CONTENT BASED RECOMMENDATION AND OPINION RETRIEVAL IN BLOGOSPHERE

Ms.Krutika P.Bang[*]

Prof.A.B.Raut[**]

ABSTRACT

Blogs are most popular way for the peoples to express opinion. Web Blog Mining which is the efficient and effective way of analyzing the sentiments of consumer reviews pertaining to specific products becomes desirable and essential. Blogs provides information but it hard to reach information automatically because blogs are full of un-indexed and unprocessed text that reflects the opinions of people. To evaluate the system, we experiment on specific domain blogs and collect user's feedbacks.. This paper covers the web mining approach about reviewing web blogs and analysis is done from the blogs. Also opinion regarding the particular blog is being obtained by the people.

Keyword: Blogosphere, BRank, Crawler, Parser, Scrapper.

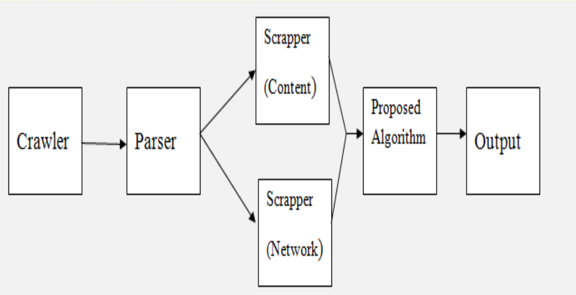
[*] M.E. Second year CSE, H.V.P.M's C.O.E.T., Amravati. S.G.B.A.University (MS), India.

[**] H.V.P.M'S, C.O.E.T., Amravati. S.G.B.A. University (MS), INDIA.

Introduction

In recent years, blogging has become a common way for people to publish content on the Internet. Because blogs are easy to use, people can rapidly share their daily diaries, discuss the latest news, and express their opinions on numerous topics.[1] Web blogs commonly described as blogs are “frequently modified Web pages in which dated entries are listed in reverse chronological sequence”. A blog consists of a title, subscription information, and multiple posts Bloggers are the people who write them use this venue to freely express their opinions and emotions, making blogs increasingly popular. Analyzing the personal entries could even provide opportunities for governments and companies to understand the public in a way that was previously costly or even unavailable. A blog post typically has the post date and text, and might also include hyperlinks, images, and other media. A post might include comments or trackbacks from other bloggers, indicating user interest in that post’s topic [2]. The *blogosphere* is the collection of all blogs and their interconnections, which can serve as a social network as participating bloggers form an online community. Because of the increasing number of blogs and their unique characteristics, developing techniques for searching and mining them has become important. Blog recommendation engines use mined information from diverse sources, including blogs, to make personalized, relevant recommendations to different individuals.Aggregating numerous blogs that offer diverse opinions on the same topic provides valuable collective wisdom and can, for instance, help individuals make a collective judgment about a particular product that they’re considering.

Working:



Crawler : A crawler is a program that visits Web sites and reads their pages and other information in order to create entries for a search engine index. Crawlers are typically programmed to visit sites that have been submitted by their owners as new or updated. Entire sites or specific pages can be selectively visited and indexed.

Parsing : A parser is a software component that takes input data (frequently text) and builds a data structure often some kind of parse tree, abstract syntax tree or other hierarchical structure giving a structural representation of the input, checking for correct syntax in the process.

Scrapper : Web scraping is a computer software technique of extracting information from websites. Usually, such software programs simulate human exploration of the World Wide Web by either implementing low-level Hypertext Transfer Protocol (HTTP), or embedding a fully-fledged web browser. Web scraping is closely related to web indexing, which indexes information on the web using a web crawler and is a universal technique adopted by most search engines.

Algorithm: BRank: A Social-Relation-Based Algorithm

In a social blog network, BRank computes popularity scores to rank a single community's blogs. BRank modifies the surfing probability in PageRank algorithm,

$$P_{A \to B} = \frac{1}{Outdegree\ of\ blogA}, \qquad (1)$$

to consider social relationships in its original random walk model, where the probability for a visitor to go from A to B ($P_{A\to B}$) is decided by the out-degree of A. We adjust the probability that a blog reader will follow a link in blog A to another blog B using a new formula,

$$P_{A \to B} = \frac{R_{A \to B}}{\sum_{X \in O(A)} R_{A \to X}}, \qquad (2)$$

where $O(A)$ means blogs linked by A. In BRank, the probability is determined by the relationship scores ($R_{A\to B}$). In Equation 2, X indicates the blogs to which blog A links.

The relationship score $R_{A\to B}$ represents the relation strength from A to B. It's decided by three factors. The first is the type of blog relationship (comment, trackback, blogroll, or citation).

Different blog relationships are assigned different weights (W_{Rtype}) because they have distinct meanings for a blogger. In our experiments, $W_{comment}$ is set to 0.25 and others are set to 1.

The second factor is the number of the corresponding relationship. Here, we simply use the degree of the number (R_{NRtype}) to express the relationship's strength. Instead of the actual numbers, we use the actual numbers' natural log. The final factor is the blog quality score (BQ_k), which combines the normalized blog features, including the number of subcategories, the number of custom categories, the last article date, the commented post count, the tracked post count, and the average blog/post life cycle.

The blog quality score shows a blog's basic activity. That is, a higher quality score for a blog indicates that the blog's relationships are stronger than ones with a lower score and that it therefore might receive more support from other bloggers. We assume that the probability of a user moving to a blog with a higher quality score is greater than that of moving to others. This quality score is also converted to the natural log value for calculation. The relationship score combines all kinds of relationships between two blogs. The relationship score from blog A to blog K is defined as follows:

$$R_{A\to K} = \sum_{Rtype} W_{Rtype} * RN_{Rtype} * BQ_k. \qquad (3)$$

We compute the relationship score for each directed node pair in the social blog network. A directed node pair could be connected by several support edges, a bidirectional interest edge, or both kinds of edges. We then apply the random walking on the network with the modification of the propagation probability. We can thus define BRank as follows:

$$BRank(A) = \frac{1-d}{n} + d * \sum_{X\in I(A)} BRank(X) * P_{X\to A}; \qquad (4)$$

where $I(A)$ represents the set of blogs linking to A, and d is the damping factor as in the original PageRank algorithm.

Generally, the blogosphere allows anonymous comments and cross-BSP trackbacks. We consider only relationships among users in the same BSP. Therefore, beyond the blog relationships, we consider several countable features that take the effects of anonymous users into account so as to evaluate the importance of different relationships [1].We include the number of posts, the number of all comments, and the number of all trackbacks to represent the anonymous effects on the blog content. Next, we can normalize the BRank scores of blogs in a single BSP. The normalized BRank scores range from 0 to 1. Next, we can apply the global feature to augment the general linking effect in the Web.

Application :

Analysis of Public Awareness

One useful application of blog mining is to evaluate what people say about a company. An effective way to find and analyze blogs gives companies a better understanding of their customers' concernsand helps them evaluate their image, which in turn offers areas of improvement at an early stage for better decision-making, particularly on

customer-related activities.

Companies can also mine blogs about a particular product. For example, using our framework, we developed a preliminary prototype and applied it to the collection and analysis of blogs related to the iPod, a popular portable musicplayer.First, the blog spider connected to hosting sites and blogrings and downloaded the blogs relevant to the iPod, based on their content and groups. The blog parser processed and extracted useful information, such as company names, product names, and opinions. The blog analyzer then reviewed each blog's content and its relevancy to the iPod. Our analysis showed that 49 percent of all the bloggers in this data set didn't mention the word iPod in their blogs, although they had joined blogrings focused on the product. This finding proved that we couldn't use traditional keyword-based retrieval techniques to identify the bloggers who only indicated their preference by joining social communities but not by blogging about it. Finally, the blog visualize presented a high-level display with analysis results. Figure 2 shows a blog visualizer output with many interesting findings—for example, different attitudes toward the iPod don't keep bloggers from interacting with one another. Bloggers with a positive, neutral, or negative attitude toward the product were mixed together in many blog communities. These findings could provide useful information for areas such as online

marketing.

Analysis of Online Social Activities

Bloggers have formed many communities online. Their interests, demographics, opinions, and beliefs make up the focus of these communities, where they share ideas by reading and commenting on each other’s blogs. Unfortunately, inappropriate messages that express hatred or extremism can also easily circulate in blogs. By applying network analysis, we can find these communities and identify the roles bloggers play—namely, leaders, followers, or gatekeepers.We applied our framework to identify and analyze a selected set of 28 racist hate groups (820bloggers) on Xanga, one of the most popular blog-hosting sites. After the blog spider collected entries on these online hate groups’ blogs, the blog content analyzer extracted their content and linkage information (based on membership and subscription information).The blog network analyzer then performed social network analysis on the information, and eventually identified two large communities that consisted of some smaller communities. The blog visualizer generated graphical analysis displays. By showing the structural relationships in the network, such analysis can help identify bloggers who participate in multiple blogrings or subscribe to several other blogs in the community. It can also facilitate analysis for law enforcement officers and social workers who need to study and monitor such activities.

Analysis of Public Opinion

Another important blog mining application is *news monitoring*. People increasingly use blogs to supplement news distribution for several reasons: anyone can update a blog at any time, blogs represent the views of different individuals without filtering (factors such as the target audience’s preferences or political constraints influence mainstream media), and blogs are interactive.Readers can easily post comments to express their views, or they can write their own blogs.

Let’s take the 2005 London bombing as an example. On 7 July 2005, the date the bombing took place, Annie Mole’s blog kept an hourly updateof the aftermath starting at 9:55 a.m. The blog immediately attracted numerous comments from the public about their reactions to the event. Another example is a presidential election. An effective blog mining tool can help candidates better understand what voters like or dislike about

them as well as that about their opponents. For example, BlogPulse's Trends Search (www.blogpulse.com/trend) shows users
a term's frequency in blogs over a six-month
timeframe. Figure 3 shows search results for
two Taiwanese politicians, Chen Shui-bian and Ma Ying-jeou, from September 2006 to February 2007. Chen and Ma, the heads of Taiwan's two largest political parties, were both involved in scandals during that period. You can visualize the progress of the incidents in the trend search result. Most of the frequency spikes during the two terms (in Chinese) reflected major events in the scandals.

Conclusion :

Thus by applying this algorithm we can obtained the blogs as an output which is most popular amongst its category and interested people in the particular domain with get popular blog as an recommended blog.

References:

[1] Chih-Lu Lin and Hung-Yu Kao "Blog Popularity Mining Using Social Interconnection Analysis" *National Cheng Kung University, Taiwan.*

[2] Michael Chau, Porsche Lam, and Boby Shiu, *University of Hong Kong*Jennifer Xu, *Bentley College* Jinwei Cao, *University of Delaware "Blog mining framework "* .

[3]A. Qamra, B. Tseng, and E.Y. Chang, "Mining BlogStories Using Community-Based and Temporal Clustering," *Proc. 15th ACM Int'l Conf. Information and Knowledge Management* (CIKM 2006), ACM Press,2006, pp. 58–67.

[4] K. Fujimura, T. Inoue, and M. Sugisaki, "The EigenRumor Algorithm for Ranking Blogs," *Trusting Agents for Trusting Electronic Societies*, LNCS 3577, Springer, 2005, pp. 59–74..

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

[6] Y. Fu et al., "Finding Experts Using Social Network Analysis," *Proc. Int'l Conf. Web Intelligence*, 2007, pp. 77–80.

[7] R. Blood, R., "How Blogging Software Reshapes theOnline Community," *Comm. ACM*, vol. 47, no. 12,2004, pp. 53–55.

[8] R. Kumar et al., "Trawling the Web for EmergingCybercommunities," *Computer Networks*, vol. 31, nos.11–16, 1999, pp. 1481–1493.

[9] S. Baker and H. Green,. "Blogs Will Change Your Business," *Business Week*, 2 May 2005

[10] M. Chau and H. Chen, "Personalized and Focused Web Spiders," *Web Intelligence*, eds., N. Zhong, J. Liu, and Y. Yao, eds., Springer-Verlag, 2003.

[11] B.A. Nardi et al., "Why We Blog," *Comm. ACM*, vol. 47, no. 12, 2004, pp. 41–46.

[12] N.Agarwal and H. Liu, "Blogosphere: Research Issues, Tools, and Applications," *ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) Explorations*, vol. 10, no. 1, 2008, pp. 18–31.

[13]Y-R. Lin et al.,"Blog Community Discovery and Evolution Based on Mutual Awareness Expansion," *Proc. Conf. Web Intelligence*, ACM Press, 2007, pp. 48–56.

[14] A. Chin and M. Chignell, "A Social Hypertext Model for Finding Community in Blogs," *Proc. 17th Conf. Hypertext and Hypermedia*, ACM Press, 2006, pp. 11–22.

[15] D. Gruhl et al., "The Predictive Power of Online Chatter," *Proc. Int'l Conf. Knowledge Discovery and Data Mining*, ACM Press, 2005, pp. 78–87.