BLOGRANK: Ranking on the blogosphere

Apostolos Kritikopoulos
Athens University of Economics and Business
Patision 76, Athens, Greece
+30 6977687978
apostolos@kritikopoulos.info

Martha Sideri
Athens University of Economics and Business
Patision 76, Athens, Greece
+30 2108203149
sideri@aueb.gr

Iraklis Varlamis
Athens University of Economics and Business
Patision 76, Athens, Greece
+30 2108203160
varlamis@aueb.gr

Abstract
Although, the Blogosphere is part of the World Wide Web, weblogs present several features that differentiate them from traditional websites: the number of different editors, the multitude of topics, the connectivity among weblogs and bloggers, the update rate, and the importance of time in rating are some of them. Traditional search engines perform poorly on blogs since they do not cover these aspects. We propose an extension of Pagerank, which analyses and extends the link graph, in an attempt to exploit some of the weblog features. The analysis of the weblogs' link graph is based on the assumption that the visitor of a weblog tends to visit relevant or affiliated weblogs. Our algorithm, BlogRank, models the similarity between weblogs by incorporating information on common users, links and topics and generates a global ranking for all weblogs in a set. To validate our method we ran experiments on a weblog dataset, processed and adapted to our search engine:

http://spiderwave.aueb.gr/blogwave

Our experiments suggest that our algorithm enhances the quality of returned results.

Keywords
Ranking, related blogs, graph-based ranking, PageRank

1. Introduction
Although the graph formed by the hyperlinks between weblog posts is part of the web graph, the ranking algorithms for web pages (i.e. PageRank) seem to be insufficient for the following reasons:

- The number of links between weblog entries is very small. Thus, the weblog entries graph is very sparse and the ranking algorithms do not perform well.
- Weblog-specific information (time, topic, editor etc.) is not exploited in its full extent.

To improve the ranking results of PageRank on Weblogs we propose the following method:

- We first process the weblog graph from to provide a denser graph.
- We assign weights to the new edges taking into account several criteria, such as similarity in topics and contributors between the source and target nodes, the number of explicit hyperlinks between nodes, and the difference in time of creation between the source and target node. The weights are assigned in a way that new hyperlinks are promoted.
- We modify in the standard PageRank algorithm to incorporate these criteria

We test the efficiency of our ranking method in a sample weblog dataset provided by Nielsen BuzzMetrics, Inc. For this reason we have developed an experimental search engine over the dataset and allow web users to provide human judgment on the results. We use both implicit and explicit human evaluation measures in order to evaluate our algorithm.

The contributions of our work comprise: a blog ranking algorithm that exploits many of the weblog intrinsic features and reveals a new way for ranking weblogs, a blog search engine that can be extended to cover larger parts of the blogosphere, a metric of user satisfaction that exploits implicit [1] and explicit user feedback. With the aid of this metric we are able to measure the relevance of a page to the query and therefore evaluate the ranking algorithm. For this paper we made available a test service for weblogs. The service can be accessed at:

http://spiderwave.aueb.gr/blogwave

2. The BLOGRANK algorithm
The output of our algorithm is a ranking of all weblogs in the dataset. This overall ranking is used by our search engine for the presentation of results: matching entries from highly ranked weblogs are presented first. In the case of entries from the same weblog, most recent entries are presented first. BlogRank is a generalized approach of Pagerank [1]. The BlogRank of a Weblog A is given by the formula:

$$B(A) = (1\text{-}E) + E\,(FN(U_1 \rightarrow A)*B(U_1) + \ldots + FN(U_n \rightarrow A)*B(U_n)) \quad (1)$$

where: B(A) is the BlogRank of weblog A,
B(Ui) is the BlogRank of weblog Ui which link to weblog A,
E is a damping factor between 0 and 1 (normally is 0.85)
FN(Un→A) is the probability that a user who visits weblog n selects weblog A as a next visit, and denotes a factor which shows how much the weblog Un « fancies » weblog A.
The following equation holds:

$$\sum_{j=1}^{t} FN(Uz \rightarrow j) = 1 \quad (2)$$

where: z is a weblog with t outlinks (to other weblogs)
FN(Uz→j) is the possibility that the user will choose weblog j

If we assume in BlogRank that FN(Uz→j)=1/N where N is the total number of outlinks in weblog z, then we can easily derive the PageRank formula. We strongly believe that a user is not attracted equally by every outlink that exist in post of a given weblog. The most probable case is that the user was driven to a post because she was looking for topic or she is interested for the main subject of the post. It is logical to hypothesize that she is most probably going to continue her quest, by selecting similar post or news. From all the outlinks of weblog z, the significant function FN(Uz→j) favours those posts of the j weblog that:

a) belong to common categories with the weblog z
b) same users have posted as in weblog z
c) link to the same news posts as mentioned in weblog z

Before we apply the BlogRank, we expand the connected graph of the weblogs by adding bidirectional links between the weblogs that share same categories, users and news. Then we apply weights to every connection. The utility function that gives the possibility of user to move to weblog j once in z is :

$$FN(U_{z\to j}) = \frac{F_{z\to j}}{\sum(F_{z\to x})} \qquad (3a)$$

Where

$$F_{z\to k} = L_{z\to k} + w_T * T_{z\to k} + w_A * A_{z\to k} + w_N * N_{z\to k} + w_D * D_{z\to k} \qquad (3b)$$

and

L is the number of links from weblog z to weblog j
T is the number of common tags/categories between z and j
A is the number of authors that have posted in both z and j
N is the number of couplings of z and j to news URLs.
D equals to 24*60 / average(posting time difference in minutes), between z and j (only for hyperlinked posts)

$\mathbf{W_T}$, $\mathbf{W_A}$, $\mathbf{W_N}$, $\mathbf{W_D}$ are the weights we use in each one of the factors T,A,N,D respectively.

We generated 43 different rankings using different values for the parameters of formula 3b. We used human judgments to decide on the importance of the top-40 weblogs of every ranking. Experimentally we adjusted weights of formula 3b to the following values: W_T=1.70, W_A=1.10, W_N=4.80, W_D=0.40. The aim was to maximize user satisfaction from the top ranked results, and consequently from the results of every individual query. Although the selected weights are not fine-tuned, and an extended evaluation is under consideration, the first results we have available show that BlogRank outperforms the rest of the algorithms we tested.

3. Evaluation

In our experiments we do not assume a priori knowledge neither on the ranking of documents nor on their relevance to every possible query. In order to perform the evaluation process we use the Success Index (SI) metric, a number between 0 and 1 which was presented in [2]:

$$SI = \frac{1}{n}\sum_{t=1}^{n}\frac{n-t+1}{d_t * n} \qquad (4)$$

where: ***n*** is the total number of the posts selected by the user
$\boldsymbol{d_t}$ is the order in the list of the ***t***-th post selected by the user

The basic advantage of Success Index is that it does not require the user to vote for her satisfaction. BlogWave records the posts clicked on by the user, and the order in which they are clicked.

During our experiment period, users of the BlogWave service made queries and got results ranked either using BlogRank or Pagerank. The algorithm used for ranking was randomly selected every time. Users selectively clicked on the results as a means of evaluation. Comparison of user implicit evaluation proved that BlogRank considerably improved the quality of the retrieved information.

4. Conclusions and future work

We have proposed a method for using link graph characteristics, time and common attributes between the posts to enhance the quality of the results of the ranking mechanism for each weblog's importance. Our experimental results are quite encouraging. Of course, more experimental evaluation of our method, as well as tuning of its parameters is needed.

We plan to process other aspects of the posts graph, more specifically, instead of grouping posts by weblog we plan to group posts "by topic" and "by author", thus forming a graph of interconnected topics and a graph of interconnected authors. The strength of each connection will be based on the number of real links between posts of each topic or author. Both author and topic graphs are directed, strongly connected and have many nodes. Using the biased surfer model we can estimate the probability of a surfer to follow a link to another topic or to another author's post thus revealing the most authoritative authors or topics [3].

Acknowledgments

Our thanks to our students, who helped us with the experiments. To Nielsen Buzmetrics, for the weblog dataset.

References

[1] Joachims, T. , Granka, L. , Pan, B., Hembrooke, H, Gay, G., (2005). Accurately interpreting clickthrough data as implicit feedback, Proceedings of the 28th annual international ACM SIGIR conference.

[2] Kritikopoulos, A., Sideri, M. (2005). The Compass Filter: Search Engine Result Personalization Using Web Communities, Lecture Notes in Computer Science, Springer, Volume 3169, pp. 229 – 240

[3] Nakajima, S., Tatemura, J., Hino, Y., Hara, Y., Tanaka, K., (2005), "Discovering Important Bloggers based on Analyzing Weblog Threads", 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, WWW 2005.

[4] Page, L., Brin, S., Motwani, R. & Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford, USA.