

# Competent Search in Blog Ranking Algorithm Using Cluster Mining

Robin Singh Bhadoria

*M.Tech (Software System), Samrat Ashok Technological Institute (Govt. Autonomous), Vidisha (MP)*

**Abstract:** In this paper, we focus on blogger search in blogs in particular, who are the top authors for blogs of a particular topic. All blogs and their interconnections are collectively called the blogosphere. We emphasis on a new graph structure known as blogger graph, is proposed in the blogosphere based on the bloggers information.

We concentrated on blogger search in blogs – in particular, who are the top authors for blogs of a particular topic. All blogs and their interconnections are collectively called the blogosphere. Results show that we can identify the top bloggers with a high precision.

## 1. INTRODUCTION

A blog (weblog) is a Web 2.0 communication tool, generally a website maintained by an individual user or by a group of users with regular entries of commentaries, descriptions of events, videos, etc. Most of the blogs provide commentary or news on a particular subject or topic. People mostly write about their personal life experiences, product reviews, technology trends, political opinions, etc. A typical blog entry contains text, images, videos and links to other blogs, web pages, and other media related to the blog topic. A blog reader can post her comments on the blog. This service makes the blogs a highly interactive discussion platform.

### 1.1 Blogs

In this section, we discuss about a typical blog entry structure and important terminologies related to blog. A blog entry consists of contents, comments, trackback links, tags, author information, the posting time information, etc. The important terminologies related to blogs are the following:

- **Blogger:** Blogger is a person who maintains the blog and posts her entries on a blog.
- **Blog-roll:** Every weblog contains a list of other weblogs that the author reads regularly. These lists are known as blog rolls.
- **Permalink:** A permalink, or permanent link, is a URL that points to a specific blog post entry after it has passed from the front page to the archives. A permalink remains unchanged indefinitely.
- **Comment:** A reader posts his own view to a specific post within the site. The comment systems are usually implemented as a chronologically ordered set of responses, i.e., the most recent comment is on the top.
- **Trackback Link:** Trackback is an automatic communication that occurs when one blog references another blog entry. If both weblogs are enabled with trackback functionality then a reference from a post on weblog A to another post on weblog B will update the post on B to contain a back reference to the post on A. This automated referencing system gives authors and readers an awareness of who is discussing their content outside the comments on their site.

## 1.2 Blogosphere

All blogs and their interconnections are collectively called the blogosphere. Blogosphere implies that blogs exist together as a connected community or as a social network in which everyday authors can publish their opinions.

Blogosphere continues to grow considerably, in 2006 around 35 million blogs was there, while in year 2007 this number reached to around 72 million. In 2008, around 184 million new users have started a blog, 346 million users read blogs, 77% of active Internet users read blogs and 120,000 blogs were created every day. The blogosphere growth is shown in Figure 1.1

[1].

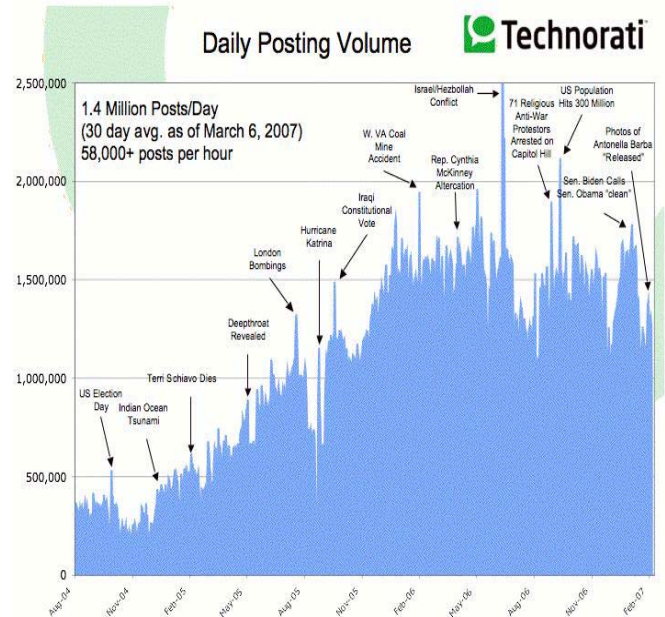


Figure 1.1: Blogosphere Growth Taken From Technorati

Today blogs become more popular than news sites and any other internet information source. Blogosphere structure can be understood by Figure 1.2

In blogosphere, blogs network can be classified mainly into two categories, post network and blog network.

- In post network, nodes are individual blog posts and edges can be comment link, trackback link or any other link between these posts.
- In blog network, nodes are blogs comprising of a number of posts, authors and topics. Edges in this graph are aggregated edges between posts.

Example of blog network and post network from a blogosphere is shown in Figure 1.2.

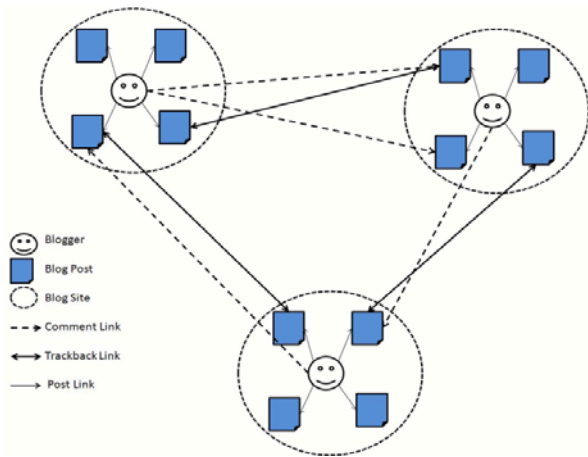


Figure 1.2: Blogosphere Structure

## 2. BLOGGER GRAPH

In a blogger graph there is two components namely blogger graph node and blogger graph edge.

- **Graph Node:** In this graph an individual blogger is a graph node. This node contains many other information about the blogger like his age, gender, address, profile URL etc. This node also contains all the blog posts posted by the blogger.
- **Graph Edge:** In blogger graph, edge between two nodes is the aggregated edges between the blog posts of these two blogger nodes.

An example of a blogger graph is shown in the Figure 2.1

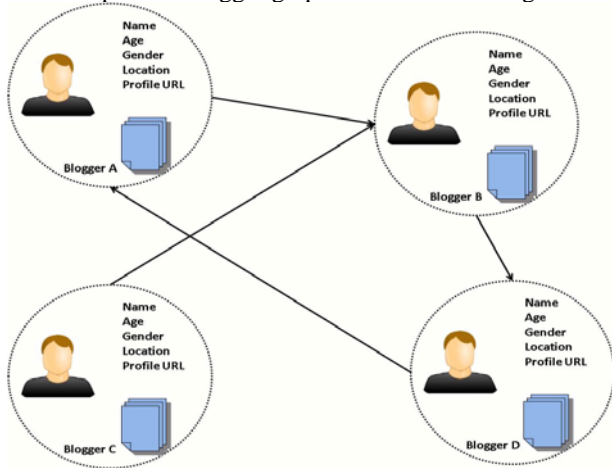
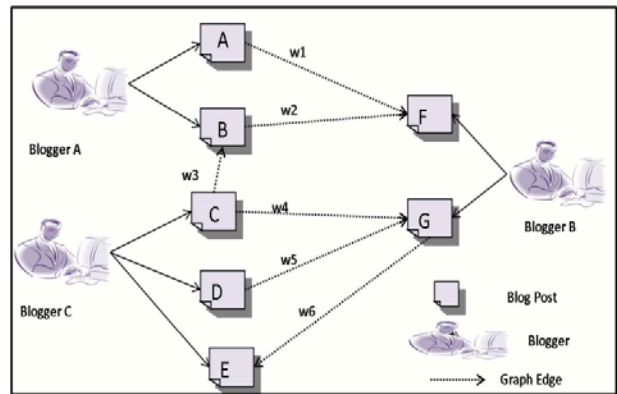


Figure 2.1: Blogger Graph

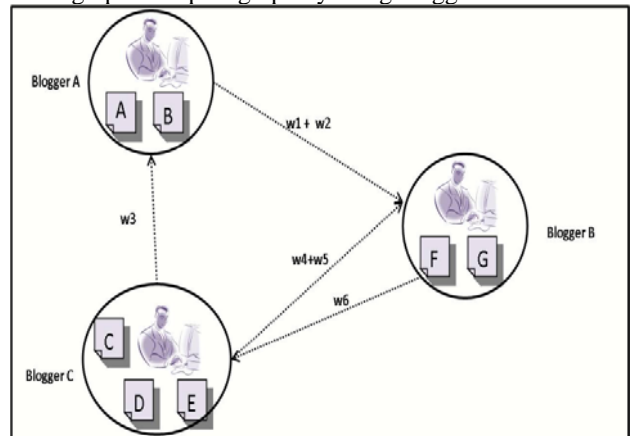
This section explains the transformation of a post graph into blogger graph by an example. In post graph (Figure 2.2) , the posts are graph nodes and edge can be link comment link, similarity link, trackback link etc. In our example a post is represented by document sign and edges between two posts is represented by dotted arrow. Edge weights in this graph are  $w_i$ . In this example there are 3 bloggers and total 7 posts posted by these bloggers. So in post graph total 7 nodes are there and 5 edges between these posts. So the average links per node is  $5/7 = 0.71$ .



Post Graph

Fig 2.2

We aggregate the blogger information of post graph and build the blogger graph. In blogger graph a single node contains multiple blog posts. Edges in the blogger graph are the aggregated edges of post graph. So in our example, blogger graph (Figure 2.3) contains 3 nodes and 4 edges and edge weights become the sum of the edge weights of corresponding post graph edges. So the average links per node for blogger graph is  $4/3 = 1.33$ . This blogger graph is denser than post graph. In this way we are able to produce a denser graph than post graph by using blogger information.



Blogger Graph

Figure 2.3: Transformed Blogger Graph

## 3. BLOG RANKING ALGORITHM

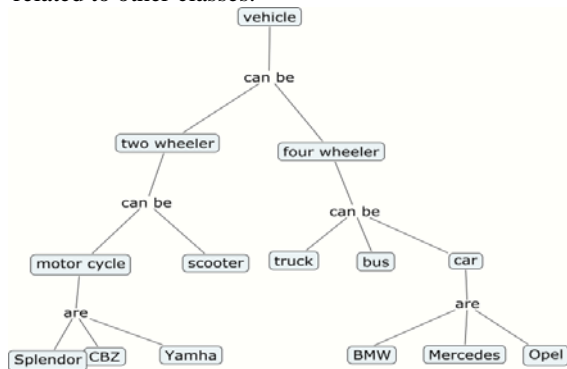
Most of the ranking algorithms those using the interconnected graph. depends on the density of graph [9]. But as already discussed that weblog graph is a sparse graph, so use of these web ranking algorithms give poor performance in the case of blogs. This is mainly due to the fact that bloggers usually write about their personal opinion on a topic without linking to the opinion of other bloggers. If some bloggers provide links then also these links generally point to some newspaper articles, videos etc. and not to weblog entries. So there is need for a new approach for ranking the blogs and already there has been done some work on rank algorithm for blogs, proposed a new approach for ranking of non interconnected documents by creating new links between documents based

on their content similarity. They have generated new directed links based on the probability assigned by the language model induced from one document to the term sequence comprising another. This approach is mainly for non-hypertext documents, proposed a blog ranking method known as iRank. This method ranks the blogs based on how important they are for information propagation. This algorithm is based on inferred implicit structure of blogs. For implicit structure they have used some similarities between blogs mainly blog and link similarity, posting time similarity, text similarity. Algorithm assign high ranking to the blogs that serve as a source of information and later linked with many other blogs.

**3.1 Ontology**

Ontology [14] is an explicit specification of a shared conceptualization. An ontology is a description of the concepts within a domain and relationships that can exist between those concepts. So an ontology can be used to describe a domain. Ontologies describe individuals (instances), classes (concepts), attributes, and relations. An ontology has many important components, few of them are following:-

- Instances or objects: An instance of an ontology is the basic(ground level) object of the ontology.
- Class: A class in an ontology is a group of objects those have identical properties. Classes can be organized into a hierarchy.
- Attributes: An object or class has some properties, features, characteristics, or parameters known as attributes.
- Relation: Relation is the ways in which one class can be related to other classes.



**3.2 Clustering Algorithms**

The main objective of the clustering algorithm is to group the similar objects in the same cluster. If two objects lie in the same cluster then the similarity between these two objects should be more than the similarity between the objects of two different clusters.

Clustering algorithm can be mainly divided into two main categories:-

- Partition Algorithm: Partition clustering, divide the data set into a set of disjoint clusters.
- Hierarchical Algorithm: Hierarchical clustering algorithm proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters into smaller ones. Hierarchical clustering can be divide into

two groups- agglomerative and divisive. Agglomerative clustering is a bottom up approach while divisive clustering is a top down process.

Clustering algorithms like K-means [16], is a partitional clustering algorithm. The K-means algorithm uses some simple iterative steps to classify entire data set into some fixed number of clusters (assume k), defined a priori. Suppose that the k initial cluster centers are given, the algorithm iterates as follows:

- It computes the Euclidean distance from each of the documents to each cluster center and assigns a document to the cluster with the smallest distance.
- Recalculate the k new cluster centers as the centroid of its constituent documents.
- Repeat steps (1) and (2) until the convergence is reached.

**4. PROPOSED SYSTEM**

Crawling Blog is a very different process than web crawling although blog is a subset of web. Because blogs are dynamic in nature so most of the blog storage services provide the Really Simple Syndication (RSS) feed of all the stored blogs. An RSS document may contain full or summarized text of the blog content, plus blog metadata. From this metadata we got the information about publishing date, author name, author profile, etc. RSS feed provides lots of benefits for both publishers and readers. A blog publisher can syndicate the content automatically and a reader can subscribe for timely updates from favored websites or for aggregate feeds from many sites into one place. In web crawling we have to follow the outlinks present on that page but in case of blogs there is no need to follow the outlinks. There are some services like blog.gs and weblog.com that maintain a list of updated blogs. So, in case of blogs crawler we do not need to parse the HTML for getting outlinks information. Instead we fetched the updated blogs list and parsed this list using RSS XML parser. We used the weblogs.com to get the list of updated blogs. It provides two types of list of updated blogs URLs, one is last one hour updated blogs and other is last 15 minutes updated blogs.

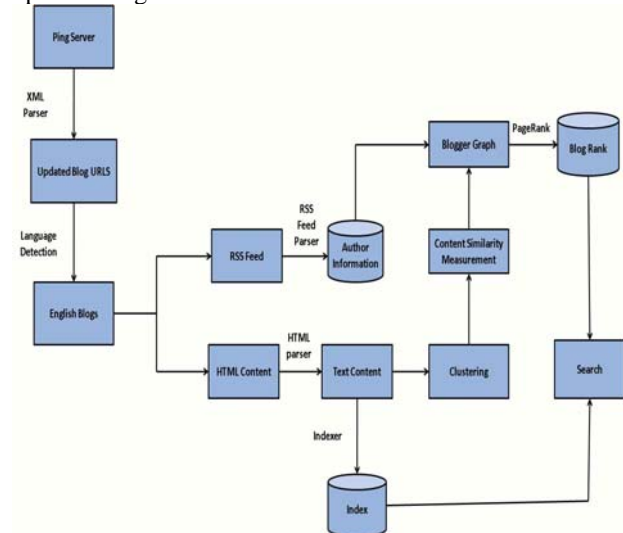


Figure 4.1: Proposed System

To calculate dissimilarity between two blog posts we used the keywords extracted from the blog posts and the ontology. We first calculated the dissimilarity between each keyword pair of the two documents and used this for calculating the dissimilarity between the documents.

**5. RESULT**

*Data set* : We worked in sports category blogs. For blog data we manually downloaded the blogs for cricket, tennis, football from different blog sites. A total of 1645 blogs were downloaded for these three categories. Details of these blogs is shown in Table 5.1

Category	Number of Blogs
Cricket	595
Tennis	450
Football	600

Table 5.1: Blog Dataset information

**5.1 Clustering Results**

We applied K-Means and PDDP clustering algorithm on the blog data. To measure the quality of a cluster we used the purity measurement. Purity of a cluster is defined as the fraction of blogs belonging to the dominant category in that cluster.

$$Purity(c_i) = \frac{max_j(|c_i|_{class=j})}{|c_i|}$$

where Purity (c i) is the purity of a cluster i. (|c i| class=j) denote number of items of class j assigned to cluster i.,|c i| size of cluster i.

For K-Means clustering we used randomly selected centroid points to initialize the algorithm. Since we have three types of blogs in our dataset we fixed a priori number of cluster to 3. We used TMG software [27] for K-means clustering. Purity results of the 5 K-Means clustering are tabulated in Table5.2.

Experiment No.	Cricket Purity	Tennis Purity	Football Purity	Average Purity
1	84.44	70.15	81.25	78.61
2	83.25	72.39	79.86	78.5
3	81.96	73.43	80.95	78.78
4	80.64	71.42	83.46	78.51
5	85.43	70.15	78.72	78.1

Table 5.2: Purity of K-Means Clustering

Confusion matrix for one of the K- Means clustering experiment is shown in Table 5.3.

Cluster Number	Cricket	Tennis	Football	Purity
1	76	4	10	84.44
2	22	80	12	70.15
3	2	16	78	81.25

Table 5.3: Confusion Matrix for K-Means Clustering

Because in this case penalty score was assigned. In the last case when both the keywords were not present in the ontology the dissimilarity function value is more than the first three cases i.e. both the keywords are highly dissimilar.

Keyword 1	Keyword 2	Dissimilarity Function Value
Sachin Tendulkar (Y)	Rahul Dravid (Y)	2
Sachin Tendulkar (Y)	Sania Mirza (Y)	14
Sachin Tendulkar (Y)	Sonia Gandhi (N)	31
Aamir Khan (N)	Sonia Gandhi (N)	62

Table 5.4: Dissimilarity Function Value for Different Keyword Pairs

**6. CONCLUSION AND FUTURE WORK**

In this thesis, we have proposed a framework for author search in blogs. Web search algorithms don't perform well in case of blogs because of sparseness of blog graph. In our work, we proposed a new graph structure in blogs named as blogger graph. For building this graph we used the blogger information that is available in every blog post. This graph is denser then the post graph. To make the graph further denser we used the semantic similarity between the posts. We calculated semantic similarity between the blog posts using the ontology. Then we added the new edges in the graph based on the semantic similarity. Then we applied the PageRank algorithm on this modified graph. To evaluate the system, we experimented on sports blogs and collected users feedback. The results are encouraging from our experiments.

**REFERENCES**

- [1] Blogosphere growth. <http://www.sifry.com/alerts/archives/000334.html>.
- [2] Apostolos Kritikopoulos, Martha Sideri, and Iraklis Varlamis. Blogrank: ranking weblogs based on connectivity and similarity features. In AAA-IDEA '06: Proceedings of the 2nd international workshop on Advanced architectures and algorithms for internet delivery and applications, page 8, New York, NY, USA, 2006. ACM.
- [3] Adamic L. Adar E., Zhang L. and Lukose R. Implicit structure and the dynamics of blogspace. In Presented at the Workshop on the Weblogging Ecosystem at the 13th International World Wide Web Conference, New York, 2004.
- [4] Marti A. Hearst, Matthew Hurst, and Susan T. Dumais. What should blog search look like? In SSM '08: Proceeding of the 2008 ACM workshop on Search in social media, pages 95–98, New York, NY, USA, 2008. ACM.
- [5] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Stanford InfoLab, November 1999.
- [6] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In Journal of the ACM, volume 46, pages 668–677, 1999.
- [7] Taher H. Haveliwala. Topic-sensitive pagerank. In WWW '02: Proceedings of the 11th international conference on World Wide Web, pages 517–526, New York, NY, USA, 2002. ACM.
- [8] Paolo Massa and Conor Hayes. Page-rerank: Using trusted links to re-rank authority. In WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, pages 614–617, Washington, DC, USA, 2005. IEEE Computer Society.
- [9] Amy N. Langville and Carl D. Meyer. Deeper inside pagerank. In Internet Mathematics, volume 1, page 2004, 2004.
- [10] Oren Kurland and Lillian Lee. Pagerank without hyperlinks: structural re-ranking using links induced by language models. In SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pages 306–313, New York, NY, USA, 2005. ACM.



- [11] Ko Fujimura. The eigenrumor algorithm for ranking blogs. In In Ding-Zhu Du and Jie Sun, editors, *Advances in Optimization and Approximation*, page pages. Academic Publishers, 2005.
- [12] Jung-Hoon Kim, Tae-Bok Yoon, Kun-Su Kim, and Jee-Hyong Lee. Trackback-rank: An effective ranking algorithm for the blog search. In *IITA '08: Proceedings of the 2008 Second International Symposium on Intelligent Information Technology Application*, pages 503–507, Washington, DC, USA, 2008. IEEE Computer Society.
- [13] Nilesch Bansal and Nick Koudas. Blogscope: spatio-temporal analysis of the blogosphere. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1269–1270, New York, NY, USA, 2007. ACM.
- [14] Christopher Brewster, Kieron O'Hara, Steve Fuller, Yorick Wilks, Enrico Franconi, Mark A. Musen, Jeremy Ellman, and Simon Buckingham Shum. Knowledge representation with ontologies: The present and future. In *IEEE Intelligent Systems*, volume 19, pages 72–81, Los Alamitos, CA, USA, 2004. IEEE Computer Society.
- [15] H.W.Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2 (1955).
- [16] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.
- [17] Daniel Boley. Principal direction divisive partitioning. In *Data Mining and Knowledge Discovery*, volume 2, pages 325–344, Hingham, MA, USA, 1998. Kluwer Academic Publishers.
- [18] Christopher H. Brooks and Nancy Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 625–632, New York, NY, USA, 2006. ACM.
- [19] Beibei Li, Shuting Xu, and Jun Zhang. Enhancing clustering blog documents by utilizing author/reader comments. In *ACM-SE 45: Proceedings of the 45th annual southeast regional conference*, pages 94–99, New York, NY, USA, 2007. ACM.
- [20] Nitin Agarwal, Magdiel Galan, Huan Liu, and Shankar Subramanya. Clustering blogs with collective wisdom. In *ICWE '08: Proceedings of the 2008 Eighth International Conference on Web Engineering*, pages 336–339, Washington, DC, USA, 2008. IEEE Computer Society.
- [21] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. In *INFORMATION PROCESSING AND MANAGEMENT*, pages 513–523, 1988.
- [22] Shunyao Wu, Jinlong Wang, Huy Quan Vu, and Gang Li. Text clustering with important words using normalization. In *JCDL '10: Proceedings of the 10th annual joint conference on Digital libraries*, pages 393–394, New York, NY, USA, 2010. ACM.
- [23] Gate tool. <http://gate.ac.uk/>.
- [24] Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third ACL Applied NLP*, Trento, Italy, 1992.
- [25] Hamish Cunningham, Hamish Cunningham, Diana Maynard, Diana Maynard, Valentin Tablan, and Valentin Tablan. Jape: a java annotation patterns engine, 1999.
- [26] Jaana Kekalainen Kalervo Jarvelin. Cumulated gain-based evaluation of information retrieval techniques. In *ACM Transactions on Information Systems* 20(4), pages 422–446, 2002.