JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-10, May 2019 E-ISSN: 2347-2693

Algorithm for Mining above and Below Average Utility Blogosphere Users in a Blog Network

Shashank Khare[1*], Sapna Choudhary[2]

[1,2]HOD Deptt. Of Computer Science & Engineering, Shri Ram Group Of Institutions, Jabalpur (M.P), India

Corresponding Author: ishashank29@gmail.com, Tel.: +919424314321

DOI: https://doi.org/10.26438/ijcse/v7si10.2934 | Available online at: www.ijcseonline.org

Abstract— In the past few years weblogs have become a major channel for publishing content over the Internet. With the popularity of social media as a medium to communication, everyone around the world has started using weblogs as part of their communication strategy. However there remains a void of literature on mining information from blogging, and users still do not have a solid understanding of how and why people are using this tool. This is an exploratory study into the world of blogging, and it aims to add some insight as to what is going on in the blogosphere.
As data mining is an important tool for gathering information in any field. Applying this tool in the field of blogosphere is somewhat we are here to discuss about. The thesis aims at gathering information related to the users and documents being published over the internet. We wish to know the documents and the users that are highly active in the blogosphere. This study of our can be conducted by mining high utility documents and users in the blogosphere.
This study we have conducted on a new blogging website created by us by using ASP.Net 4.0 as the tool and then applying the code for mining and reporting of the data.

Keywords—Blog Network, Blogs, Content Power User (CPU), Power User, Document Content Power.

I. INTRODUCTION

During the last years, we are experiencing an unprecedented growth of the use of World Wide Web for both scientific and commercial purposes; especially in the commerce sector. Now people are encouraged to conduct all their transactions online. Today advances in telecommunications networks it made possible the transmission of large amounts of data in a short period of time that resulted in the accumulation of data on the Internet. All these data, which indicate the user's behavior, can be stored in database files specially created for this purpose.

There is a need, therefore, once the data is there to find ways of extracting information out of them; a way to dig into the large files for patterns of users' behavior must be found. The answer to this problem was the development of data mining techniques for blog users, which are the subject of this paper.

Social Networking Websites (SNWs) provides a medium for users to express themselves beyond physical features and labels, to share experiences, discuss interests, and influence one another in a selective network. In addition, social networking Websites are not constrained by the same geographic boundaries as real life networks; allowing users to make and develop relationships with individuals of similar interests around the world. Lastly, SNWs provide an optimal format for users to keep a "personal narrative going" in which they "integrate events which happen in the external world, and sort them into an 'ongoing' story about the self" (Marsh, 2005).

In recent years, SNWs havepopularized the construction and presentation of personal identity online. Social networks provide a platform for communication and the extension of consumer influence. SNWs are "one of the fastest growing arenas of the World Wide Web" and Facebook, MySpace, and LinkedIn are currently among the most visited Websites in the United States of America (Trusov, Bucklin, Pauwels, 2009).

The rest of the paper is organized as follows. The Section 2 contains Previous work. Section 3 provides Scope and Problem Formation. We proposed and algorithm for Above and Below Average Utility Blogosphere Users in Section 4. Experimental results are provided in Section 5. Finally, Section 6 and 7 concludes the paper with Future directions.

II. RELATED WORK

In this section, we review the previous methods to identify power users in a social network, the methods to measure influential power of a user, and two link-based ranking algorithms. We also discussed their applicability to identifying CPUs in a blog network.

Previous studies on social networks have proposed various methods to identify power users in a social network. In particular, the problem of identifying power users in a social network has been studied for a long time in the field of viral marketing. Its primary goal is to determine a small group of customers that can produce the maximum marketing effect.

Due to modern-era individualism and recent developments in Internet technology, self-expression and networking are rapidly mobilizing content development from offline to online. As a result, online social networks have emerged where individuals write documents, exchange information, and form relationships online. The blogosphere is a primary example of such online social networks.Previous paper[1] defined a blog network as the social network established through blogs and their relationships. Examples include facebook.com, myspace.com, linkedin.com, blogger.com, cyworld. com, and blog.naver.com. As the number of blog users increases, companies have become interested in providing products and services that utilize blogs. To ensure the success of a business geared toward blog networks, encouraging users within a blog network to actively utilize blog services is an essential prerequisite (Wegonet, 2004). Within a blog network, there are special users who contribute inducing other users to actively utilize blog services. These users are Content Power Users. If such influential users can be identified, it is possible to establish diverse business policies centered on these users that will increase the usage of blog services more effectively.

In previous studies, in a social network power users are determined which was based on the topology and characteristics of a social network (S. Bagchi, G. Biswas, K. Kawamura, 2000; Karmeshu, D. Goswami, 2008; Y. Lin, 2007; N. Agarwal, H. Liu, 2009; L. Xing 2008). In a blog network, however, topology does not seem to reflect the actual influence relationship between users, and thus topology-based selection may not correctly identify power users. There have been excellent research results on identifying special users with influential power greater than others in social networks (P. Domingos, M. Richardson, 2001, 2002; D. Kempe, J. Kleinberg, E. Tardos, 2003). Typical examples are the independent cascade model, the linear threshold model, the general model that combines the independent cascade model and the linear threshold model together, mining of network values of customers, and information diffusion in blogosphere. They employed different definitions of power users and proposed methods for determining their own power users. If a user has put a trackback link to another user's document or put a comment on someone else's document in a blog network, it is not because she is being influenced by her neighboring users but because she is being influenced by a single user who has that particular document. Thus, the models and approaches which consider the sum of influences received from multiple surrounding users when computing the user's power are inappropriate for explaining influence power in blog networks.

Unlike the linear threshold model[4], the independent cascade model assumes a user is independently influenced by her surrounding users. Therefore, the independent cascade model[9] is more appropriate to model influence power in blog networks. There are two weaknesses in applying the independent cascade model, however. First, accurate assigning of probability between every pair of users is a prerequisite in producing correct power users with this model. It is extremely difficult, however, to compute the assimilation probability accurately in real applications. Second, the power users defined in the independent cascade model do not capture the importance of the quality of contents of CPUs.

The CPUs we are seeking in this research, however, are those users that, within their blogs, maintain good-quality contents and thus induce a large amount of activities of other bloggers. Thus, in order to identify such CPUs properly, we should consider all the activities of various types performed by users in a blog network.

Next chapter will introduce the objectives of the project, as well as the research interests. In addition, the research methodology, which was followed for the project's successful completion will be discussed and analyzed. Finally, there is a layout of its contents.

III. PRESENT WORK

3.1 Scope

For the determination of Content Power Users we start by firstbuilding a blog network composed of bloggers and their actions.

We discuss two ways of constructing a blog network by capturing different influence relationships: one based on bookmarks and the other on user activities. We argue that a user's action on another user's document may be viewed as the former being influenced by the latter and that user actions, in particular reproductive actions like trackbacking, are better suited for capturing the propagation of influence in a blog network. Second, we propose a method to compute the content power of a document. The influence power of a document is greater with more activities of other users on the document and with more activities of other users on the documents that are re-produced from the original document. Therefore, the document content power is computed from both direct and indirect user activities. Third, the content power of a user is computed by adding the content power of all the documents owned by the user. Since a document with longer exposure tends to receive more user activities, the raw value of the document content power is inversely adjusted with exposure duration. Finally, we determine n CPUs in a blog network, by selecting top n users from the highest user content power.

3.2 Problem Formulation

Investigating on the influential relationship in a blog network, we work around the development of blog network. Based on the bookmarks or user actions, two different influence relationships can be defined. First there can be an influence relationship between a user and a blog, via use of bookmarks. But using bookmarks for the construction of blog network may fail to capture the dynamic and fast-changing influence relationships in a blog network.

Second a relationship can exists a user and owner of a document via way of an action performed over the original document of the owner. The actions like trackback or scrap reflect the propagation of influence and can be considered for direct or indirect influence of a document in a blog network.

DCP defined as degree of influence of a document can be calculated by adding up the weighted frequencies of other users' activities induced by that particular document. We need to have the parameters that can help out in degree calculation 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.

3.3 Objectives

Mining of high average utility documents and users as the intention for this work we shall proceed to calculate the following details out of the blogosphere.

1. Calculate the indirect influence of the users.
2. Calculate the direct influence of the uses.
3. Calculate the high average utility users and documents based on a threshold value supplied by the user.
4. Design the network in Petri net (Graphical Tool) or Microsoft Visio.

3.4 Research Methodology

The method used shall be using the following terminology.

Ui represents user i. Di represents the set of documents owned by Ui, and Di,j represents document j of user i. Document Content Power (DCP) is defined as the content power of a document, and DCP(Di,j) represents the DCP of document Di,j . Similarly, user content power (UCP) is defined as the content power of a user, and the UCP(Ui) represents UCP of user i.

Action Type (AT) represents the types of actions (i.e., comment, trackback, and scrap) a user can perform in a blog network. An action of type k is denoted as Ak. When computing the content power of a document, different weights may be assigned depending on the types of actions. The weight for Ak is denoted as .

Ui: User i
Di,j: Document j of User i
Di ={D1, D2, …}: Set of documents owned by Ui.
DCP(Di,j): Document Content Power of Di,j.
UCP(Ui): User Content Power of Ui.
AT = {A1, A2,…}: Types of user actions.
Ak: Action of type k.
: Weight of Ak.

$$DCP\left(D_{i,j}\right) = w_D \times D_{DCP\,(D_{i,j})} + W_I \times I\_DCP(D_{i,j})$$

$$D_{DCP\,(D_{i,j})} = \sum\nolimits_{A_k \in AT} W_{A_k} \times Count(D_{i,j}, A_k)$$

$$Count\left(D_{i,j}, A_k\right) = Frequencies of A_k on D_{i,j}$$

$$I\_DCP\left(D_{i,j}\right) = \sum\nolimits_{j'} DCP(D_{*,j'})$$

$$UCP(U_i) = \sum_i IED_{D_{i,j}} * DCP\left(D_{i,j}\right)$$

The actual work starts over here for mining of high average utility document and user, after supply of threshold value for evaluation.

IV. IMPLEMENTATION

4.1 Construction of a Blog Network

In this section, we investigate influence relationships in a blog network and how to capture them in a blog network representation. A blog user is often provided with a functionality that enables her to keep track of the blogs of her interest, which makes it easy and convenient for her to visit those blogs. Such functionality is called *bookmark*, *blogroll*, or neighbor.

A blog user performs actions on a document in someone else's blog, such as *read*, *comment* , *trackback* and *Scrap*.

In figure 1.1 Read and comment are reading and putting comments on someone else's document, respectively. Trackback is writing a new document related to someone else's document while putting a link to the original document in one's own blog.

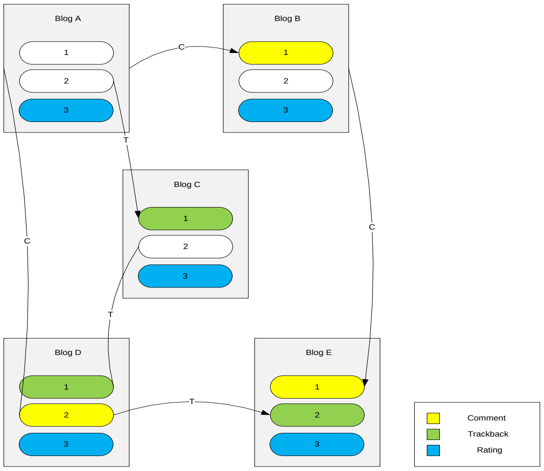


Figure 1.1 Example of Blog Network

4.2 Blogs and Blogging

Simply stated, “blogs are a diary or journal maintained on the Internet by one or more authors or contributors” (Catalano, 2007, p. 248). Blogs can take many different forms and encompass any topic imaginable, from personal to political, from business to religion. Blogs can provide product support and are a gateway for ushering new ideas into the public domain. The authors are called bloggers, and the community in which blogs exist is called the blogosphere. Weblogs, more commonly known as “blogs,” are one of those many new tools and have become a way for people to be able to post anything they desire on the Internet.

Blogs are an extension of websites, and have become commonplace because the majority run on a simple and free format, are easy to use.

Blogging is a tool used in computer-mediated communication that not only allows consumers to learn about companies and products from other consumers, but also allows users to interact directly and create dialogue with other consumers and companies alike. This allows consumers to ask more questions about products and ideas, as well as discuss the benefits or value of a company or product with others that are familiar with it.

4.3 Algorithm

Input:

1. A set of m items I = {i_1, i_2, … , i_j, … , i_m}, each i_j with a measured value p_j, j = 1 to m;
2. A transaction database D = {T_1, T_2, … , T_n}, in which each transaction includes a subset of items with quantities.
3. The minimum average-utility threshold λ.

Output:

1. The minimum average-utility threshold.

Step 1: Create a blog.
Step 2: Perform any action like Read or Comment and trackback which has weight 10 and 8 respectively.
Step 3: Identify the action performed by the users.
Step 4: Compute Document Content Power for each and every blog by $W_D * D\_DCP(D_{i,j}) + W_I * I\_DCP(D_{i,j})$
Step 5: Now compute User Content Power by $UCP(U_i) = \sum_j DCP(D_{i,j})$
Step 6: Enter the minimum support
Step 7: I = {i1, i2, …, im} is a set of items.
Step 8: D = {T1, T2, …, Tn} be a transaction database where each transaction Ti ∈ D is a subset of I.
Step 9: o(ip, Tq), local transaction utility value, represents the quantity or profit of item ip in transaction Tq.
Step 10: s(ip), external utility, is the value associated with item ip in the Frequent Items Table. This value reflects the importance of an item, which is independent of transactions.
Step 11: u(ip, Tq), utility, the quantitative measure of utility for item ip in transaction Tq, is defined as o(ip,Tq) × s(ip).
Step 12: u(X, Tq), utility of an item set X in transaction Tq, is defined as $\sum_{i_p \in X} u(i_p, T_q)$, where X = {i1, i2, …, ik} is a k-item set, X ⊆ Tq and 1≤ k≤ m.
Step 13: u(X), utility of an itemset X, is defined as $\sum_{T_q \in D \wedge X \subseteq T_q} u(X, T_q)$.
Step 14: Generate results based on the results found.
Step 15: Input gain value for mining of erasable itemsets.
Step 16: Compare gain value with gain of retrieved itemsets of high utility, and eliminate those with value less than the given gain value.
Step 17: Generate results for the new set of itemsets.

4.4 Implementation in terms of User Access

Over here we shall discuss about an example how the whole application shall work:

5. Only authenticated users can create a blog or do any of the other activities on the website, so the users need to get registered via registration page and this shall affect the User Table structure, where the information shall be saved.
6. Secondly the users can create a fresh blog or create a blog in reference (i.e. it is having a backtrack link) or create a blog that is having a matter copied from other blog document (i.e. it is considered as a scrap in our terms). All the above documents are having weights in concern to them. Weight of 8 to any trackbacked link and 5 to any scrap. These are indirect weights for the document.
7. Moving on any user who has logged in can now go in for multiple options
   a) Comment on any document which adds weight of 10 per comment per document which is known as Read attribute of the document. This adds a direct weight for the document.
   b) Become a member for the document.
   c) Give rating to the comments which again adds indirect weight to the document. The rating is on scale of 1 to 5 which is considered by averaging them.
   d) Exposure time is one of the important weight for the document, because this value let us know how much time has been devoted by the user on it.

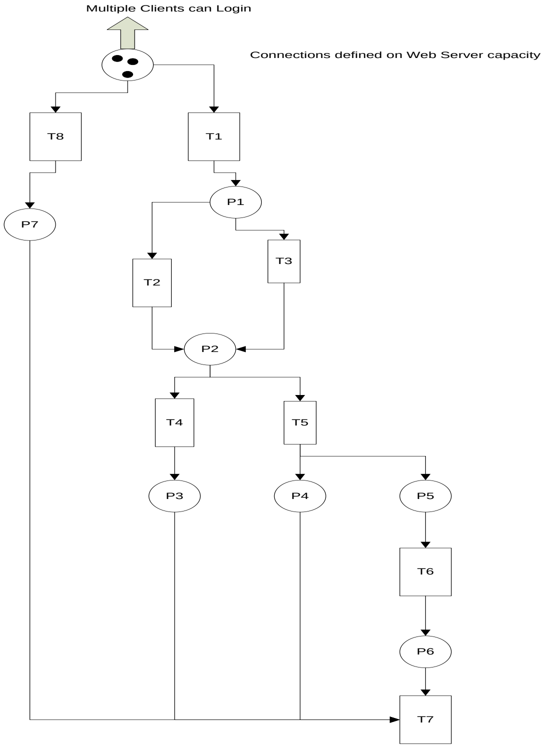

Multiple Clients can Login
Connections defined on Web Server capacity
T8
T1
P1
P7
T3
T2
P2
T4
T5
P3
P4
P5
T6
P6
T7

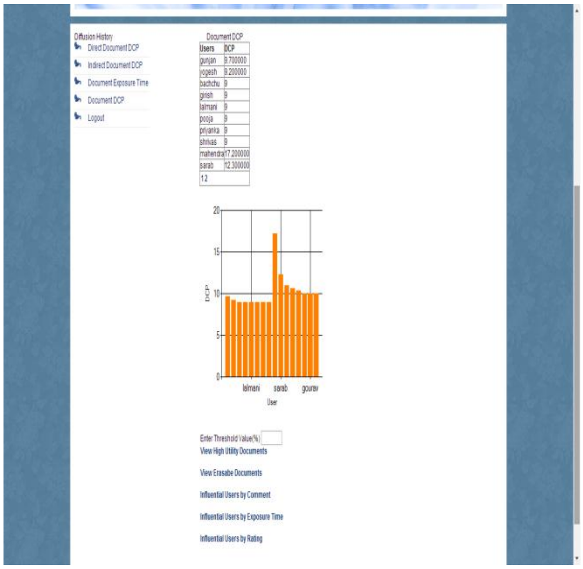
Figure 2 Process Diagram

Table 1: Detailed information about process Diagram

Figure 1. P1 Successful Login	Figure 2. T1 Login Website
Figure 3. P2 Blog Id	Figure 4. T2 View Blog
Figure 5. P3 Increase Expo Time	Figure 6. T3 Create Blog
Figure 7. P4 Add Comment	Figure 8. T4 Exposure time
Figure 9. P5 View / Read Comment	Figure 10. T5 Comments
Figure 11. P6 Add Rating	Figure 12. T6 Rating
Figure 13. P7 View Reports	Figure 14. T7 End
Figure 15.	Figure 16. T8 Admin Login

1. Summing up the weights we get a DCP value for all the documents.
2. Now onwards the high utility and erasable documents can be retrieved from the values, based on the threshold value supplied by the user.

V. RESULT



Document DCP

Users	DCP
gunjan	9.700000
yogesh	9.200000
bachchu	9
girish	9
lalmani	9
pooja	9
priyanka	9
shrivas	9
mahendra	17.200000
sarab	12.300000
12	

Figure 3 A graph showing the high utility documents in respect of the threshold value supplied by the user i.e 45% in this case.

VI. CONCLUSION

The paper proposed an algorithm for High average utility blogs and user in a blog network.

We proposed a normalization method based on the exposure time of a document. By soliciting domain experts for user study, we revealed that the proposed method performs best in finding those users who actually major contribute to revitalizing the blog network.

Today, virtually anyone has the power to share his or her views with the world. Realistically, however, very few bloggers will achieve this. Most will undoubtedly toil in obscurity. Having said this, there is the potential for a star to rise in the blogosphere. Glenn Reynolds of Instapundit.com, for example, was not widely known before he started his blog, and now he is the most famous blogger in the world. Gaining a large audience is the result of incisive commentary, frequent posts, a recognizable persona, and probably, some luck as well.

Concluding on this paper we can see that there can be maintenance of the server and blog documents on the system on having the high utility documents along with the erasable documents; that too on the choice of the user which supplied in the form of threshold value.

VII. FUTURE SCOPE

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]. 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" in September 2011.

[2]. N. Agarwal and H. Liu, Modeling and Data Mining in Blogosphere. San Rafael, CA: Morgan and Claypool, 2009.

[3]. C. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[4]. 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, 2007, pp. 191–200.

[5]. R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in Proc. Int. Conf. Knowl. Discov. Data Mining, ACM SIGKDD, 2006, pp. 611–617.

[6]. D. Gruhl, R. Guha, D. Nowell, and A. Tomkins, "Information diffusion through blogspace," in Proc. Int. Conf. WWW, 2004, pp. 491–501.

[7]. D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in Proc. ACM Int. Conf. Knowl. Discov. Data Mining, SIGKDD, 2003, pp. 137–146.

[8]. M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in Proc. ACM Int. Conf. Knowl. Discov. Data Mining,SIGKDD, 2002, pp. 61–70.

[9]. J. Goldenberg, B. Libai, E. Muller, 2001 "Talk of the network pp 211-223.