MINING OF HIGH VALUE BLOGOSPHERE USER

Shubham Baronia, Anurag Jain
Radharaman Institute of Technology & Science,
Bhopal, Madhya Pradesh, India

ABSTRACT
Blogosphere is an upcoming field that has motivated users for vital participation. These users not only enhance other people but also can help in the growth of business and business models. The aim is to discuss the content power user determination and then the focus shall be on how to get high utility documents from the blogosphere for the content power users. We shall be presenting our own blogging site and shall maintain the users data for deriving the results from them. The users blogging belong to the college and thus result set generated is almost actual.

INDEX TERMS – *Blogging, blogs, CPU, high utility set, DCP.*

I. INTRODUCTION

The world is growing smaller day by day with the increase in internet and the number of people who access it. From small to big, from highly educated to less educated, from technical to non-technical, almost every genre of people are accessing internet in some way or the other. The internet has almost made a virtual world exist along with the real. People now know each other more through their profile that exists on the various social networking sites than personally.

The internet has also provided the right of expression in true sense. Write comment whatever one wants whether it is any personal thoughts, experiences, and web links. One of the prominent social networking that has gained popularity in recent time is blog .It's like an online dairy which can be about and be used for anything at all, it can be used for news, reviews, products etc. for a business, organization etc. This is great as it helps the user stay in touch with the website with new and up to date information. There was a time when people use to maintain personal journals now blogs are providing the facility to make it public as per the user's wish, it could be accessed by a few or the entire world.

In these social networking sites there are few users who are more active than the others; the third party companies who basically maintain the blog website are utilizing these users to provide more services. These users also promote other users to be active and thereby making the blog more user optimized. The success of the blogosphere depends entirely on the activities of the user. The main activity of any blogging site is maintaining registered users who can post their own documents that they want to share with others. It not only includes writing documents but also reading what others have posted and commenting if they want. One could also share documents posted in other profiles.

All these social networking sites are maintained against advertisements. The companies that maintain these sites are paid on the basis of either maintaining advertisements as one surf through the web pages or on the click action of these advertisements. So in order to earn the site need to keep a track of its different user involvement as the more the user is active the more revenue could be earned. Our objective is to develop and test a methodology to identify power users in blogging networks on the basis of a simple metric of their activity level. The term power user is used to define those users whose content in the network have an influence on other users directly or indirectly thereby promoting the entire activity level of the website.

In the early studies of mining power users within a network topology was used. This is beneficial to compare within the users who exhibit a greater influence than the other. But it is not that effective when used in a blog network.

II. RELATED WORK

As per the earlier studies there are various ways through which we could judge the power users in a network. The idea of identifying power users have come up from the term viral marketing which is an idea that spreads--and an idea that while it is spreading actually helps market your business or cause. Topology based methods were initially used for determining power users, but it is not successful for blog networks as they and hence the complexity of the n/w increases. Methods like linear threshold model which is based on a threshold value given to the user and a weight given to the relationship b/w the users, then independent cascade model which uses probability could also be used for calculating power users. A method that makes use of both the models has also been proposed.

But both has some or the other disadvantage of its own. One counts in the influential factor of all the member accounts which is mostly of no use and for the other giving an appropriate probability value every time is not possible. The two link-based algorithm Page Rank and HITS which were developed for information retrieval could also be used for retrieval of power users in a network.

In order to measure the influential users in blogging n/w, we first create one blog network and study the influence of one user over the other. There could be various patterns of influence, a user may directly or indirectly influence other users. There are various ways in which a single user interacts with the other, they are reading, scraping, commenting or trackbacking. Reading refers to simply opening any other users' document , scraping means copying the document into one's own profile, commenting is writing comments on the document and trackback means writing a new document related to the original document along with it mentioning a link to the original document. Trackback and scrap leads to reproducing the original document which could further be reproduced by others.

Influence relationship between various blog users could either be judged in a static manner by studying their bookmarks or by the more dynamic way of user activities. Since the actions of power users are pretty fast so capturing the dynamic nature of the user would be more appropriate in calculating the influence relationship. The various blogging site provides a facility for the user to keep a track of the blogs of one's interest. Now scrap or trackback are considered dynamic as it is an activity from a document in a blog to another or blog to a blog is a bookmark. When a user performs any sort of action on another user's document it is believed that the former is being influenced by the latter. In order to capture the flow of these user activities a blog network is usually build according to user actions.

In order to sieve in the most influential users we track the activities of the users on various documents of the user and then calculate the content power user by adding the content power of the various documents. For calculating document content power there are two points to consider , one is the direct document content power that is directly accessing a user's document or indirect content power document that stands for accessing the document after being shared or trackbacked by other.

1.1. Document Content Power

Document content power is calculated by weighted frequencies of other users' activities on the particular document. To compute this, a diffusion history table is maintained which stores the various activities on the document. The diffusion history table consists of a user id column than the corresponding document id with the type of activity, the users who are either a member of the user and their corresponding activities on the document.

Once the document content power is calculated the user content power is computed by adding the weights of various documents of a particular user. Usually the DCP increases when a document is exposed for a longer duration. This leads mostly to an older document having more DCP than a newer document. In order to balance this, the inverse of the exposure time of a document is multiplied with the DCP to calculate the UCP.

$$UCP(Ui) = \sum_j IED_{D(i,j)} * DCP(D_{i,j}) \quad (1.a)$$

The above equation is used for calculating the gross user content power.

Here, UCP- User Content Power

IED- inverse of the relative exposure duration

DCP- document content power.

After the UCP of all the users are calculated, the top n users with the highest UCP are selected as the power user of the particular blog network.

Most of the networking sites generate their revenue from advertisements. So, its very important to know for the parties who run these networks who all in their networks are power users in order to utilize them to not only generate revenue but also use them for branding their blog. The more quality content these power users write more activity is generated in the network. So after determining the power, the information is used for advertisement as well as marketing.

Special features could also be provided for the power user group like music skein etc. Some sort of compensation could also be given to the power user group members. Usually the members are aware that they belong to power user group and they need keep on writing quality content in order maintain in that group. Since this method is more based on activities generated in a particular profile for a given span of time, it is extremely important for the power users to keep on maintaining their blogging site to be in the power user group. Even the top ranking users are aware that if they could generate more activities they'll be promoted to the power user group. This way the network could always maintain a handful of power users in order to utilize them for their own benefit as well.

III. PROPOSED METHOD

In order to find out the content power user we initially create our own blog network wherein we maintain the blogging activity of each user. Since our study is based on the utility pattern of each user we keep a track of their actions in a table. The table keeps a record of all the user members, their respective actions in user's profile like comment or scraping etc., the new blogs created by the user and their exposure time and similar activities.

Different weight combinations could be used to determine the content power user, for e.g. If one wants to calculate the power user based on the amount of blogs that are trackbacked then the weight of trackbacking could be given higher priority than weight given to comment and so on. This is important because it's required to determine the users who could contribute to the maximum profit of any website. A threshold value is calculated depending on the DCP of various users. Once this is calculated the users are divided into high utility and low utility group. The benefit of using threshold for calculating the power user group is that the entire amount of calculation reduces even on changing the threshold value or the database.

Initially the methods that were used only comprised of a fixed database. And the entire calculations had to be made every time any sort of updating was done. This problem is removed in our proposed method and any sort of deletion, modification does not burden with extra calculation. Moreover the present day databases are based on "build once mine many" concept that is one can create a database once and then several mining procedures could be implemented for mining. In the real world scenario, especially in blog network where the number of users and their usage pattern keep on changing on a daily basis , it is extremely required to have a method that flexible to this change , moreover the threshold also need to be changed as and when required. On using our proposed version of mining these problems are easily minimized.

Our main aim is to mine frequently appearing itemsets from the database so that it could be put to some use. For this we'll do utility mining wherein we discover the itemsets which is being most influential in the network. Mining high utility itemsets from the databases refers to findingthe itemsets with high utilities. By the term high utility here we mean the power users in the network, i.e. the users who are more active than the others.

We provide a user defined threshold value, any itemset beyond that value will be considered as high utility itemset and any itemset less than the value will come under low utility itemset. In this way the user can distinguish between the power users. Different weights are given to different activities of the user on the basis of that the user content power is calculated.

The following formula is used for the same

$$UCP = 10 * \sum_i^n R(i,n) + 8 * \sum_j^p T(j,p) + 5 * \sum_k^q S(k,q) \qquad (1.b)$$

Where ,

UCP = User Content Power

R= Reading action of the user and the weight used for this is 10

T= Trackback action of the user and the weight used for this is 8

S= Scraping action of the user and the weight used for this is 5

Indirect DCP of Document

TrackId	DCP
3	8
4	8

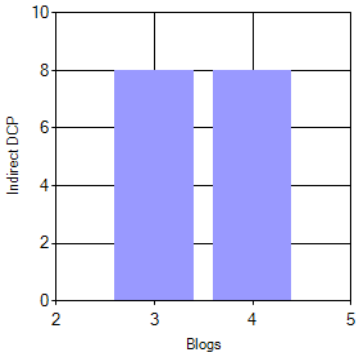

Figure 1. Indirect DCP of Document

Direct DCP of Document

DCP	BlogId
20	3
10	4

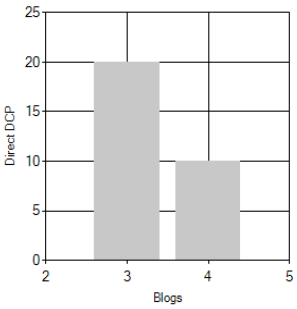

Figure 2. Direct DCP of Document

Document DCP

BlogId	DCP	User
3	28	gunjan
4	18	sameer

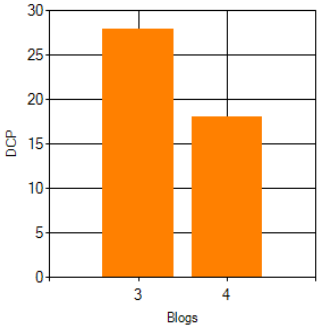

Figure 3. Document DCP

Exposure Time in Seconds

BlogId	UserId	ExpoTime
3	gunjan	125
4	sameer	431

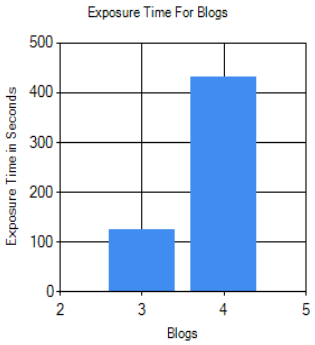

Figure 4. Exposure time for Document

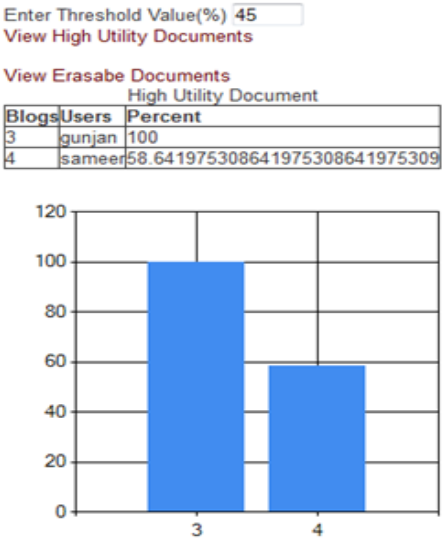

Figure 5. High Utility Document

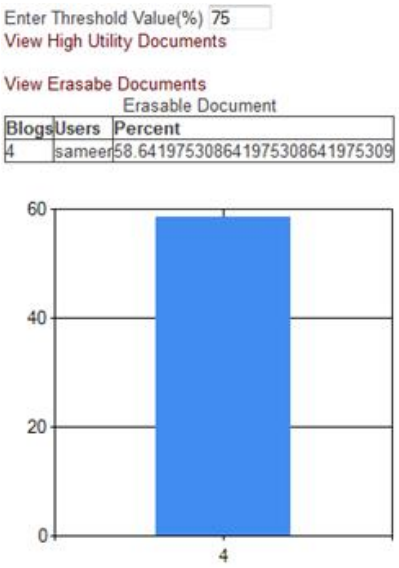

Figure 6. Erasable Document

The main motive of the study is to search all the High utility document and erasable document, on the basis of the threshold value supplied by the user.
The above figure 5 gives the high utility documents which are having more importance than the others, while the figure 6 shows the set of erasable documents that have least importance and are of no use.

IV. CONCLUSION

The paper is used to define the high utility documents with document power users. The determination of power users has been done based on the direct and indirect document content power of all the documents based on the users.

1. Unlike the concept of power users derived from the structural topology and characteristics of social networks, the new concept of "content power user" was proposed, which reflected the influence of a user within a blog network more appropriately.
2. We proposed a method of measuring the content power of a document and that of an individual user. In order to measure content power correctly, we proposed a normalization method based on the exposure time of a document.
3. By soliciting domain experts for user study, we revealed that the proposed method performs best in finding those users who actually contribute to revitalizing the blog network.

4. We proposed several business models that use the concept of CPUs to stimulate activities in a blog network.

In this paper we have proposed new definition of power users. Note that the research reported in this paper, is the process of developing a method to identify power users who persuade the largest number of users to take actions in a blogosphere.

V. FUTURE WORK

- For the future work, there is a lot of interesting research issues related to erasable documents mining.
- First, we will take efforts towards more efficient algorithms by adopting useful ideas from many proposed algorithms of mining frequent patterns.
- Second, there have been some interesting studies at mining maximal frequent, and top-k frequent patterns in recent years.
- Similar to frequent patterns, the extension of erasable documents to these special forms is an interesting topic for future research.

REFERENCES

[1]. Mr. shubham baronia, "Survey Paper Power User Determination", International Ideal Journal For Computer Science & Research Publications, Volume 11, December-2015 267 ISSN 2278 - 7097
[2]. 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, Volume 41, No.5 September* - pp. 853-862, 2011.
[3]. X. Song, Y. Chi, K. Hino, and B. Tseng, "Mining in social networks information flow modeling based on diffusion rate for prediction and ranking," in Proc. Int. Conf. WWW, pp. 191–2000, 2007.
[4]. M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in Proc. ACM Int. Conf. Knowl. Discov. Data Mining, SIGKDD, pp. 61–70, 2002.
[5]. N. Agarwal and H. Liu, Modeling and Data Mining in Blogosphere. San Rafael, CA: Morgan and Claypool, 2009.
[6]. D. Gruhl, R. Guha, D. Nowell, and A. Tomkins, "Information diffusion through blogspace," in Proc. Int. Conf. WWW, pp. 491–501, 2004.

AUTHORS BIOGRAPHY

Shubham Baronia was born in Mandla, India, in 1990. He received the Bachelor of engineering degree from the University of RGPV, Bhopal, in 2011 and he is currently pursuing Master of Technology degree from the University of RGPV, Bhopal, both in computer science engineering. His research interests include web mining.