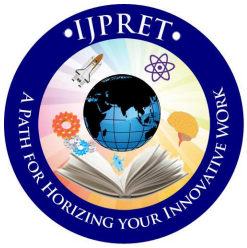
Research Article *ISSN: 2319-507X*
Krutika Bang, IJPRET, 2013; Volume 1(8):167-176 *IJPRET*



INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

WEB BLOG MINING SURVEY

KRUTIKA P. BANG[1], PROF. A. B. RAUT[*2],

1. M.E. First year CSE, H.V.P.M's C.O.E.T., Amravati. S.G.B.A.University, India.
2. H.V.P.M'S, C.O.E.T., Amravati. S.G.B.A. University (MS), INDIA.

Accepted Date:

27/02/2013

Publish Date:

01/04/2013

Keywords

Web Blog,
Knowledge discovery task,
Clustering,
Matrix Factorization,
Ranking.

Corresponding Author

Ms. Krutika P. Bang

Abstract

Knowledge discovery in blogs is different from knowledge discovery in areas such as databases or Web documents due to blogs' unique characteristics, which introduce additional mining challenges. Although researchers have investigated several techniques to address different aspects of blog discovery, no comparisons among key knowledge discovery techniques for blogs exist. This article examines three prominent techniques that are frequently applied to discovery in blogs — clustering, matrix decomposition, and ranking. We are comparing them in terms of effectiveness in combating present challenges and their ability to accomplish challenging tasks required for effective blog mining.

INTRODUCTION

Web Blogs—commonly described as blogs—are "frequently modified Web pages in which dated entries are listed in reverse chronological sequence". Bloggers—the people who write them—use this venue to freely express their opinions and emotions, making blogs increasingly popular. Analyzing these personal entries could even provide opportunities for governments and companies to understand the public in a way that was previously costly or even unavailable. Although the blogosphere contains a lot of useful information, the data is noisy because blog entries are unstructured and might cover a wide variety of topics. By analyzing the freely expressed opinions of bloggers via blog mining, marketers, for example, can get closer to customers and learn more about their opinions on certain products, companies, or political issues. However, because so many blogs exist, manually monitoring and analyzing them is a labor-intensive and time-consuming task. In addition, we can apply knowledge discovery algorithms to determine why such topics are popular and categorize them according to blogger profiles and communities. Blog recommendation engines, such as the one built in to Google Reader (www.google.com/ reader), 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. Several commercial Web sites, such as Blog pulse (www.blogpulse.com),Technorati(www.technorati.com), are available for mining and analyzing blog content. They provide services that include searching blogs based on query keywords, ranking blogs according to popularity, and identifying trends in keywords seen in the blogosphere. However, because blogs are very dynamic, we can't easily apply traditional Web mining techniques to them.

I. THE BLOG MINING FRAMEWORK

Blog mining is an important way for people to extract useful information. As discussed

earlier, blogs are very dynamic, so it isn't as straight forward to apply traditional Web mining techniques to them. However, we've created a general framework for different tasks.

This framework consists of a blog spider, a blog parser, a blog content analyzer, a blog network analyzer, and a blog visualizer.

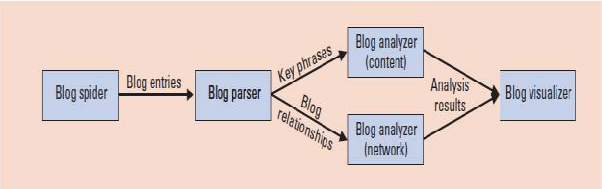


Fig.1 A general blog mining framework

A. Blog Spider

Users can focus on certain types of blog content, but they can't monitor thousands, much less millions, of blogs simultaneously. However, blog *spiders* can simplify this task by monitoring and downloading content from multiple blog hosting sites. Blog spiders are similar to standard Web page spiders in most aspects, except that they must be more timely. Because blogs update frequently, a blog spider must find and download the latest to the hour or even the minute. However, it's often difficult and costly to set up a system to store and monitor the numerous blogs online. An alternative is to connect to popular blog search engines such as Techno rati (www.technorati.com*)*, Google Blog Search (www.blogsearch.google.com), and Blog Pulse (www.blobpulse.com), perform a "meta search," and then combine the results.

B. Blog Parser

A blog *parser* extracts information from blogs, including names of people, products, and organizations. It also includes other patterns, such as dates, times, number expressions, dollar amounts, email addresses, and URLs. Developers can create tools to extract information from blogs based on traditional Web page word segmentation tools such as mutual information, decision trees, or neural networks. However, writing conventions in blogs differ from those found in traditional Web pages. Bloggers often write in a rambling and unstructured narrative style.

C. Blog Analyzers

We can further analyze extracted key phrases by using standard text mining techniques, such as classification and clustering. Blogs can skew positively or negatively toward a product, depending on their content or the bloggers' personal opinions. A blog analyzer associates a phrase that expresses a positive attitude. The blog analyzer then uses these vectors to classify and cluster the blogs into meaningful categories. Blog analyzers can classify blogs that describe a new product into three categories: positive, negative, and neutral. They can use standard models—support vector machines, feed forward/back propagation neural networks, or naïve Bayesian classifiers—to conduct classifications. Because blogs differ from traditional Web pages, blog analyzers can also use other features—comment content, blogger profiles, and links to other blogs—in addition to term features to conduct classifications. For clustering, blog analyzers can group blogs into different categories based on their features. They can measure content similarity between blog entries based on a *similarity score* such as the cosine product between the term features and the additional features such as those used in classification. The blog analyzer is also capable of analyzing the network relationships among bloggers.

D. Blog Visualizer

A blog *visualizer* presents content and network analysis results to users—for example, the use of folders or map displays to present classification and clustering results so that users can explore blogs related to their areas of interest.

Blog visualizers can display the relationships among bloggers in two dimensions with network display techniques. Through these techniques, users can easily identify relevant blogs, important communities, and key bloggers in anet work.

II. KNOWLEDGE DISCOVERY TASK

In addition to addressing the challenges inherent to blogs, relevant literature indicates that knowledge discovery techniques must accomplish particularly challenging *tasks* to accurately and effectively discover knowledge.

A. Visualization

One general knowledge discovery challenge is quickly summarizing discovered results in a concise, easy-to-understand, intuitive format. Knowledge discovery algorithms output doesn't necessarily meet these criteria and could be very large and convoluted. So, researchers developed visualization techniques to better convey these algorithms results.

B. Identifying Authoritative and Reliable Sources

Measuring blog participants authority is becoming increasingly important, particularly when professional application areas are involved. Authoritative bloggers might be experts on a particular subject or persuasive forces that influence others. Weighing blog posts with respect to the author's authority is a significant challenge for knowledge discovery algorithms & verifying the authenticity of data shared on blogs. Information frequently flows from blog to blog, making it difficult to track the information's origin, provenance, or credibility. Discovery algorithms that incorporate techniques such as pattern mining to discover information flows between blogs can use such patterns to give users more reliable data.

C. Accommodating Multiple, Diverse Goals

Given the variety of uses for knowledge discovered from blogs, a particular discovery technique should be capable of discovering knowledge in multiple facets. For instance, if we can use the same technique for topic detection and trend analysis, this technique would be more widely applicable to a large user base.

D. Using Structured and Unstructured Blog Content

Tagging is a popular phenomenon in blogs. Users tend to employ descriptive tags to annotate the blog content they're interested in. In addition to mining unstructured text in blogs, discovery algorithms could leverage structured information such as tags to discover social interests, for instance.

III. KNOWLEDGE DISCOVERY STRATEGIES

We examined three of the most popular strategies for discovering knowledge in blogs as shown in figure 2:

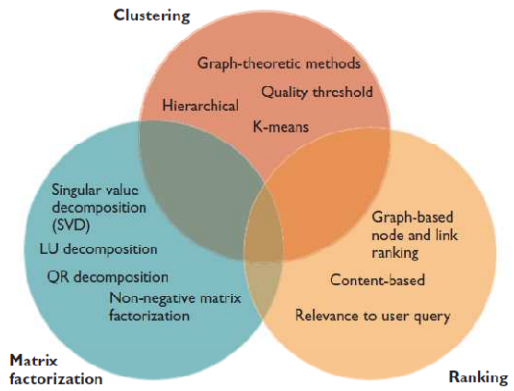


Fig 2: *Strategies for knowledge discovery in blogs.*

A. Clustering

Clustering is one of the most common machine learning techniques researchers have applied to knowledge discovery in blogs. Clustering assigns a set of observations to subsets, referred to as clusters, such that observations in the same cluster are similar according to prespecified criteria. Data clustering algorithms can be hierarchical or partitional. K-means clustering and quality threshold (QT) clustering are both partitional clustering algorithms. Clustering algorithms might require users to specify the number of clusters the algorithm should produce in the input data set. An important step in clustering is to select a distance measure, which will determine how the algorithm calculates two elements similarity. This influences cluster shape because some elements might be close to one another according to one distance and farther away according to another. Once the algorithm has established clusters, another important step in clustering is to determine the membership of newly arrived data into preexisting clusters.

Recently, co-clustering algorithms have emerged that cluster along both rows and columns. For example, a user might be interested in finding similar documents and their interplay with word clusters. When dealing with sparse and high-dimensional data, co clustering is scalable to large matrices.

B. Matrix Factorization

Matrix factorization is another common tool researcher's use for knowledge discovery in blogs. This technique involves

decomposing a matrix into some canonical form. Many different matrix decompositions exist, such as LU decomposition (which writes a matrix as the product of a *lower* triangular matrix and an *upper* triangular matrix), singular value decomposition (SVD), Cholesky decomposition, and QR decomposition, and each is useful for particular problems.

Finding the appropriate interpretation for the singular vectors isn't always easy in practice. Finally, updating SVD results can be difficult if the graph evolves over time. Researchers have developed techniques, such as the Colibri methods, to address SVD's challenges. However, their effectiveness for knowledge discovery in blogs is still undetermined.

C. Ranking

A user might choose to drive knowledge discovery in blogs by specifying criteria for ranking retrieved blog entries. Ranking blogs is quite similar to ranking Web pages. Page Rank and Hyper-Text Induced Topic Selection (HITS) are two popular techniques for Web page ranking that exploit the link structure between such pages. These algorithms focus on a directed graph setting that describes resources via nodes and hyperlinks. Link-popularity-based algorithms, however, might not work well for blog mining because blog pages aren't well linked and bloggers might try to exploit such a system to boost their rank. This observation has led to several new blog ranking techniques that exploit both links and content for ranking. A user might want to retrieve blog entries ranked according to opinions on a certain topic expressed in those entries. Similarly, the influence Rank algorithm enables blog ranking according to how influential each one is compared to others as well as the originality of the opinions the blogs specify. In the context of blogs, researchers have also explored link-popularity-based graph algorithms that use techniques such as random walks and random sampling combined with ranking for both static and dynamically changing graphs.

IV. APPLICATIONS

A. 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 companiesa better understanding of their customers' concerns and 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. First, the blog spider connected to hosting sites and blog rings and downloaded the blogs, 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 .rings focused on the product. This finding proved 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. 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.

B. 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 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 blog rings 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.

C. Analysis of Public Opinion

Another important blog mining applications *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.

V. CONCLUSION

Despite the extensive recent activity on knowledge discovery in blogs, more research needs to occur before we have algorithms that are scalable, accurate, and robust, and provide interpretable results. Blog data is no longer just numerical or discrete. Because bloggers now express themselves through videos, photos, and tweets in addition to text-based posts, algorithms must combine mining capabilities for different social media to effectively mine the blogosphere.

REFERENCES

1. N. Agrawal 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.

2. 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.

3. K.C. Sia et al., "Efficient Computation of Personal Aggregate Queries on Blogs," Proc. Knowledge Discovery and Data Mining Conf., ACM Press, 2008, pp. 632–640.

4. H. Qian and C.R. Scott, "Anonymity and Self-Disclosure on Weblogs," J. Computer-Mediated Comm., vol.12, no. 4, p. 1. 1

5. B. Nardi et al., "Why We Blog," Comm. ACM, vol. 47, no. 12, 2004, pp. 41–46.

6. R. Blood, R., "How Blogging Software Reshapes the Online Community," Comm. ACM, vol. 47, no. 12, 2004, pp. 53–55.

7. R. Kumar et al., "Trawling the Web for Emerging Cyber communities," Computer Networks, vol. 31, nos.11–16, 1999, pp. 1481–1493.

8. S. Baker and H. Green,. "Blogs Will Change Your Business," Business Week, 2 May 2005, pp. 44–53.