

A More Effective Approach For Mining Powerful Users In A Blogosphere

Mrinalini Tiwari^[1], Hare Ram Shah^[2]

Research Scholar^[1], Associate Professor^[2]

Department of Computer Science & Engineering
Gyan Ganga Institute of Technology and Sciences, Jabalpur
Madhya Pradesh - India

ABSTRACT

Blogosphere is a well known term which has its own impact on the Internet. It is a very popular social media consisting of various users all around the world. The research aims at collecting information related to different classes of users and documents published over the blogging network. We primarily focus on finding active users, which are highly motivated towards the use of the blogging network. The main idea behind this research is that the active users can be helpful in promoting businesses. We have applied web data mining strategies on the blogosphere to gather most active users.

Keywords:- Blogging network, Blogosphere, Active users, Web data mining.

I. INTRODUCTION

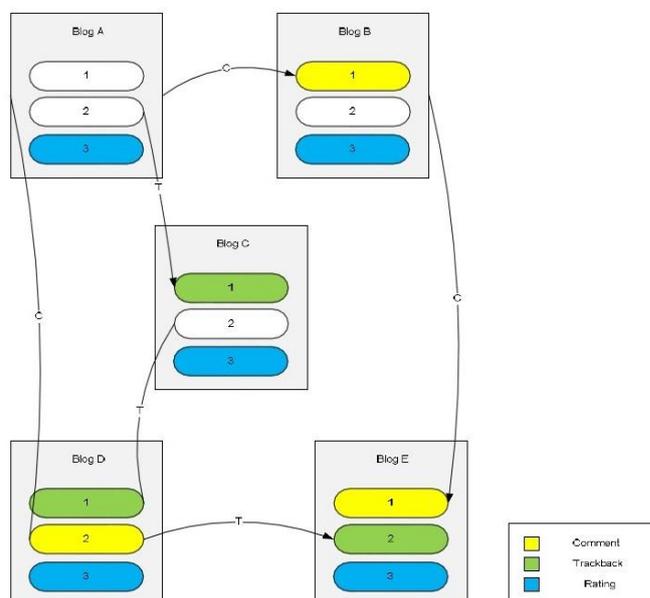
Nowadays, social websites are very much prominent among different classes of users as a mode of communication. Blogosphere is an example of social website. It consists of different blogs written by users. A user can create his/her own blog and publish it over the Internet. In this way, communication between different users can be established through this medium. The blogs of some users are capable to induce other users as well. Such class of users is known as content power users. The subject of this paper is to find those powerful users.

Blogging is an effective medium that allows various users to learn about ideas of other users and also provides a way to interact directly with them. In this way, a business or a product can be promoted through the blog.

A. Creating a blogosphere

We discuss two techniques to construct a blogosphere by capturing different influential aspects: one according to the bookmarks and the other on the basis of different user activities.

Bookmarks let user to keep track of the blogs he is interested in. With the help of bookmarks, he is able to visit the blog whenever he wishes to. The various user activities include reading blog, commenting, trackback, scrap, and rating of comments and so on. The figure shows the relationships between various documents in a blogosphere.



II. PREVIOUS 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 discuss their applicability to identifying CPUs in a blog network.[1]

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.[1]

In early studies, power users in a social network are determined 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. As explained in Section II, however, they have limitations on identifying content power users (CPUs) since they do not properly consider various types of activities performed by users in a blogosphere.

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 neighbouring 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.

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.[1]

III. PROBLEM FORMULATION

We need to discuss two types of relationship between different documents available in the blogosphere. Bookmarks and comments will directly influence the document. The other user activities will also have impact on the original document. These will indirectly influence the document.

We have to take into account both the direct as well as indirect influence on a particular document, thereby calculating the overall document content power. We can say that total document content power of a particular document is the sum of direct as well as indirect content power.

Then, we can calculate high utility documents and users on the basis of a provided threshold value.

IV. IMPLEMENTATION

We have created an artificial blog network and calculated the desired values. The procedure is as follows:

1. The users will get themselves registered on the registration page in order to create new blog or can see other blogs.
2. Secondly the users can create a fresh blog or create a blog in reference (i.e. it is having a backtrace 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, like weight of 8 to any trackbacked link and 5 to any scrap. These are indirect weights for the document.
3. 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 values for the document, because this value let us know how much time has been devoted by the user on it.
4. Summing up the weights we get a DCP value for all the documents.
5. Now onwards the high utility and erasable documents can be retrieved from the values, based on the threshold value supplied by the user.

V. RESULTS

The implementation has been done using Windows 7 as operating system, Visual Studio 2010 as Integrated Development Environment for C# programming and ASP.Net 4.0 as web designing tool, Microsoft SQL Server 2008 as backend for storing data, Internet Information Services used as Web Server for deploying.

Taking in consideration the implementation done by Seung-Hwan Lim, Sang-Wook Kim, Sunju Park, and Joon Ho Lee in their paper, going forward development on the same path, the implementation has been done considering it.

Following the pattern we move ahead and the implementation goes along this way:

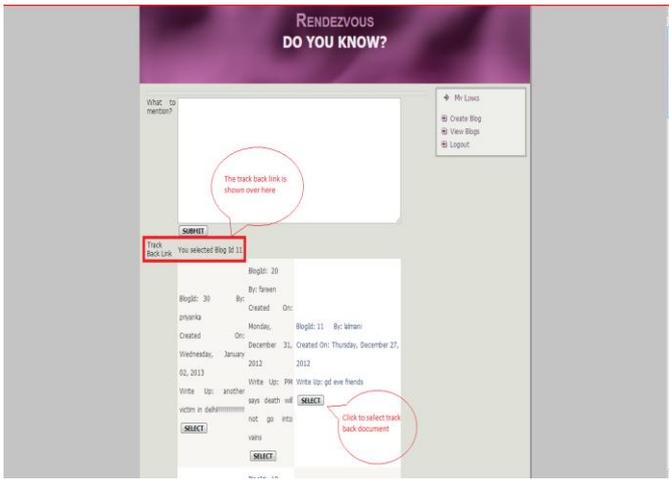


Fig. 1: Screen displaying new blog creation and also the latest blogs created.

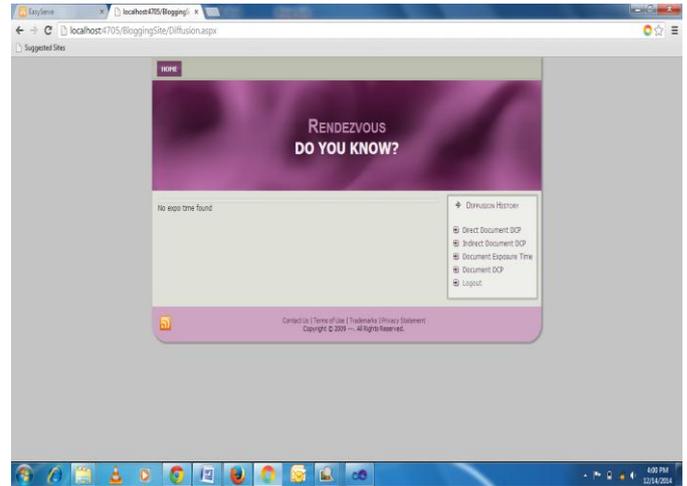


Fig. 4: Diffusion history menu to view different reports

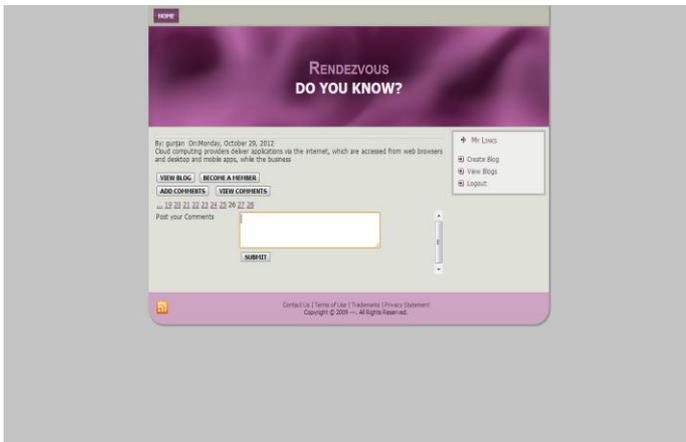


Fig. 2: Screen displaying how the comments will be done.

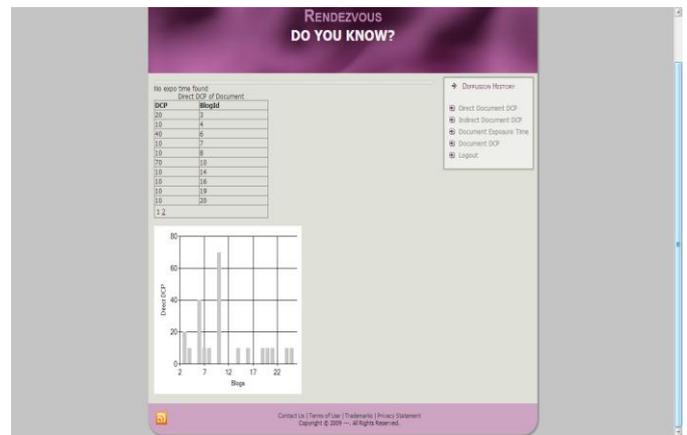


Fig. 5: Figure showing Direct Document DCP values.

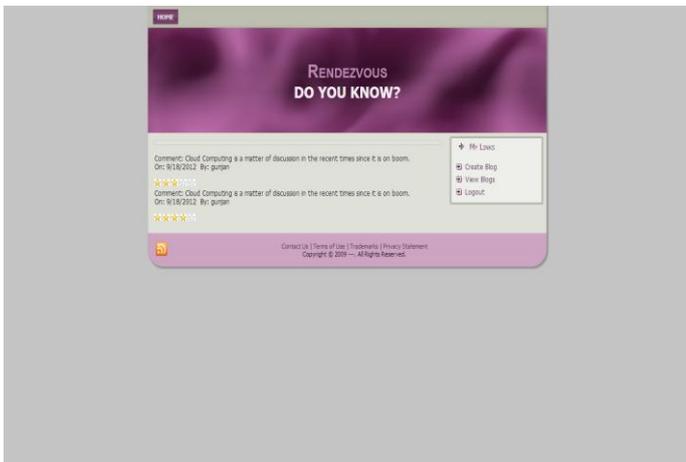


Fig. 3: Screen displaying the comments and lets user to rate comments

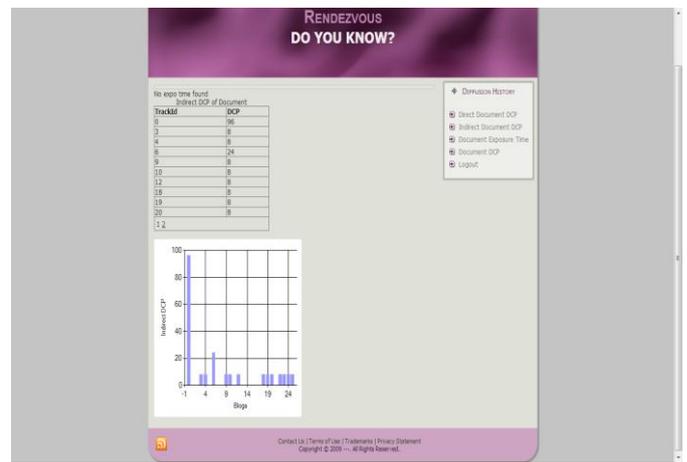


Fig. 6: Figure showing Indirect Direct Document DCP values.

