Volume 2, No. 6, June 2011

Journal of Global Research in Computer Science

ISSN-2229-371X

RESEARCH PAPER

Available Online at www.jgrcs.info

SEARCHING IMPROVEMENT IN BLOGS USING DATA MINING TECHNIQUES

Robin Singh Bhadoria
M.Tech (Software System)
Samrat Ashok Technological Institute, Vidisha (MP)
ssr_robin@yahoo.co.in

Abstract: To evaluate the system, we experiment on sports blogs and collect user's feedbacks. This paper focus 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. In these approach significant intervals for each Website are computed first (independently) and these interval used for detecting frequent patterns and then the Analysis is performed on Significant Intervals and frequent patterns.

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. In this paper, a new graph structure known as blogger graph is proposed in the blogosphere based on the bloggers information. Results show that we can identify the top bloggers with a high precision.

INTRODUCTION

A lot of research has been done in the area of Web usage clustering, which directly or indirectly addresses the issues involved in data mining for extraction of web navigational patterns, ordering relationships, prediction of web surfing behavior, and clustering of web usage sessions based on web logs. Many research studies have looked at capturing users' web access patterns and store them in log files for different purposes.

Some techniques of weblog data mining use cookies to identify site users and user sessions. Cookies play role of markers that are used to tag and track site users automatically. Another approach to identify users is to use a remote agent, as described in, uses Java agents that is run on the client side in order to send back accurate usage information to the Web server. The major disadvantage of both techniques is that they rely on implicit user cooperation, which doesn't exist in many cases. There is a constant conflict between the Web user's desire for privacy and the Web provider's desire for collecting information about the visitors. In fact many users disable the browser features that allow storage of cookies or Java agents, which makes such techniques impractical. We assume that

user identification is not facilitated by the techniques described above.

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 terminology related to blogs is the following:

a. Blogger: Blogger is a person who maintains the blog and posts her entries on a blog.
b. Blog-roll: Every weblog contains a list of other weblogs that the author reads regularly. These lists are known as blog rolls.
c. 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.
d. 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.
e. 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.

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, and 77% of active Internet users read blogs and 120,000 blogs were created every day. The blogosphere growth.

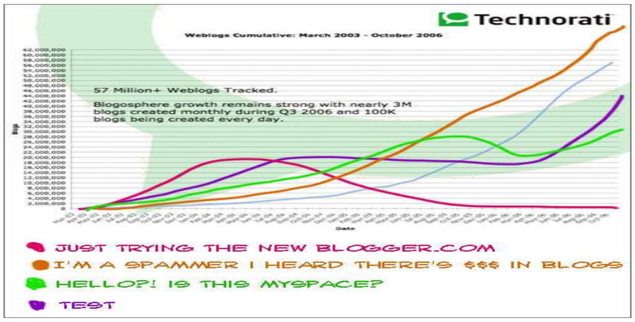


Figure.1Blogosphere Growth Taken From Technorati

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.

a. In post network, nodes are individual blog posts and edges can be comment link, trackback link or any other link between these posts.
b. 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.

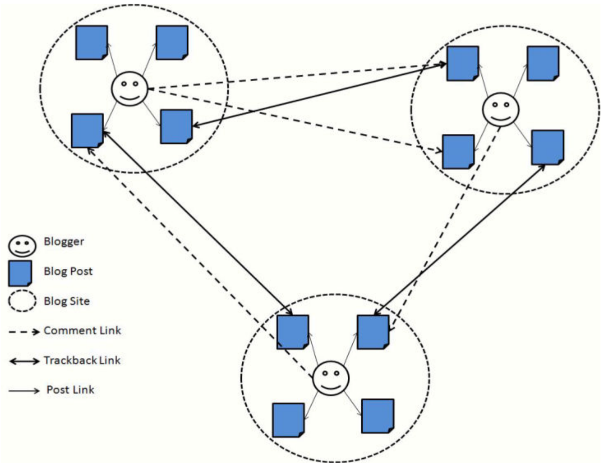


Figure.2 Blogosphere Structure

BLOGGER GRAPH

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

a. Graph Node: In this graph an individual blogger is a graph node. This node contains much 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.
b. 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

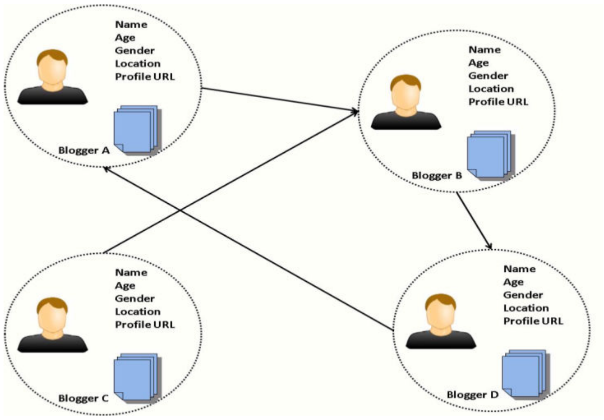
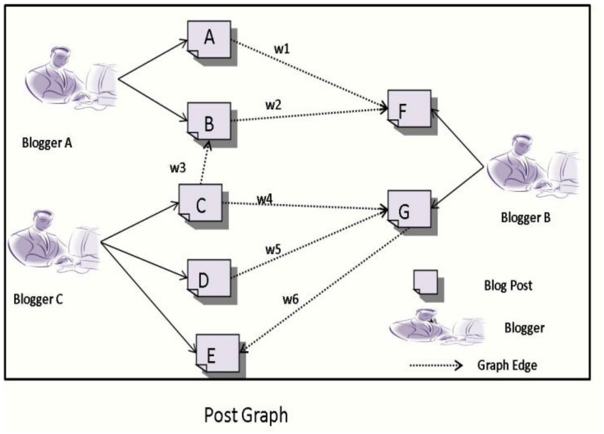


Figure.3 Blogger Graph

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 out example a post is represented by document sign and an edge between two posts is represented by dotted arrow. Edge weights in this graph are wi .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.



We aggregate the blogger information of post graph and build the blogger graph. In blogger graph a single no de contains multiple blog posts. Edges in the blogger graph are the aggregated edges of post graph. So in our example, blogger graph (Figure2.3) 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.

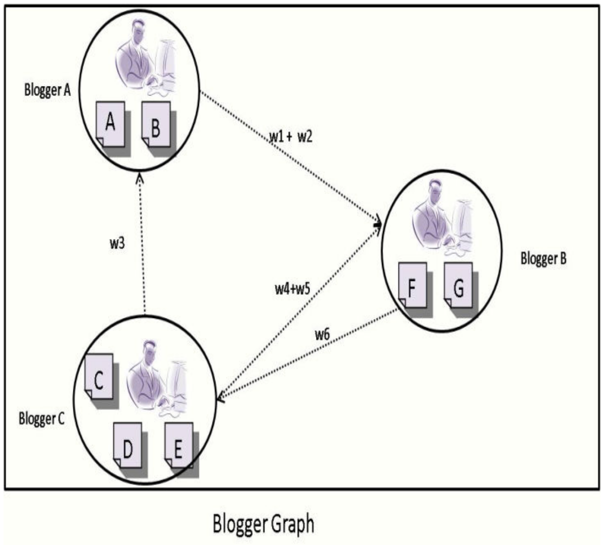


Figure.4 Transformed Blogger Graph

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.

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:-

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

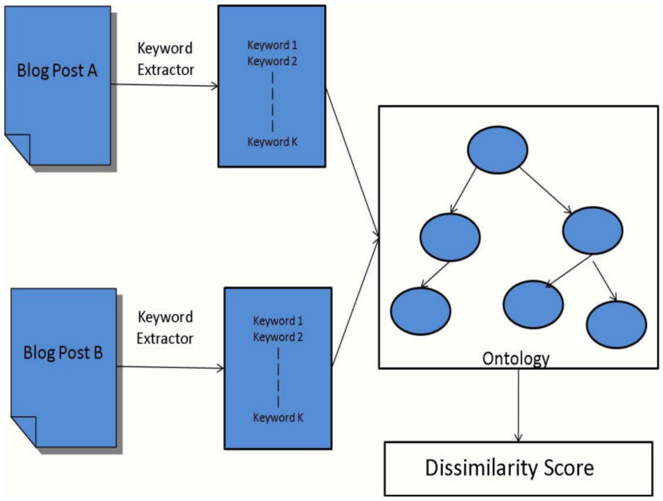


Figure.5 Dissimilarity Calculation Framework

We calculated dissimilarity function value between the keywords, i.e, if the dissimilarity function value is high means keywords are highly dissimilar.

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:-

a. Partition Algorithm: Partition clustering, divide the data set into a set of disjoint clusters.
b. 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 dividing into two groups- agglomerative and divisive. Agglomerative clustering is a bottom up approach while divisive clustering is a top down process.

Hungarian Algorithm

Hungarian algorithm is an algorithm to solve the assignment problem. The assignment problem, also known as the maximum weighted bipartite matching problem
can be stated as follows:-

Assignment Problem: Given a set of n workers, a set of n jobs, and a set of ratings indicating how well each worker can perform each job, determine the best possible assignment of workers to jobs, such that the total rating is maximized and every worker has been assigned exactly one job. More generally, given a bipartite graph made up of two partitions V and U , and a set of weighted edges E between the two partitions, the problem requires the selection of a subset of the edges with a maximum sum of weights such that each node $v_i \in V$ or $u_i \in U$ is connected to at most one edge.

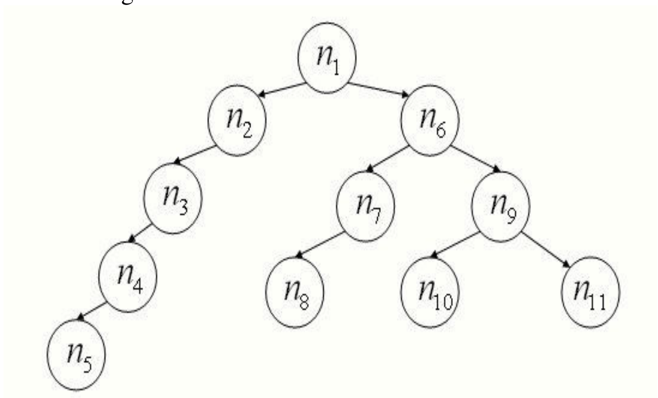


Figure.6 Example for Hungarian Algorithm

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.

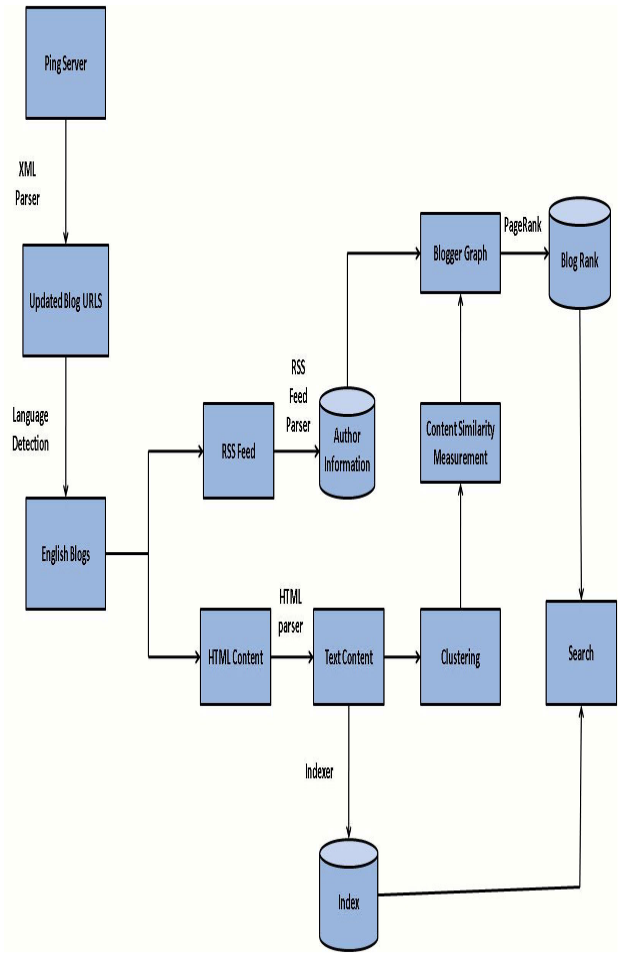


Figure.7 Proposed System

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.

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

Table. I Blog Dataset information

Category	Number of Blogs
Cricket	595
Tennis	450
Football	600

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.
(lc il class=j) denote number of items of class j assigned
to cluster i.,lc il 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.

Table .II Purity of K-Means Clustering

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

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

Table.III Confusion Matrix for K-Means Clustering

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

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.

Table .IV Dissimilarity Function Value for Different Keyword Pairs

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

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.

For future work we have the following suggestions

a. Our dataset size was very small because we downloaded the blogs manually for sports category. The automated system we have implemented for data collection needs lots of improvement. We got lots of noise and spam data from that system. So work can be done to improve the data collection stage.
b. We build the ontology manually for sports category. There is lots of research is going on in field of automatic ontology building from the text corpus.
c. In our experiments, we used the different threshold values for adding the edge in the blogger graph and find out the best result for high values of k. Due to small size of our dataset the value of threshold is not best tuned. So experiments need to be done on a larger dataset and parameters need to be best tuned.
d. We only used content similarity score for adding edges in the blogger graph. The system can be improved by using the other links like comment links, links between posts. There can be other attributes of similarity between posts like posting time similarity, writing style similarity. These attributes can also be used for calculating similarity between the posts.
e. This system can also be used for community finding in blogs. The blogger graph structure can be used for finding out the similar interest bloggers and then find out the communities based on interest similarity.

REFERENCES

[1] Blogosphere growth.
http://www.sifry.com/alerts/archives/000334.html.

[2] A. Nanopoulos, D. Katsaros and Y. Manolopoulos, "A data mining algorithm for generalized web prefetching," IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 5, pp. 1155-1169, 2003

[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] X. Jin, Y. Zhou, and B. Mobasher, "Web usage mining based on probabilistic latent semantic analysis," in KDD '04: Proceedings of the 2004 ACM SIGKDD

international conference on Knowledge discovery and data mining. New York, NY, USA: ACM Press, 2004, pp. 197-205

[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 So ciety.

[13] Nilesh 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 Intel ligent 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. Bro oks 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 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] Lucene to olkit. http://lucene.apache.org/.

[27] Tmg tool. http://scgroup.hpclab.ceid.upatras.gr/scgroup/Projects/TMG/.

[28] Omniclusterer clustering library. http://www.tcllab.org/canasai/software/omniclusterer/.

[29]C-map tool. http://cmap.ihmc.us/.

[30] Jaana Kekalainen Kalervo Jarvelin. Cumulated gain-based evaluation of in-formation retrival techniques. In ACM Transactions on Information Systems 20(4), pages 422–446, 2002.

[31] Renate Iváncsy, István Vajk, Frequent Pattern Mining in Web Log Data Acta Polytechnica Hungarica Vol. 3, No. 1, 2006

[32] Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques, chapter 1.2, page 5. Morgan Kaufmann Publishers, 2005

[33] 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.

[34] 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.