Quantification of Social Blog Network Using B-Rank Technique & Blog Recommendation

Syed Fazal Rahaman¹, M. Srilakshmi², Sayeed Yasin ³,
Department of Computer Science and Engineering
Nimra College of Engineering & Technology.
Vijayawada, A.P.-521456
Email: fazi8606@gmail.com¹, srilum.tech@gmail.com², sdyasin761@gmail.com³

Abstract--Now a days, 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. Because blog growth is massive, blog readers can find numerous “hot blogs” on blog hosting sites, yet have no idea which ones contain the most informative content. Some blog service providers (BSPs) list hot blogs based on their number of visitors, but this indicator is weak and insufficient for determining blog popularity, which is a crucial issue in the massive blogosphere.

This project extracts real-world blog data and analyzes the interconnections within blog communities. Such interconnections reveal blogger behaviors and visibilities, as well as blog popularity — all of which might relate to blog quality. Social blog network model on this interconnection structure between blogs and, on this basis, developed a popularity ranking method, called BRank.

While conducting several experiments to analyze various explicit and implicit interconnection structures and discover variability in the impact of interactions in different communities should be made.

1. Introduction

1.1. Quantification

Quantification has several distinct senses. In mathematics and empirical science, it is act of counting and measuring that maps human sense of observation. And experiences into members of some set of numbers.

Quantification in this sense is fundamental to the scientific method.

i. In logic: quantification is the binding of a variable ranging over a domain of discourse. The variable thereby becomes bound by an operator called a quantifier. Academic discussion of quantification refers more often to this meaning of the term than the preceding one.

ii. In grammar, a quantifier is a type of determiner, such as all or many, that indicates quantity. These items have been argued to correspond to logical quantifiers at the semantic level.

Analyzing interconnections among blog communities reveals blogger behaviors that could help in assessing blog quality. A new approach to ranking blogs uses a social blog network model based on blogs’ interconnection structure and a popularity ranking method, called B Rank. Experiments show that the proposed method can discriminate blogs with various degrees of popularity in the blogosphere.

A blog consists of a title, subscription information, and multiple posts that display in descending order by publication date. A blog post typically has the post date and text, and might also include hyperlinks, images, and other media. Bloggers interact with readers and each other in various ways. 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 blog roll, which typically appears as a list of links on a blog’s main page. Also, hyperlinks contained within a blog post give additional information for readers who would like to read related news or blog posts.

By examining the nature of news blogs and the news stream they monitor, we have proposed more suitable metrics for evaluating the topical relevance, timeliness, specificity and credibility of news blog postings than the current standard model. In all cases, it is important to consider the news article to which a blog post links as well as the blog post itself in computing these metrics. Implementing these models effectively involves shallow, automatable analysis of blog post content. In this analysis, it is important to distinguish original post content from quotations, and links to news stories from links to other blog posts. The effort involved is warranted by the increasing importance of blogs in both shaping and reflecting discourse about important events. By using these metrics we can increase the efficiency of searching engine and we can provide reliable blogs information related to user requirements.

2. Problem Review

A blog consists of a title, subscription information, and multiple posts that display in descending order by publication date. A blog post
typically has the post date and text, and might also include hyperlinks, images, and other media. Bloggers interact with readers and each other in various ways. 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. Also, hyperlinks contained within a blog post give additional information for readers who would like to read related news or blog posts.

This project extracts real-world blog data and analyzes the interconnections within blog communities. Such interconnections reveal blogger behaviors and visibilities, as well as blog popularity—all of which might relate to blog quality. Social blog network model on this interconnection structure between blogs and, on this basis, developed a popularity ranking method, called BRank.

3. Objective

A new approach to ranking blogs uses a social blog network model based on blogs’ interconnection structure and a popularity ranking method, called BRank. Analyzing interconnections among blog communities reveals blogger behaviors that could help in assessing blog quality.

4. Methodology

As Figure 1 shows, we use interactive blog behaviors to construct a blog network model that describes the blog-specialized linking structure. Our model includes four types of explicit links: comments, trackbacks, and citations, and blogrolls. These link types represent distinct relationships among bloggers. Furthermore, bloggers might know and visit some blogs based on search engine results for specific interests and topics. These visitors don’t always leave comments on posts that they read, but they share common interests in general. Thus, we assume that these are implicit links that is, interest links among these bloggers. These explicit and implicit relationships could represent interactions among bloggers and also help identify blogger popularity. Call the specific network that these blogging relationships form a social blog network (SBN).

When apply BRank, a weighted and interaction-aware linking analysis, to this network, it lets us rank blogger popularity, quantifying a blog’s impact as a popularity measurement score for blog ranking. To show the performance of our proposed network model and the ranking algorithm, while performing an experimental analysis using the blog data set from three prominent BSPs should be made and then compare our results with BlogLook and manually ranked results.

5. Input Parameter/Size of Sample

5.1. 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 outdegree of blogA AB _ _ = 1, to consider social relationships in its original random walk model, where the probability for a visitor to go from A to B (PA→B) is decided by the out-degree of A. Adjust the probability that a blog reader will follow a link in blog A to another blog B using a new formula,

\[ P R \frac{R A \rightarrow B}{R A \rightarrow B} \left( \sum_{X \in O(A)} \frac{X}{X} \right) = \frac{1}{N}, \]

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

![Figure 1. Linking relationships in the blogosphere. Interactive blog behaviors are used to construct a blog network model of the blogspecialized linking structure.](image-url)
The relationship score \( RA \rightarrow B \) 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 (\( W_{\text{Rtype}} \)) because they have distinct meanings for a blogger. Normally \( W_{\text{comment}} \) is set to 0.25 and others are set to 1. This setting simply represents how easy it is to make a relationship (for example, leaving a comment is the easiest way to support a blogger by using the blogging interface).

The second factor is the number of the corresponding relationship. Here, simply use the degree of the number (\( R_{\text{N}} \)) to express the relationship’s strength. Instead of the actual numbers, while using the actual numbers’ natural log. The final factor is the blog quality score (\( BQ \)), 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. While using the time span between the last date and first date for all posts to represent a blog’s life cycle. These metrics are automatically extracted from data sets.

5.1.2. A Blog Network Model

Our proposed method is based on blog characteristics, including linking structure and blog interest similarity features. The blogosphere’s linking structure is different from that of general web pages, as the former’s network structure should consider the social interconnections among bloggers. When a blogger performs blog interactions such as comments and trackbacks, the link information is generated in the blog pages. An example of blog-specific linking structure is as follows: when blogger A comments on a post in blog B, a link to A is coded into B’s post page. Due to this inserted link, A receives a vote from B in PageRank. However, this might not actually represent support interactions from B. Indeed, the sender of the comment, A, sends a vote to B in this case. The blog sphere’s linking structure is thus different from that of general web pages. When A sends a trackback to B, reciprocal links are established.

The link represents that A sent a vote to B, even though links are present in both blog pages. Posts connected by trackback interactions are typically related. If they’re interested in a topic, visitors can traverse the trackbacks to read relevant posts. So, A, who sends the trackback, is also supported by B, who lists the post link. We regard this kind of interaction as a reciprocal relationship. Blog readers can determine the topics of blogs while browsing the blog posts. If readers are interested in a certain topic, they’re likely to read blogs containing similar content. We hence assume that if A and B are similar — that is, if they share common interests — there will be some probability for A’s readers to read B.

Links mentioned frequently might refer to specific hot topics. Also, we regard the co-occurrence of hyperlinks in the posts of different blogs as an interest similarity relationship in our proposed model.

5.1.2. A Social Blog Network

Our network model defines two of relationships between blogs. Interactions or links from A to B indicate that A is a reader of B, and we take these explicit relationships as support relationships. Implicit interest links in A and B constitute a similarity relationship between them. Our model regards the four kinds of interactive behaviors — comments, trackbacks, blogrolls, and citations — as support relationships. These relationships are social relations the bloggers perform. In addition, we propose virtual associations between blogs as similarity relationships. Assume that if A and B are similar, it’s possible that readers of A will visit B, or readers of B will visit A. Also, hyperlinks in blog posts might refer to topical information. Thus, we extract the hyperlinks. When common interest links exist between A and B, we establish a similarity relationship between them.

This virtual relation is reciprocal in that the two blogs’ contents are related to each other. In the social blog network model, each node represents a blog. Each edge between two nodes represents a relation of the two blogs. The blog network has three general types of blog edges:

- Support edges are formed by support relationships,
- Interest edges are generated by virtual interest similarity relationships,
- Hyperlink edges represent the links in content between a blog and a Web page. The system creates interest edges by counting the common links between two blogs. We assume an interest similarity relationship only if the number of common links is greater than three.

5.2. B Rank:

5.2.1. A Social-Relation-Based Algorithm

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

\[
P_{A \rightarrow B} = \frac{1}{\text{Outdegree of blog} A},
\]  

(1)

To consider social relationships in its original random walk model, where the probability for a visitor to go from A to B (\( PA \rightarrow B \)) 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,

\[ P_{A \rightarrow B} = \frac{R_{A \rightarrow B}}{\sum_{X \in O(A)} R_{A \rightarrow X}} , \tag{2} \]

where \( O(A) \) means blogs linked by A. In B Rank, the probability is determined by the relationship scores \( RA \rightarrow B \). In Equation 2, \( X \) indicates the blogs to which blog A links. The relationship score \( RA \rightarrow B \) represents the relation strength from A to B. It’s decided by three factors. The first is the type of blog relationship (comment, trackback, blog roll, 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. This setting simply represents how easy it is to make a relationship (for example, in our observation, leaving a comment is the easiest way to support a blogger by using the blogging interface). The second factor is the number of the corresponding relationship. Here, we simply use the degree of the number \( R_{N Rtype} \) 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. We use the time span between the last date and first date for all posts to represent a blog’s life cycle. These metrics are automatically extracted from data sets. 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:

$$R_{A \rightarrow K} = \sum_{\text{Rtype}} W_{\text{Rtype}} * RN_{\text{Rtype}} * BQ_k.$$  (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:

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

### Table 2. Blog information and important interaction.

<table>
<thead>
<tr>
<th>Blog</th>
<th>#Blog</th>
<th>#Post</th>
<th>Comment</th>
<th>Trackback</th>
<th>Citation</th>
<th>Blogroll</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wretch</td>
<td>592,123</td>
<td>6,680,087</td>
<td>16,527,101</td>
<td>316,263</td>
<td>154,190</td>
<td>236,168</td>
</tr>
<tr>
<td>Yahoo</td>
<td>294,352</td>
<td>792,335</td>
<td>1,599,940</td>
<td>137,232</td>
<td>253,920</td>
<td>110,637</td>
</tr>
<tr>
<td>Yarn</td>
<td>84,536</td>
<td>1,895,319</td>
<td>2,310,052</td>
<td>104,594</td>
<td>65,125</td>
<td>15,583</td>
</tr>
</tbody>
</table>

### Table 3. Ground-truth evaluation criteria.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommendation</td>
<td>Did the blog recommend other articles?</td>
</tr>
<tr>
<td>Famous person</td>
<td>Is the blogger famous?</td>
</tr>
<tr>
<td>The content of the reply</td>
<td>Does the reply’s content consist of informative description or merely gossip or nonsense?</td>
</tr>
<tr>
<td>Reply</td>
<td>Do you want to reply to the post?</td>
</tr>
<tr>
<td>Citation</td>
<td>Do you want to cite the blogger’s post?</td>
</tr>
<tr>
<td>Blogroll</td>
<td>Do you want to add the blogger to your blogroll?</td>
</tr>
<tr>
<td>Advertisement</td>
<td>Did you support the blog (by clicking the advertisement)?</td>
</tr>
<tr>
<td>RSS</td>
<td>Do you want to monitor the blog by RSS?</td>
</tr>
<tr>
<td>The content of the homepage</td>
<td>Is the homepage attractive?</td>
</tr>
<tr>
<td>The growth of popularity</td>
<td>Do you think the blog’s popularity is increasing or decreasing?</td>
</tr>
</tbody>
</table>

In this Blog analysis, and the link analysis on the interconnection network is an important approach to reveal the roles and impact of elements in the network. Jon Kleinberg proposed the Hyperlink-Induced Topic Search (HITS) algorithm.

- The major concepts in HITS are hubs and authorities. The hub score represents the quality of links to other pages about that topic, whereas the authority score indicates the quality of the page content. Ko Fujimura and his colleagues proposed the Eigen Rumor algorithm for ranking blogs.
- Eigen Rumor is based on HITS and can be applied to communities with observable membership identifications. Information providers have two scores: authority and hub; the algorithm adds the reputation score to indicate the provided information’s quality. In addition to web pages, Yupeng Fu and his colleagues presented an expertise propagation algorithm to find the potential experts on a specific topic using co-occurrences of people from web pages and email communication patterns.
- In a related work on analyzing blogs, Bonnie Nardi and her colleagues analyzed the text of blog posts and comments to identify a community’s feelings.
- Alvin Chin and Mark Chignell used centrality measures and visualizations to detect communities in blogs.
- Their study aimed to improve the classification of emotions from blog posts. Thomas Lento and his colleagues applied the logic regression model and network visualizations to analyze blog data in the Wallop system.
- They investigated the relationship between blog interactions and bloggers’ willingness to continue updating blogs. Tadanobu Furukawa and his colleagues defined the regular reading relation, which describes reading behavior among bloggers.
- They examined four kinds of social networks to predict the regular reading relation and analyze information diffusion. Noor Ali-Hasan and Lada Adamic studied bloggers’ online and real-life relationships in three blog communities.
- They analyzed the different kinds of links, including blogrolls, citations, and comments, and
discovered that few blog interactions reflect real-life relationships. Several studies proposed new algorithms. Eytan Adar and his colleagues introduced the concept of implicit links representing similarity among blog posts.  

- Along with the implicit links, they presented iRank, a blog ranking algorithm. Apostolos Kritikopoulos and his colleagues added implicit links to increase the density of the blogs’ graph based on similarities among topics and users.  

- They modified PageRank into BlogRank, an algorithm for ranking blogs. They don’t consider comments in their work. As the most frequent interactions, comments show their importance in our work. Another algorithm, B2Rank, developed by Mohammad A. Tayebi, is also a modification of PageRank.  

- B2Rank assigns each blog a personality score and an operation score; the scores are mainly obtained from blogrolls and citations. Nitin Agarwal and colleagues investigated the behaviors of influential bloggers and presented a preliminary model to quantitate them.  

- Whereas their work focuses on the community of blogs, our own study targets individual blogs.

5.3. Blog Ranking Approaches

Many websites have developed blog-related ranking technologies. Users can now search blogs by query words at Google Blog Search (http://blogsearch.google.com.tw) and Technorati (http://technorati.com). BlogLook (http://bloglook.urs.tw), a prominent blog ranking service in Taiwan, lets bloggers provide their URLs for ranking. BlogLook’s data contain more than 210,000 blogs. Instead of analyzing the actual blog content, it retrieves blog information from several blog search engines, including Technorati’s Authority, as well as the number of links and subscriptions, and so on. Such services, however, don’t comprehensively consider detailed information — such as content, comments, links, and citations within blogs. Although BlogLook uses the number of blog interactions to calculate the authority score for Technorati, it ignores comments, and it treats all the various interactions the same. Popularity and Interactive Behavior to define a blog’s popularity as its ability to prompt interactive behaviors in other bloggers. On famous and highly popular blogs, content is typically informative and useful for readers. Posts in such blogs could include insightful discussions or innovative opinions, depending on subjective judgments. However, analyzing this content to understand the relationship between bloggers and readers isn’t a trivial and objective task. Thus, the need to find other factors that reveal reader behavior. A blog that receives comments from many bloggers is influential in that the post’s author has an impact on the readers and commentators. In this way, the referenced post shows the impact of the writing quality on its popularity. Also, bloggers subscribe to other blogs that are informative for them, and we could also interpret this interaction as having an impact on the subscribers. Given these observations, to make some assumptions in measuring blog popularity:

- A popular blog might have more support relationships with other blogs than do lesser known blogs  

- Users might cite a blog or a blog post in other web pages, including personal websites, forums, or other Web content.  

- When a blog is frequently referenced on the Internet, we can assume it’s highly popular. PageRank is a link-analysis algorithm that assigns a weighting to each element of a hyperlinked set of documents to measure element importance within the set.

For a set of hyperlinked pages, PageRank assigns a score to each page representing its relative importance — that is, its popularity or reputation — among the set of pages. The blogosphere’s linking structure is similar to that of websites, but with some additional characteristics. Therefore, a general ranking method based on the general linking structure isn’t appropriate for the blogosphere’s structure. As the “Related Work on Blog Analysis” sidebar describes, several projects have used additional information — such as comments, citations, and blogrolls — to rank blogs. In our work, we augment such efforts with the interactions among bloggers to represent the visiting probabilities in PageRank. A New Model As Figure 1 shows, we use interactive blog behaviors to construct a blog network model that describes the blog-specialized linking structure. Our model includes four types of explicit links: comments, trackbacks, citations, and blogrolls. These link types represent distinct relationships among bloggers. Furthermore, bloggers might know and visit some blogs based on search engine results for specific interests and topics. These visitors don’t always leave comments on posts that they read, but they share common interests in general. Thus, we assume that these are implicit links — that is, interest links — among these bloggers. These explicit and implicit relationships could represent interactions among bloggers and also help identify blogger popularity. We call the specific network that these blogging relationships form a social blog network (SBN). When we apply BRank, a weighted and interaction-aware linking analysis, to this network, it lets us rank blogger popularity, quantifying a blog’s impact as a popularity measurement score for blog ranking. To show the performance of our proposed network model and the ranking algorithm, we performed an experimental analysis using the blog data set from three prominent BSPs. We also compared our results with BlogLook and manually ranked results.
In the social blog network model, each node represents a blog. Each edge between two nodes represents a relation of the two blogs. The blog network has three general types of blog edges:

- Support edges are formed by support relationships,
- Interest edges are generated by virtual interest similarity relationships, and
- Hyperlink edges represent the links in content between a blog and a Web page.

The system creates interest edges by counting the common links between two blogs. We assume an interest similarity relationship only if the number of common links is greater than three. Figure 2 shows common links is greater than three. Figure 2 shows

![Figure 2. Social blog network. Each node represents a blog. There are three types of edges — support, interest, and hyperlink — and each represents a blog relation.](image)

<table>
<thead>
<tr>
<th>Table 2. Blog information and important interaction.</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Blog</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Wretch</td>
</tr>
<tr>
<td>Yahoo</td>
</tr>
<tr>
<td>Yam</td>
</tr>
</tbody>
</table>

In this project (see http://140.116.245.199/blogdata), while focus on three well-known BSPs in Taiwan: Wretch Blogs (www.wretch.cc/blog), Yahoo Blogs (http://tw.blog.yahoo.com), and Yam Blogs (http://blog.yam.com). The crawling process for the top blogs ran from September 2007 through May 2008. While chosing the crawling entries from the top bloggers as listed in several authoritative blog sharing and ranking sites. Guaranteed that the data set contains the popular blogs and the blogs that interact with them. After the blog pages were retrieved, we extracted the blog posts, comments, trackbacks, citations, and blogrolls from the pages. Table 2 shows the data-set statistics.

<table>
<thead>
<tr>
<th>Table 3. Ground-truth evaluation criteria.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criteria</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Recommendation</td>
</tr>
<tr>
<td>Famous person</td>
</tr>
<tr>
<td>The content of the reply</td>
</tr>
<tr>
<td>Reply</td>
</tr>
<tr>
<td>Citation</td>
</tr>
<tr>
<td>Blogroll</td>
</tr>
<tr>
<td>Advertisement</td>
</tr>
<tr>
<td>RSS</td>
</tr>
<tr>
<td>The content of the homepage</td>
</tr>
<tr>
<td>The growth of popularity</td>
</tr>
</tbody>
</table>
Table 3 shows the rating criteria, while establishing these features based on my criteria for determining blog popularity. To consider the ranking evaluated from the manually labeled scores as the ground truth in our evaluation and denote it as $\text{RankH}$.

5.4. Applications

Blogs have become increasingly popular, and new blogs are generated every day. Many of the contents are useful for applications in various domains, such as business, politics, research, social work, and linguistics. Whenever the question arises how a product, a personality, or some other specific entity is perceived by the public, the blogosphere is a very good source of information. This is what usually interests business users from marketing.

6. Conclusion and Future Work

To detect and measure the hot blogs is an important task for understanding the social impacts in a large social network. In our future work, we’ll consider a more unique factor “time” for our SBN. By analyzing the SBNs in different time snapshots, we could monitor the SBN’s changes and observe the structural feature evolution over time. We will combine the technique of the community detection and community evolution tracking to our SBN and $\text{BRank}$ to help us to identify the impact of hot blogs.

7. Acknowledgments

This research was supported by grant from Department of Computer Science and engineering in V R Siddhartha Engineering College, Vijayawada.

References

1. L. Page et al., The PageRank Citation Ranking: Bringing Order to the Web, tech. report, Stanford InfoLab, 1999.

Sites Referred
http://java.sun.com
http://www.sourceforgegde.com
http://www.networkcomputing.com/