International Journal of Research in Advent Technology, Vol.2, No.4, April 2014
E-ISSN: 2321-9637

Content Based Recommendation of Blogs

Ms.Krutika P.Bang [*1] Prof.A.B.Raut *[2]
1. M.E. Second year CSE, H.V.P.M's C.O.E.T., S.G.B.A.University Amravati.(MS), India.
2. H.V.P.M'S, C.O.E.T., S.G.B.A. University Amravati. (MS), India.
krutikabang@gmail.com[1]

Abstract- Now a day's social media plays very important role in various domains. There are number of resources available on the Internet to express the opinions, ideas emotion and interests. 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.

Keyword: Blogosphere, BRank, Crawler, Parser, Scrapper

1. 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. Given this convenient platform, the number of blogs is increasing at a dramatic rate.[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 that display in descending order by publication date. 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. In addition, bloggers can add their favorite blogs to a blogroll, which typically appears as a list of links on a blog's main page [2]. Also, hyperlinks contained within a blog post give additional information for readers who would like to read related news or blog posts.

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. Blogs act as rich sources of knowledge that can serve a variety of purposes.

Because of the increasing number of blogs and their unique characteristics, developing techniques for searching and mining them has become important. Individuals can use mined information from blogs to determine topics that are popular at a particular point in time. Blog recommendation engines use mined information from diverse sources, including blogs, to make personalized, relevant recommendations to different individuals. Companies can use knowledge discovered in blogs to profile consumer preferences and obtain direct feedback about products through blog-style product reviews. 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. Analytical tools applied to mine blogs for commercially available products can be helpful in indicating sales volumes and predicting market trends.

2. WORKING

The framework or working is as shown in figure1.

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. Crawlers apparently gained the name because they crawl through a site a page at a time, following the links to other pages on the site until all pages have been read.

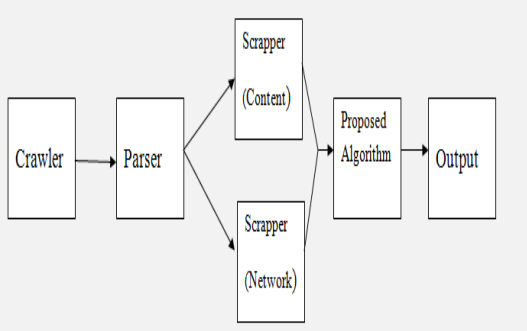

Figure 1. Framework for blog mining

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. The parser is often preceded by a separate lexical analyser, which creates tokens from the sequence of input characters; alternatively, these can be combined in scanner less parsing. Parsing is complementary to templating, which produces formatted *output.* These may be applied to different domains, but often appear together, such as the input (front end parsing) and output (back end code generation) stages of a compiler.

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, such as Internet Explorer or Mozilla Firefox.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. In contrast, web scraping focuses more on the transformation of unstructured data on the web, typically in HTML format, into structured data that can be stored and analyzed in a central local database or spreadsheet. Web scraping is also related to web automation, which simulates human browsing using computer software. Uses of web scraping include online price comparison, contact scraping, weather data monitoring, website change detection, research, web mashup and web data integration.

3. 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,

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

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,

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

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:

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

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:

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

(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. Given the lack of identification mapping between BSPs, there's no effective and trustworthy method that considers the blog interaction for comparing blogs between different BSPs in a global view. We thus 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 consider anonymous effects in the features. 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.

4. 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 KongJennifer 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.

[16] T. Lento et al., "The Ties that Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop Weblogging System," Proc. Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, ACM Press, 2006;www.blogpulse.com/www2006-workshop/papers/Lento-Welser-Gu-Smith-TiesThatBlog.pdf

[17] N. Bansal et al., "Seeking Stable Clusters in the Blogosphere," Proc. Very Large Databases Conf., ACM Press, 2007, pp. 806–817.