due to which active, inactive users can also be recognized and which can do some activities on a specific blog. Data can be collected by internet, blog data is used here. Do classification technique of data mining with the help of coding which is used to classify the users, user id then clustering is done. Clustering means grouping of similar type of objects in same cluster which is different with other group cluster object. Object of cluster1 has different with respect to object of other group cluster. Following diagram represents the flow chart of doing the research

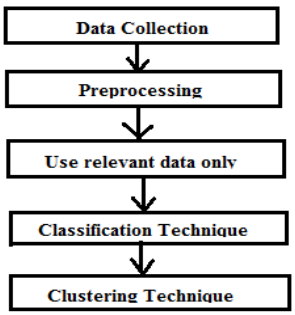


Fig 3.2 Flowchart of Research Methodology

By applying above research methodology flow chart in research getting accurate and good results. Influential users can easily identified which show some impact on other user in a particular blog through which user get influenced and increases rapidly daily. Influential users are different with other normal users because influential user encourages the normal users to do some activities on a blog. Direct Document DCP method is used to recognize the comments of a specific blog and Indirect Document DCP method is used to what to write in blog. Document exposure time is used to calculate the start and stop view time of each blog. Users can comment, give rating, track back also on a specific blog. Which helps to represent the content is influenced or not and high utility of blog. Provide some threshold value which is used to identify high utility, influential document etc. Influential users can be calculated by some features like comments, track back, rating in a specific blog etc. Through which blog utility increases many user influences by a blog and influential user also helps to motivate the other normal user.

4. Previous Work

The statistical Analysis method is used which represents relation between documents of a blog and where users are engaged in a blog means which activity they are doing due to which they are engaged in a blog. All the organizations know the significance of social networking of a blog via internet which can spread and advertising products and services also. Organizations introduce their own blog to support their products [1].

In blog network some users are special because they motivate other users for blogging. Motivational users are the influential user who motivates other user. Following are the major role of this paper: Content of power users can be derived by the characteristic of social networks. The new concept of CPU was introduced. Method for measuring the content power of document and also the influential users so proposed a normalization

method. By requesting field experts for user study and also proposed a method through which it can find those users who actually contribute in the blog network. This paper introduces the several business models that use the concept of CPU's to promote the activities of users in the blog network [7].

The web atmosphere gives chance to take aspects of textual information into report. It evaluates the technique for enhancing blog using the hidden information in web comments. According to the outcome of the experiments relations among commentators can be used in content clustering and get better the quality of content clusters. Post-processing technique of blog clusters is introduces which deals with the how user comments on several blogs. Cluster represents the commentators but it is different from the content clusters. Value of content cluster is better than by using hidden technique in which commentator don't act well into single clusters. [7]

Recently blog systems level 'A-list' bloggers, but they are not necessarily powerful. Proposed a new technique called QIM which is used to analyze influential bloggers score. This technique is the mixture of two mechanisms: it represents the communication between bloggers and those users who likes that particular blog and how much that information is distributed on a blog means the level of that distribution of information on blog. It also shows how many users has own that specific blog (which users use that blog on a regular basis) [2].

In current days, a lot of effort has been done in the field of Opinion Detection in blogs means detection of attitude of user. The aim is to focus on Social Network that can be utilizing for the task of Opinion Detection. It plans a structure that utilizes the main elements which means essential opinions from the blogs. As well as also highlight the tasks of opinion prediction [4].

In recent time blogs rapidly increases with simple tools for more popularity on the internet. Blogs become main information source. Bloggers launch news and products to blog visitors and current related information. Through communication users can interact with other users and can make people encourage and impressed by that specific blog. Result of a regression analysis shows that blogs has popular over other blogs users get attracted towards that particular blogs because it document is so impressive. Users give positive reply which shows positive attitude of users over a particular blog [3].

The Opinion Detection from blogs has very important task for researchers which is used to detect the opinion of users because they always differ person to person. One of the major tasks is that to find such documents that particularly include opinion which depend upon user's information requirement. This requires content processing on sentence level not on document level. An opinion detection approach has been proposed. It deals with the opinion detection problem by using some document level heuristics approach and processing documents on sentence level using different semantic similarity. The task is to divide the selected opinionated documents on the basis of positive, negative or neutral opinion on the given subject [5].

5. Implementation Work

5.1 Results and discussion

In a diffusion history all the research work is represented in the form of graph and table which actually want to implement in .net language. .Net language is used for implementation work of my research work which is more suited and problem can easily be solved by it. The following are the objectives of my research work which is implemented on .net language:-

- **Direct Document DCP: -** which is used to calculate the total comments by a user in a specific blog or a particular blog. Comments could be anything it is not related with any word because content mining is not doing here. Taken 10 factors for calculation means for blog id 3 DCP would be 10*2=20, here 2 represents the count of comments given by a user for a particular blog. In this way all the values could be calculated very easily and below graph shows the pictorial representation of it. Users can do any comment on any blog which will add weight 10 percent with per blog which is called read feature of the blog. In this way direct weight for the particular blog can added.

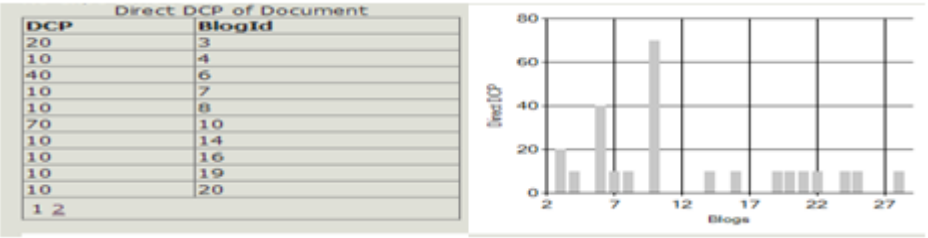


Fig 5.1 (a) Direct Document DCP

- **Indirect Document DCP: -** It is used to show what to write in a blog by a particular user with the help of track back link which is indirectly used by user for commenting on a blog. Track back means writing a new content of a blog which can be linked to someone else's blog though putting a link of a original document of any user in owns blog it is not like this that user gets motivates by their nearest users other than as user is being influenced by a particular users who has that meticulous blog. Count the trackid from the blog table for example if track id=0 it is showing 13 times in blog table then 13 * 8=104 (DCP) where 8 is the factor. In this way indirect DCP can be calculated by factor of 8. DCP=104 means indirectly with the help of track back user get influenced by that particular blog not by any other thing. Track back means already created blog by someone.

Indirect DCP of Document

TrackId	DCP
0	104
3	8
4	8
6	24
9	8
10	8
12	8
18	8
19	8
20	8

1 2

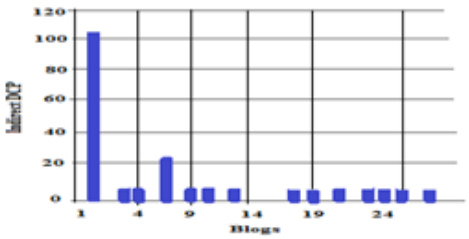


Fig 5.1 (b) Indirect Document DCP

- **Document Exposure Time: -** Exposure time in seconds. It is used to calculate the time of start view and stop view of each blog by the entire user and always used Blog Id. For example if Blog Id= 7 and expo time =4. Means how much time spends by a particular user on a single blog because if user is interested on it then only users view that blog for reading many more comments on it. Exposure time means how much time devoted by a particular user on any blogs.

Exposure Time in Seconds

BlogId	UserId	ExpoTime
3	gunjan	244
4	sameer	455
5	gunjan	87
6	mahendra	17
7	vijendra	4
8	gourav	0
9	sarab	15
10	sarab	55
12	neklata	7
15	girish	2

1 2 3

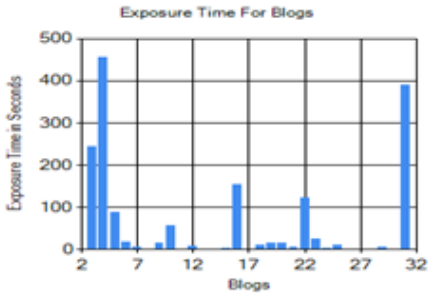


Fig 5.1 (c) Document Exposure Time

- **Document DCP: -** It can be used by calculating all the values of D_DCP and I_DCP of DCP along with exposure time. Which means it will show the result of all those users who has some calculated DCP, it will not display if DCP is 0.

Document DCP

User	DCP
gunjan	9.700000
bachchu	9.600000
laimani	9
pooja	9
priyanka	9
mahendra	17.200000
abhay	14
sarab	12.300000
vijendra	11.000000
girish	10.600000

1 2

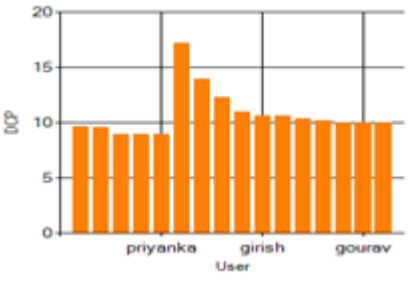
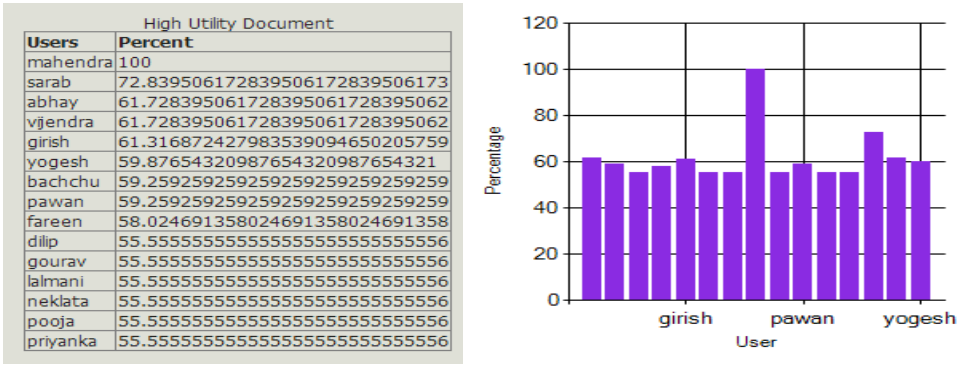


Fig 5.1 (d) Document DCP

Fig 5.1 (e) Document DCP

- **View high utility value: -** By entering threshold value 50% then click on first option which is used to display highly used documents which will show the more than 50% of the utility document not equal to or less than of it. It will always show the more than of threshold value if suppose 50% then it will represent only more than 50% not less or equals to of it. Below figure show the high utility document by all the users whose threshold value is more than 50% and graphical representation also display the pictorial representation of it.

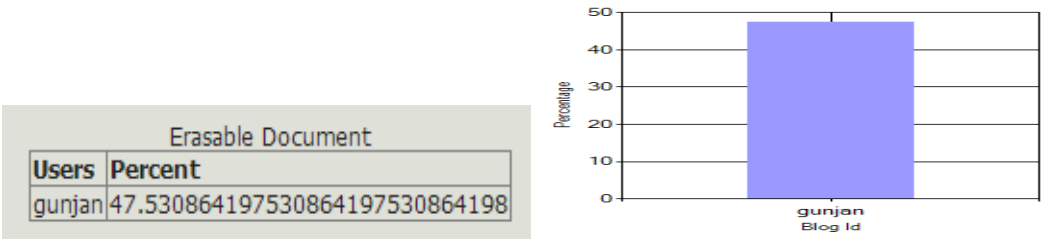


High Utility Document

Users	Percent
mahendra	100
sarab	72.83950617283950617283950617 3
abhay	61.728395061728395061728395062
vijendra	61.728395061728395061728395062
girish	61.316872427983539094650205759
yogesh	59.876543209876543209876543 21
bachchu	59.259259259259259259259259
pawan	59.259259259259259259259259
fareen	58.024691358024691358024691358
dilip	55.555555555555555555555555556
gourav	55.555555555555555555555555556
lalmani	55.555555555555555555555555556
neklata	55.555555555555555555555555556
pooja	55.555555555555555555555555556
priyanka	55.555555555555555555555555556

Fig 5.1 (f) High Utility Document

- **View erasable document: -** Erasable document means users have done nothing like no comments, no rating, no time spend by a particular user on any blog. It will display less than of threshold value if it is given 50% then it will display less than 50% not equal to it also. It will display only those users name which has not used any blog at which percent. Those documents are not used by any users either by comments, rating or time. It will show users name along with their percentage, if represents 47% this means user is not using some blogs up to 47% but it doesn't mean that user is not using any blogs. Users do their actions such as comments, rating, time spends in any blog but 60% of it will show this scenario.



Erasable Document

Users	Percent
gunjan	47.530864197530864197530864198

Fig 5.1 (g) View erasable documents

- **Influential Users by Comment: -** if choose threshold value 40% then it will display all the influential documents who has more than 40% along with influential users. It will show Blog Id's of users who influence them more. It displays how many Blog Id's have influential documents. Influential users can be recognized by comments also. Users can give influential comments also after reading all the comments by any users in a specific blog.

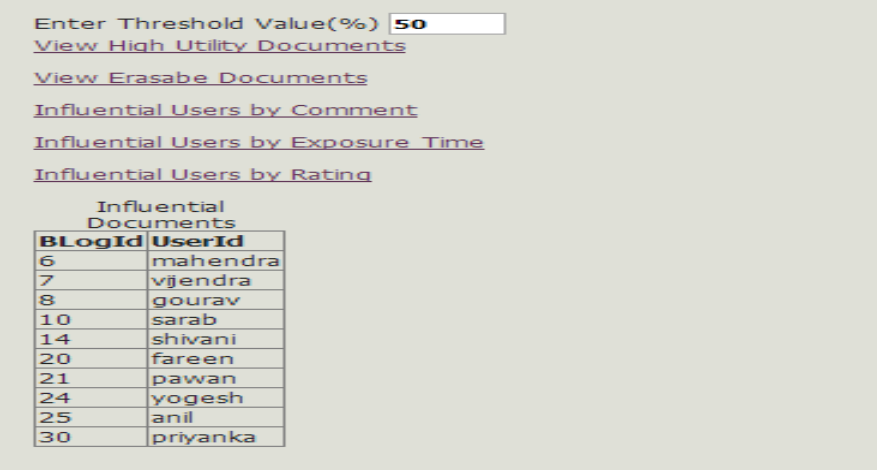


Fig 5.1 (h) Influential users by comment

- **Influential users by exposure time: -** It is used to display influential documents along with influential users. Influential users are those who comments on a specific blog more as compare to other normal users. Exposure time can be calculated by start view time and stop view time of any user related to specific blog. This means how much time spends by a particular user in each blog. Users can spend their time on any blog means he or she can view that blog by just simply reading it. Simple meaning is that if user is spending his valuable time on a particular blog then he is interested on that blog otherwise he is not.

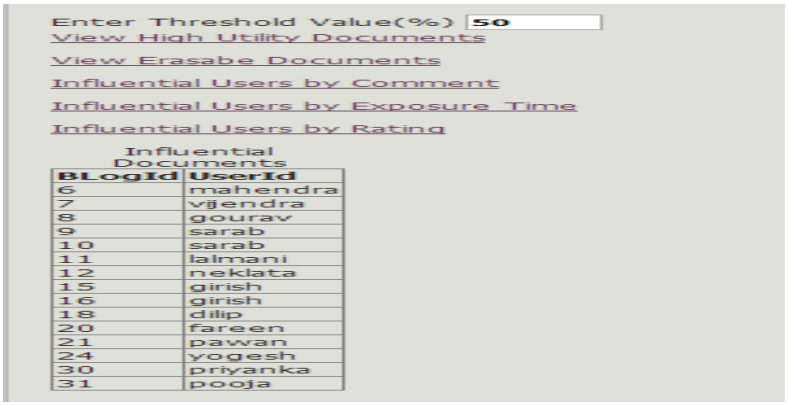


Fig 5.1 (i) Influential users by exposure time

- **Influential users by rating: -** It will display influential documents of blog along with user id. If user chose's 50% of threshold value. Below figure shows the rating of blog id with user id whose threshold value is 50%. Users give rating after reading a particular blog they can give rating starting from 1 to 5, which means 1 is for least and 5 is for more popular blog or more influential blog. 5 rating means user like that blog more as compare to other blogs. If user like that blog they give rating for that specific blog, if they don't like it they will not give any rating, user just simply read that document.

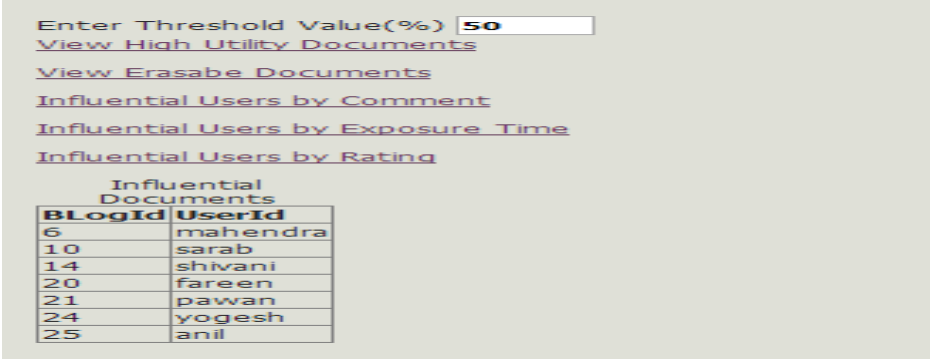


Fig 5.1 (j) Influential users by rating

6. Conclusion

The research work is used to merge two techniques of data mining which is classification and clustering in a blog network. It is used to calculate the influential users in a blog network which motivates the other normal users to rate the specific blog. Classification technique is used to classify the users whether they are active or dormant users and clustering technique is used to create a group of influential users and other normal users which are not influential ones. With the help of this method blog users can be identify by the owner of a specific blog. Some threshold value is there through which highly utility users can be easily recognized and owner can get profit in monetary terms. Some features are taken for identifying the influential users in a blog atmosphere. Due to which quality of a blog could be increase or improve by adding some features on it. The following are the some important features of this research:-

1. Introduces the influential user concept which increases the monetary in business.
2. It is used to finding those users who actually motivates the other users in a blog network.
3. Introduces the influential relationships in a blog network.
4. Identifying the influential users by some features like track back, comments, rating, like.

References

[1]. Apoorva Vikrant Kulkarni et al (2013) "Blog Content and User Engagement-An Insight Using Statistical Analysis", International Journal of Engineering and Technology (IJET) Vol.5, No.3, Jun-Jul 2013
[2]. Eunyoung Moon et al (2010) "A quality Method to Find Influencers Using Similarity Based Approach in the Blogosphere", IEEE International Conference on Social Computing/IEEE International Conference on Privacy, Security, Risk and Trust
[3]. Hsiu-Ju Chen (2009) "Blogger's Social Presence Framing and Blog Visitor's Responses", IEEE Computer Society, Eight IEEE/ACIS International Conference on Computer and Information Science
[4]. Malik Muhammad Saad Missen et al (2010) "Opinion Detection in Blogs: What Still Missing?", International Conference on Advances in Social Networks Analysis and Mining
[5]. Malik Muhammad Saad Missen et al (2009) "Sentence-Level Opinion-Topic Association for Opinion Detection in Blogs", International Conference on Advanced Information Networking and Applications Workshops
[6]. Seung-Hwan Lim et al (2011) "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, September 2011
[7]. Tomas Kuzar et al (2011) "Slovak Blog Clustering Enhanced by Mining the Web Comments", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology
[8]. Azaim Khan et al (2014) "The systematic survey on Influential users in a Blog Network", International Journal of Emerging Trends and Technology in Computer Science Vol. 3, page 237-239
[9]. Data mining concepts and techniques [2nd edition] (jiawei Han and Micheline Kamber)
[10]. http://www.ijcse.com/docs/IJCSE10-0104-51.pdf
[11]. http://www.slideshare.net/p2045i/introduction-to-data-mining
[12]. http://www.fas.org/irp/crs/RL31798.pdf
[13]. http://www.enggjournals.com/ijcse/doc/IJCSE09-01-03-11.pdf