SOCIAL NETWORK ANALYSIS USING WEB MINING IN BLOGOSPHERE

[1]*AMRUTA S. DULANGE* [2]*Dr. RAJ KULKARNI*
[1]*M.Tech, CSE Dept, Walchand Institute of Technology, Solapur, Maharashtra*
E-mail:amrutadulange2@rediff.com
[2]*Professor*

Abstract: *Blogs are most common medium over web where user posts their opinion. It is considered to be a web space of the users where they share their views, beliefs and other philosophy. The blogs are generally categorized of two types: Itemized blogs, where the user posts his views and opinions against a web news or news item and personal blogs where users posts random topics of their interest under the header of their choice. As more and more number of users publish their data over the web, it becomes significant that Meanings are extracted from blog and they are indexed properly for information retrieval. In this work we develop a crawler to read data from RSS feeds of blogs and save them locally. Finally we apply data mining technique to index the blogs for easy searching and information extraction. The work is divided into two major parts: Extraction of RSS feeds from blogs, indexing and searching. We also show through our performance that the proposed technique is faster than the existing blog indexing techniques.*

1. INTRODUCTION

Blogosphere now are very popular platform for users to post and share articles with each other, no matter traditional blogs or micro-blogs (such as twitter and plurk). Recently, there are numerous data and information aggregrated in blogs and which has becoming a very valuable database [1]. Therefore, blogs now is a new target for different applications, such as marketing [2], crime detection, politic[3], etc. For these applications, the most important issue is about how to identify the opinions in blogs[4]. For the application of politic, it is essential to understand the opinions of citizens for a public policy or a social event. This is also very helpful for the prediction of election results, especially for such countries with very clear political stance [5].

Currently, most applications in this research area is using the techniques of web mining to extract useful information in blogs. Normally, the data from blogs are semi-structured and the process of natural language processing (NLP) is always necessary. However, web mining can only be used to identify groups by using the techniques of classification or clustering, but it can't be used to illustrate the structure and relationship in a group. Social Networks Analysis (SNA) is a methodology which can be used to identify the nodes, the roles and the social relationship in social networks [6]. We believe the results of opinion groups identification will be better, if the technique of web mining and SNA can be combined together. Therefore, the objective of this paper is to propose an approach to combine web mining and SNA together for opinion gropus identification.

The rest of the paper is structured as follow. In section 2, we will provide a brief review about social networks analysis. Collecting the blogs will be proposed in section 3. The system analysis included in section 4. In section 5 methodology. Section 6 describes results and we will conclude this paper in section 7.

2. REVIEW AND RELATED WORKS

The research methodology of social network analysis is developed to understand the relationship between "actors", and the term actor can be a person, an organization, an event or an object [7]. In a social network, each actor is presented as a node and each pair of nodes can be connected by lines to show the relationships. The social network structure graph is a graph that formed by those lines and nodes, and social network analysis is therefore a methodology that used to understand the graph and the relationships and actors in the social network [8].

The most important measurements of SNA include network size, diameter, density, centrality and structure holes [9]. Size is a measurement to measure the amount of nodes or links in a network, and the measurement of diameter is to measure the amount of nodes between two nodes in a network. Density is used to calculate the closeness of a network [10]. These measurements are common used in many social network related researches and will be used in this paper as well.

Traditionally, researches about SNA are mainly focus on small group of actors and are process manually in most cases. However, with the rapid growth of Internet and web techniques, more and more data have been collected and it has become a hard task to process these data by only the mean of manually.[11] Therefore, the scholars of information technology and computer science are starting to devote related researches to deal with these research issues and web mining is consider as the most suitable techniques to analyze the data from web [12].

Web content mining, text mining or natural language processing are very useful techniques that can be used for social network analysis. For example, web content mining can be used to categorize or classify the documents of social networking website, especially for blog or text forum analysis to categorize or classify the articles of blogs. The article categorization is usually the first task for many social networks analyses or applications.

Web usage mining plays an important role in social networks analysis as well. It is useful for the social network analysis of social networks extraction. The usage data and users' communication in social networking website can be transformed to relational data for social-networks construction [12].

3. COLLECTING THE BLOGS

Our next goal was to identify blogs linking to those videos. For that purpose, we identified the most popular YouTube URL (i.e., unique identifier) for each of the 120 viral videos identified in the previous stage. We created scripts which automatically harvested all of the blog posts1 with links to these viral videos on a given day for a given video. The scripts harvested the list of blogs through the Google Blog Search tool. These searches resulted in a dataset of over 13,173 blog posts from 9,765 unique blogs linking to these viral videos during March 2007 and June 2009.

Identifying Four Types Of Blogs

For the purpose of separating our list of blogs into logical types, we gathered monthly unique-visitors traffic data from data service Compete.com. Compete.com tracks viewing data at the site level (i.e., site.domain.com and domain.com), so blogs in folders (i.e., site.domain.com/my blog) and blogs without a full domain match were excluded from our dataset. Where there was no Compete.com data, we assumed the blog had a very low unique-visitors traffic data and kept those as tail blogs. The resulting list contained 3,101 blogs. Figure 1 shows the power-law distribution of these blogs in terms of daily unique-visitors, which also helped us categorize the types of blogs into four types: elite blogs, top-political blogs, top-general blogs and tail blogs. Next, we will give justifications for the existence of each type, define them and explain how they were created.

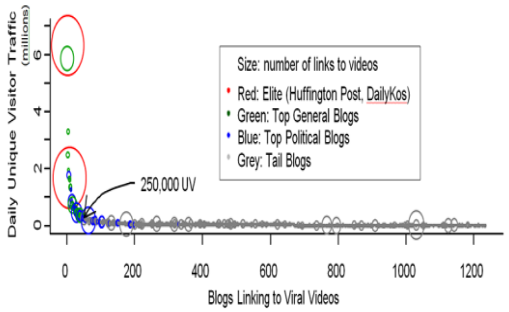

Figure 1: Power-law Distribution of Blogs Linking to Viral Videos

A. Elite Blogs

We found that Huffington Post and Daily Kos were unique blogs in our dataset, in that they have the highest number of blog posts linking to videos (64 and 49 respectively). They are recognized as influential political blogs. Also David Karpf, recognize them as influential political blogs and calls them

"the elite of the elite" [22:40]. These two blogs receive high unique visitors traffic. Furthermore, our statistical analysis showed that differentiating this group is significant.

B. Top-political Blogs

Most scholarship on the blogosphere focuses on this group of elite political blogs. Note that in the literature they are often called elite blogs or A-top blogs, and that in our study we have three types of elites: elite, top-political and top-general blogs. Our set of top-political blogs was drawn from David Karpf‟s Blogosphere Authority Index (BAI) [22,21], which is a measure of a blogs authority. Note that the rankings of blogs may change from week to week. Our set contains the top 25 conservative, and top 25 liberal blogs from the week of August 8th of 2008. Also note that since we place Huffington Post and Daily Kos in our elite group, they have been removed from our list of top-political blogs.

C. Top-general Blogs

Our set of top-general blogs was created by taking all blogs from our dataset (excluding those listed in the top-political and elite blog types) that had more than 250,000 unique visitors as listed by Compete.com. Figure 1 shows that 250,000 unique visitors is around the inflexion point, meaning, this is roughly the point when the curve goes horizontal, and therefore, anything above it seems to be more influential in terms of traffic than the ones below it.

D. Tail Blogs

Every other blog that linked to our viral videos, that is not in the other three types of blogs, is considered a tail-blogs. In other words, tail blogs would represent the blogs of users without high authority.

4. SYSTEM ANALYSIS

A. Present System

Various blog mining techniques are being proposed in the text as summarized in the related work section. The main problem with most of the techniques is that they depend upon the distance analysis and clustering result based on the occurrence of the words. The mining is purely a syntactic outcome of a sentence interpretation does not take into account of other related posts. The techniques have not proposed a clear mechanism of extracting subject and categorization of blogs. In short blog mining is presented as an aggregation result of distance in terms of sentences and not as a natural language processing technique. No past work has defined finite automata for mining, though numerous tree based approaches are proposed. The present system of blog mining technique is broadly categorized into two categories:

1) Technique based on machine learning and

2) Technique based on clustering.

In 1) a machine learning system like support vector machine is trained with known blogs with and without information. Large databases are used as training sample in such techniques. The given blogs are classified into various groups of sentences based on various distance measure by the classifier. The type 2) type of methods depends upon building a decision tree based on the clustering and occurrence of interrelated words and the words that presents the various categories.

B. Proposed System

The system is modeled in two test sets. Firstly we extract the live blogs from the news feeds from various Word Press powered sites. Here the subject matter is considered as the news item itself. The live blogs are extracted and stored offline for analysis. Secondly we consider standard blogs for analysis of the strength of the algorithm and to verify the correctness of the proposed system.

The main stages and functioning of the system is elaborated as bellow.

1. First segment the blogs into sentences and sentences into words. The words are tagged based on word Net tool for sentence segmentation and tagging.
2. Once the words are tagged, find the similarity of the blogs with respect to a specific subject matter based on the tags of the blogs.

3. The similar items to the headings are ranked higher and are sorted at the top in comparison with the other blogs.

4. The high ranked blogs are forward scanned for the indexing.
5. Blogs related to a certain heading and those posses information is now tagged with respective category and information.
6. Based on 1.4, the sentences are weighted from the start to end based on segment fragments as elaborated in 1.5.
7. Based on user query, the tagged blogs are fetched and shown to the user.

5. METHODOLOGY

In this section, we present our methodology for collecting blogs, analyzing their relationships, and presenting analysis results. Our methodology is based on a semi-automated framework that we developed in our previous research (Chau and Xu, forthcoming). Figure 2 presents the framework, which consists of four main modules, namely Blog Spider, Information Extraction, Network Analysis, and Visualization.

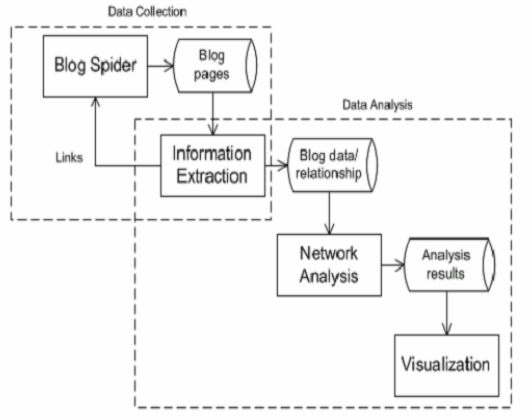

Figure.2. The Framework for Blog Collection and Analysis

The blog spider is designed to download the relevant pages from the blogs of interest in a way similar to general Web fetching. However, instead of following all extracted links, the blog spider should only follow links that are of interest, e.g. links to a group's members or other bloggers. In addition, the spider can use RSS (Really Simple Syndication) and get notification when a blog is updated. However, this is only necessary when monitoring or incremental analysis is desired and is not used in our current study.

After a blog page has been downloaded, it is processed in order to extract useful information from the page. This includes information related to the blog or the blogger such as user profiles and date of creation. This can also include relationship information between two bloggers, such as linkage, commenting, or subscription. Because different blogs may have different formats, it is not a trivial task to extract such information from blogs. Fortunately, some standard information such as name and location are oftentimes put into specific format (e.g. as a sidebar) in large blog hosting sites, and simple rules should suffice. In the current study, a pattern matching approach is employed.

The major module in our framework is network analysis including topological analysis, centrality analysis, and community analysis. We use four statistics that are widely used in topological studies to categorize the extracted network: average shortest path length, efficiency, clustering coefficient, and degree distribution. Average path length is the mean of all-pair shortest paths in a network.

Efficiency is defined as the average of the inverses of shortest path lengths over all pairs of nodes in a network (Crucitti et al. 2003). The most efficient network is a fully connected network and is often called a clique with efficiency value of 1. Clustering coefficient measures how likely nodes in a network form communities. The degree distribution, p(k), is the probability that a node has exactly k links. A random network usually has a small average path length and is more efficient because an arbitrary node can reach any other node in a few steps. Small-world networks usually have significantly higher clustering coefficients than their random network counterparts of equal size. The degree distributions of random networks are bell-shaped Poisson distributions. However, scale-free networks are categorized by power-law degree distributions, which have long flat tails (Barabási and Albert 1999).

The goal of centrality analysis is to identify the key nodes in a network. Three traditional centrality measures can be used: degree, betweenness, and closeness (Freeman 1979). Degree measures how active a particular node is. It is defined as the number of links a node has. In a directed network, the in-degree refers to the number of in-links a node has and the out-degree refers to the number of out-links. Betweenness measures the extent to which a particular node lies between other nodes in a network. The betweenness of a node is defined as the number of geodesics (shortest paths between two nodes) passing through it. Nodes with high betweenness scores often serve as gatekeepers and brokers between different communities. They are important communication channels through which information, goods, and other resources are transmitted or exchanged (Wasserman and Faust 1994). Closeness is the sum of the length of geodesics between a particular node and all the other nodes in a network. A node with low closeness may find it very difficult to communicate with other nodes in the network.

Community analysis is to identify social groups in a network. In SNA a subset of nodes in an un weighted network is considered a community or a social group if nodes in this group have denser links with nodes within the group than with nodes outside of the group (Wasserman and Faust 1994). An un weighted network can be partitioned into groups by maximizing within-group link density while minimizing between-group link density. In this case, groups are densely-knit subsets of the network. Note that community and groups here do not refer to the explicit groups (blogrings). They refer to a subset of nodes that form implicit clusters through various relationships. In these communities, members subscribe to or post comments to each other's blogs frequently even though they may not belong to the same blogrings.

The extracted network and analysis results can be visualized using various types of network layout methods. Two examples are multidimensional scaling (MDS) (Kruskal and Wish 1978) and graph layout approaches (e.g. Davidson and Harel 1996).

The major factor that may affect the scalability of this framework lies in the community analysis part in the network analysis component. As community analysis relies on the clustering of network nodes, the low scalability of the clustering algorithm selected may become the bottleneck of the framework. Fortunately, some very efficient clustering techniques have been developed (e.g. Flake et al. 2000) that can help resolve this problem.

6 .RESULTS AND ANALYSIS

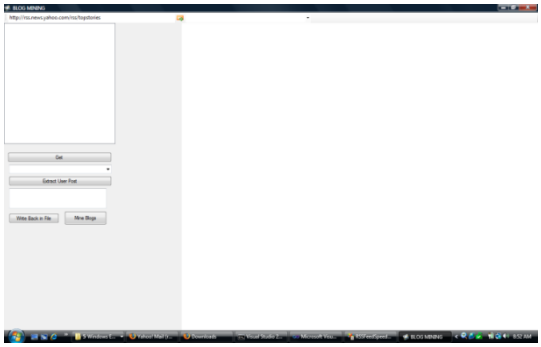

Fig.3.Interface for extracting data from BLOG

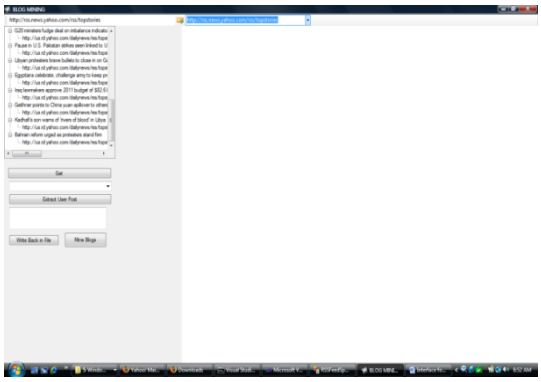

Fig.4 Fetching XML Content from the blog

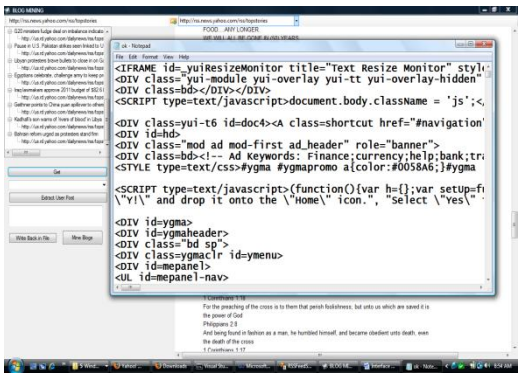

Fig.5.Once a link is selected the main story and the blogs appear

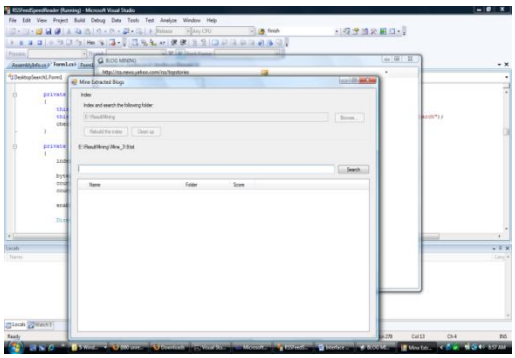

Fig.6.Once Check user Post Button is clicked, all the posts are fetched and are stored locally

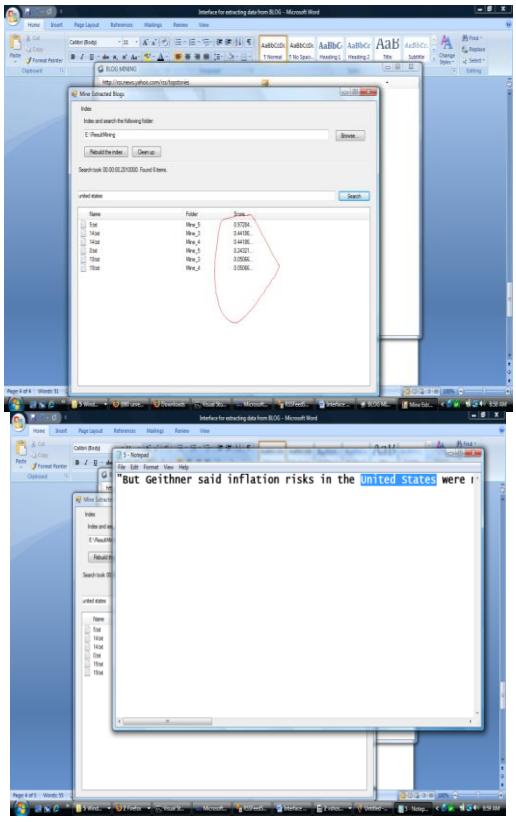

Fig.7.Search Returs fast result

7. CONCLUSION

Blog mining is an important aspect of and the user opinions are presented online, it becomes important for developing tools which can not only extract correlated blogs but also gets an overview of independent and in turn generalized overview of the blogs. Many algorithms are proposed in this direction. Most of these papers are organized to detect the categories in the blogs only and do not present a comprehensive overview of the entire technique of fetching the RSS blog data and analyze them on the fly. In this work we developed an entire lifecycle of fetching and analyzing the blogs for mining information. The technique is based on similarity of the blog with its subject matter and the presence of opinion in such correlated blogs. The result shows a significant similarity with human perception. The technique can be further improved by incorporating machine learning technique with the current algorithm for better learning of the opinions in the blogs.

8. REFERENCES

[1] T. Nanno et al., "Automatically Collecting, Monitoring, and Mining Japanese Weblogs," Proc.13th Int'l Conf. WWW, (WWW 2004), ACM Press, 2004, 320–321.

[2] N. Glance et al., "Analyzing Online Discussion for Marketing Intelligence," Proc. 14th Int'l Conf. WWW (WWW 2005), ACM Press, 2005, pp. 1172–1173.

[3] Geetika T. Lakshmanan IBM T.J. Watson Research Center Martin A. Oberhofer IBM Software Group, Germany Knowledge Discovery in the Blogosphere Approaches and Challenges

[4] M. Kobayashi, K. Takeda, "Information retrieval on the web". ACM Computing Surveys (ACM Press) 32 (2): 144–173, 2000

[5] S. Thies, Content-Interaktionsbeziehungen im Internet. Ausgestaltungund Erfolg, 1st ed., Gabler, 2005

[6] C. Marlow, "Audience, structure and authority in the weblog community," in Proceedings of the International Communication Association Conference, 2004.

[7] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in WSDM '08: Proceedings of the international conference on Web search and web data mining. New York, NY, USA: ACM, 2008, pp. 207–218.

[8] M. Chau and J. XU, "Mining communities and their relationships in blogs: A study of online hate groups," International Journal of Human- Computer Studies, vol. 65, no. 1, pp. 57–70, January 2007.

[9] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "On the bursty evolution of blogspace," World Wide Web, vol. 8, no. 2, pp. 159–178, 2005.

[10] M. Goetz, J. Leskovec, M. Mcglohon, and C. Faloutsos, "Modeling blog dynamics," in International Conference on Weblogs and Social Media, May 2009.