A Blog Mining Framework

Michael Chau, Porsche Lam, and Boby Shiu, *University of Hong Kong*
Jennifer Xu, *Bentley College*
Jinwei Cao, *University of Delaware*

New blogs appear every day, and a lot of their content is useful for applications in various domains. The authors propose a framework for mining blogs to gather this information automatically.

Weblogs—commonly described as blogs—are "frequently modified Web pages in which dated entries are listed in reverse chronological sequence."[1] 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. To mine the gold nuggets, you need the right tools. 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.[2] However, because so many blogs exist, manually monitoring and analyzing them is a labor-intensive and time-consuming task.

Intuitively, you could apply existing text and Web mining techniques to blog analysis and mining. But, because many challenges exist, you can't directly apply these techniques. First, bloggers update their work much more frequently than Web masters update traditional Web pages—often daily or even hourly.[3] Second, bloggers cover very diverse topics, so maybe only one paragraph in a particular entry could relate to someone's topic of interest—for example, a product being analyzed. In addition, blog-searching technologies aren't as effective as those for general-purpose Web searching. Finally, blog and Web page characteristics differ so much that you must use different mining techniques for each. For example, you can apply structure mining techniques on hyperlinks between general Web

 Published by the IEEE Computer Society

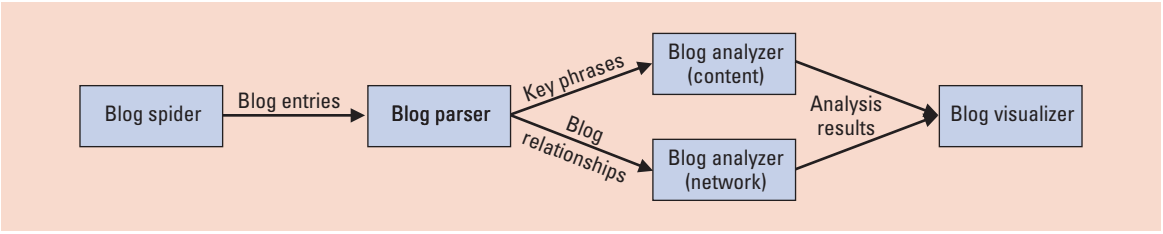

Figure 1. Structure. A general blog mining framework consisting of five components that collect, parse, analyze, and visualize blogs.

pages, but a hyperlink isn't the only way to link blogs—they can also be linked via comments or subscriptions to other blogs.[4]

Tomoyuki Nanno and his colleagues present an architecture for collecting and monitoring blogs in Japanese.[5] Other researchers have studied how marketers use text mining techniques to analyze customers' opinions and reviews on the Web.[2] However, researchers created these systems for blog collection or Web page text mining only, so you can't apply them directly to blog mining applications. In this article, we study the problem of blog mining and discuss its applications in several areas. We propose a framework along with some examples on how blog mining can help in business, management, and social studies.

The Blogosphere

Blogs are typically classified according to their purpose, but five major motivations typically drive blogging: documenting the blogger's life, providing commentary and opinions, expressing deeply felt emotions, articulating ideas through writing, and forming and maintaining community forums.[6] Blogs that serve as a diary fall into a type called *personal* blogs. Commentary and opinion blogs are usually called *issues* blogs, as most of them focus on discussing and debating current events, and blogs that articulate ideas through writing or that serve as community forums are called *topical* blogs.

Bloggers can easily link to other blogs using comments, hyperlinks, blogrolls, or TrackBacks[7]—these technologies let them interact with their readers and form virtual communities in the blogosphere, which is similar to how people form other Web communities.[8] Really Simple Syndication (RSS) is another important feature that bloggers use. With RSS, a user can subscribe to certain blogs or keywords and then receive all the relevant items at a single destination. The user can just use RSS, or other aggregators, to share the latest blog headlines or full text without having to periodically monitor for updates.[9]

The 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 straightforward to apply traditional Web mining techniques to them. However, we've created a general framework for different tasks (see Figure 1). This framework consists of a blog spider, a blog parser, a blog content analyzer, a blog network analyzer, and a blog visualizer.

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,[10] 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. Many blog search engines use spiders that depend heavily on RSS feeds. 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 Technorati (www.technorati.com), Google Blog Search (www.blogsearch.google.com), and BlogPulse (www.blobpulse.com), perform a "meta search," and then combine the results.

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

Table 1. Potential applications of blog mining.

Domain	Description	Example
Business	Assess a company's image strength or customer product reviews to identify virtual communities	Consumer products and movies
Politics	Monitor public opinion about political candidates	US presidential election
Disaster recovery	Analyze public reaction to disasters or terrorist actions	London bombing, Katrina
Social work	Trace hateful messages posted in blogs; identify individuals with suicidal intent	Analysis of racist blogrings
Cultural studies	Study a group's culture, based on age, race, and interests	Current trend analysis in youths and teenagers
Linguistics	Analyze bloggers' linguistic patterns in online writings	Studying the usage of online slang (b4, luv, reli, wat, gr8, neva, and so on)

traditional Web page word segmentation tools such as mutual information, hidden Markov models, decision trees, or neural networks.[4] However, writing conventions in blogs differ from those found in traditional Web pages. Bloggers often write in a rambling and unstructured narrative style—for example, about the many different things they experienced on the same day.[6] Thus, developers might need to customize traditional Web information extraction tools.

In addition to text, blog parsers also extract structural information from blogs, such as comments, posted links, or the bloggers' groups or *blogrings* (blogging communities). This data forms the blog linkage information that can be used in network analysis.

Blog Analyzers

You 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—for example, "like it" or "enjoy the product"—with a positive value to form a document vector that contains each term's frequencies and entry's weight. 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, feedforward/backpropagation 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 or the Jaccard measure between the term features and the additional features such as those used in classification.[11]

The blog analyzer is also capable of analyzing the network relationships among bloggers. Researchers have applied network analysis to Web structure mining with great success, and similar analysis is possible with blogs. Network analysis can include graph analysis such as finding the minimum spanning trees or graph partitions;[12] such information is useful for understanding the social distance between bloggers and different blog communities' characteristics. Researchers can also perform social network analysis to extract characteristics of network topology, centrality, and community. Topological analysis studies the blog network's structural properties and discovers how these properties affect network functions such as information diffusion and communication. Centrality analysis follows topological analysis if the extracted network is nonrandom and node degrees vary greatly. The goal of centrality analysis is to identify a network's key nodes.[13,14] This can help you identify the network's important, influential bloggers. Blog analyzers can use community analysis to identify social groups in blog communities, and they can also apply social network analysis methods, such as blockmodeling to analyze the network relationships among bloggers.

Blog Visualizer

A blog *visualizer* presents content and network analysis results to users—for example, the use

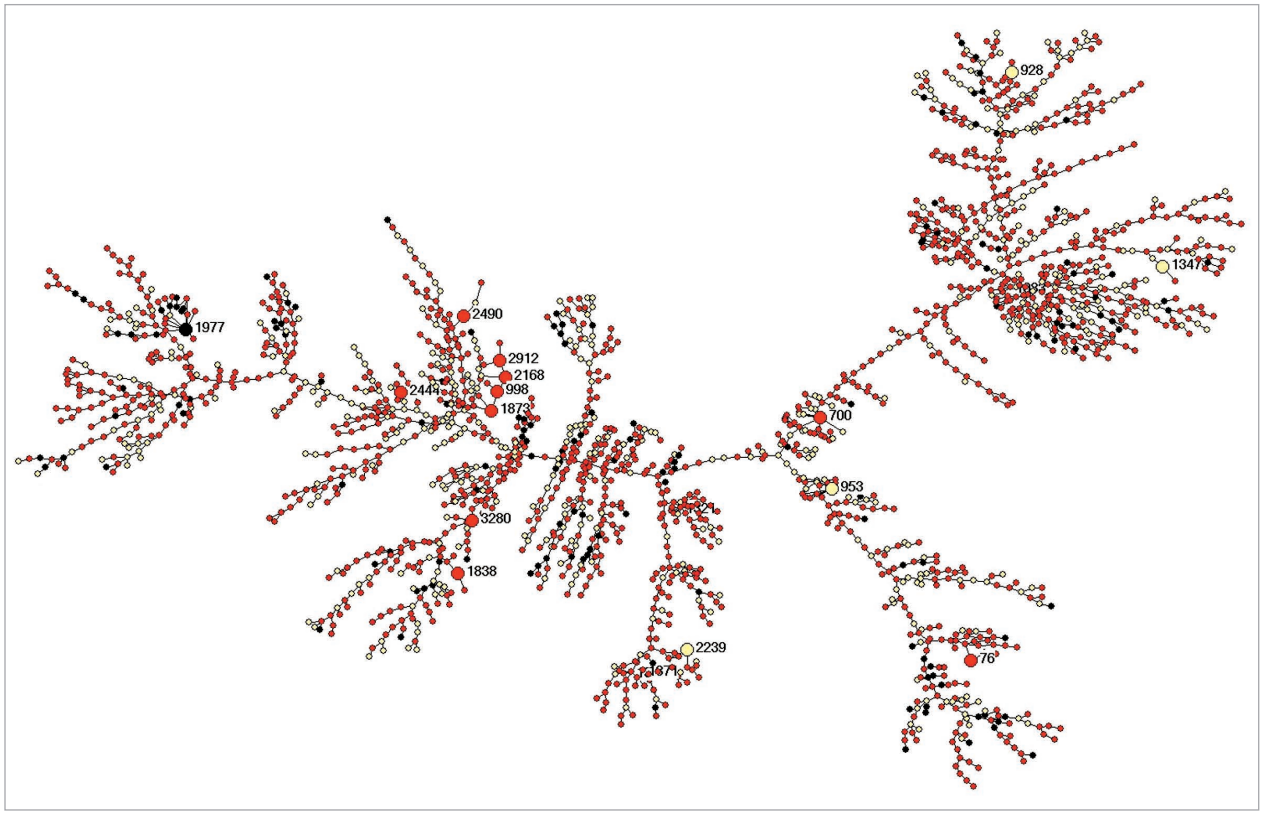

Figure 2. Visualization system. Red nodes represent bloggers with a positive attitude toward iPod, black nodes a negative attitude, and yellow nodes a neutral attitude.[15]

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.[13] Through these techniques, users can easily identify relevant blogs, important communities, and key bloggers in a network.

Applications

Table 1 shows some potential blog mining applications for various domains, such as business, politics, disaster recovery, social work, cultural studies, and linguistics.

Example 1: 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' 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. 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 music player.[15] 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 visualizer 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

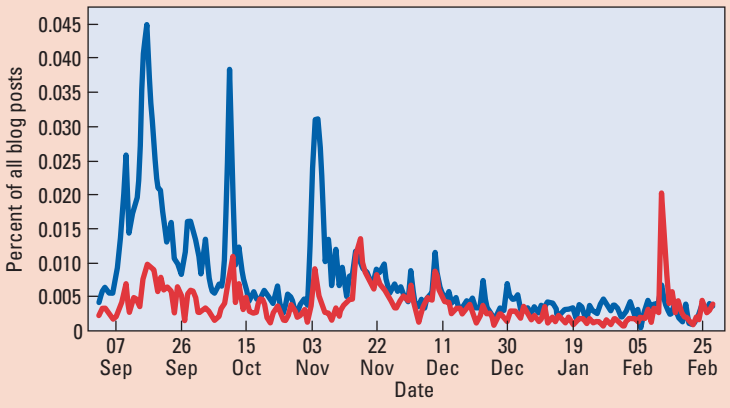

Figure 3. A search on Chen Shui-bian and Ma Ying-jeou (in Chinese) using BlogPulse's Trend Search. The blue and red lines represent the percentage of blog entries that contain the two names Chen and Ma, respectively, during a six-month period (www.blogpulse.com).

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.[15]

Example 2: 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 (820 bloggers) 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).[13] 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—similar to the one in Figure 2. 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.

Example 3: 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 (http://london-underground.blogspot.com/) kept an hourly update of 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 Trend 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.

Developers and researchers can adapt our general framework to other applications, such as analyzing blogs related to movie reviews, which we're currently studying. We predict that organizations will increasingly use blogs to collect useful information via automated techniques. One limitation of blog mining relates to blog quality: some companies pay bloggers to write positive product reviews, thus these blogs don't reflect true user viewpoints. Another problem is splogs—spam blogs that people create to promote a product or another Web site—which are also becoming increasingly popular. Blog mining applications must determine how to distinguish genuine blogs from the others.

References

1. H. Qian and C.R. Scott, "Anonymity and Self-Disclosure on Weblogs," *J. Computer-Mediated Comm.*, vol. 12, no. 4, p. 1.
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. A. Qamra, B. Tseng, and E.Y. Chang, "Mining Blog Stories 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. B. Chen et al., "Predicting Blogging Behavior Using Temporal and Social Networks," *Proc. 7th IEEE Int'l Conf. Data Mining* (ICDM 2007), IEEE CS Press, 2007, pp. 439–444.
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. B. Nardi et al., "Why We Blog," *Comm. ACM*, vol. 47, no. 12, 2004, pp. 41–46.
7. R. Blood, R., "How Blogging Software Reshapes the Online Community," *Comm. ACM*, vol. 47, no. 12, 2004, pp. 53–55.
8. R. Kumar et al., "Trawling the Web for Emerging Cybercommunities," *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, pp. 44–53.
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. D. Shen et al., "Latent Friend Mining from Blog Data," *Proc. 6th IEEE Int'l Conf. on Data Mining* (ICDM 2006), IEEE CS Press, pp. 552–561.
12. K. Ishida, "Extracting Latent Weblog Communities—A Partitioning Algorithm for Bipartite Graph," *Proc. Ann. Workshop Weblogging Ecosystem: Aggregation, Analysis, and Dynamics*, ACM Press, 2005, pp. 1–11.
13. M. Chau and J. Xu, "Mining Communities and Their Relationships in Blogs: A Study of Online Hate Groups," *Int'l J. Human-Computer Studies*, vol. 65, no. 1, 2007, pp. 57–70.
14. L.C. Freeman, "Centrality in Social Networks: Conceptual Clarification," *Social Networks*, vol. 1, no. 3, 1979, pp. 215–240.
15. M. Chau and J. Xu, "Studying Customer Groups from Blogs," *Proc. 6th WeB 2007*, (WEB2007), 2007.

Michael Chau *is an assistant professor and program coordinator in the School of Business at the University of Hong Kong. His research interests include information retrieval and digital libraries, text mining, Web mining, and security informatics. Chau has a PhD in management information systems from the University of Arizona. Contact him at mchau@business.hku.hk.*

Jennifer Xu *is an assistant professor in computer information systems at Bentley College. Her research interests include data mining, social network analysis, and information visualization. Xu has a PhD in management information systems from the University of Arizona. Contact her at jxu@bentley.edu.*

Jinwei Cao *is an assistant professor of management information systems at the University of Delaware. Her research interests include social computing, multimedia information systems, and technology-supported learning. Cao has a PhD in management information systems from the University of Arizona. Contact her at jcao@udel.edu.*

Porsche Lam *is a developer at the Royal Bank of Scotland. He has two BS degrees, one in business administration and the other in software engineering, from the University of Hong Kong, where he participated in several research projects related to Web mining. Contact him at porschelam@gmail.com.*

Boby Shiu *is a researcher in information systems at the University of Hong Kong. His interests are in intelligence solutions, and he's worked on problems with small and medium enterprises and the public sector. Shiu is also a business intelligence consultant at IBM. Contact him at kwshiu@gmail.com.*