Blog Rating as an Iterative Collaborative Process

Malamati Louta and Iraklis Varlamis

Abstract. The blogosphere is a part of the World Wide Web, enhanced with several characteristics that differentiate blogs from traditional websites. The number of different authors, the multitude of user-provided tags, the inherent connectivity between blogs and bloggers, the high update rate, and the time information attached to each post are some of the features that can be exploited in various information retrieval tasks in the blogosphere. Traditional search engines perform poorly on blogs since they do not cover these aspects. In an attempt to exploit these features and assist any specialized blog search engine to provide a better ranking of blogs, we propose a rating mechanism, which capitalizes on the hyperlinks between blogs. The model assumes that the intention of a blog owner who creates a link to another blog is to provide a recommendation to the blog readers, and quantifies this intention in a score transferred to the blog being pointed. A set of implicit and explicit links between any two blogs, along with the links' type and freshness, affect the exchanged score. The process is iterative and the overall ranking score for a blog is subject to its previous score and the weighted aggregation of all scores assigned by all other blogs.

Keywords: blog, ranking, collaborative rating, local and global rating.

1 Introduction

In the competitive industry of web search, the increase of web coverage and the improvement in ranking of results are the two main aims of any potential player. Due to the rapid increase of its content, blogosphere attracted the interest of popular web search engines (e.g. Google, Yahoo! and AskJeeves), companies that provide access exclusively to the blogosphere content (e.g. Blogpulse [1], and Technorati [2]) and researchers that focus to web search [3].

Every blog consists of a series of entries (namely posts), which carry apart from text or other media content, several hyperlinks to other entries or web pages and a timestamp information concerning the post creation. Using this linking mechanism, blogosphere is converted to an interconnected sub-graph of the web, with links to the surrounding web graph too. Similarly to normal links, blog links are used as suggestions or as a means to express agreement or disagreement [4] to

Malamati Louta and Iraklis Varlamis
Harokopio University of Athens, Department of Informatics and Telematics
176 71, Athens, Greece
e-mail: {louta,varlamis}@hua.gr

M. Wallace et al. (eds.): Semantics in Adaptive and Personalized Services, SCI 279, pp. 187–203.
springerlink.com

a blogs' content. However, due to the ease of the publishing mechanism, they have been utilized to bias search engine results (e.g. splogs, google bombs etc).

Since publishing in blogs comes at no cost for web users, the content and number of links provided by individual writers world-wide can easily surpass those in registered websites. This change affects the structure of the web graph and forces search engines to adapt. The ranking mechanisms of web search engines have two main options, concerning blog links: a) to completely ignore them, in order to avoid spamming and b) to take them into account. In the latter case, they have to tackle several trust related issues.

This work perceives hyperlinks in blogs as recommendations to blog readers and models the network of hyperlinked blogs as a continuous process where respectful or disrespectful sources recommend other trustful or distrustful ones. The overall ranking score for a blog is computed on top of all its incoming links (inlinks). Moreover, the time information, which is attached in blog posts, is exploited in order to compute hyperlink freshness and re-calculate the overall score for a blog.

In the following section we provide reference to research works on web document ranking that make use of various web page information and to works that emphasize on the additional information that blogs carry. In section 3 we give an overview of blog information and the fundamental concepts of our rating model. Section 4 presents the mathematical formulation of our proposed model and suggests a model for attaching rating semantics to blogs. Through the experimental evaluation of our designed mechanism in section 5, we demonstrate the first results from the application of our model in a collection of blogs and present some interesting findings. Finally, section 6 contains the conclusions from this work and our next plans.

2 Related Work

Ranking on the web is primarily based on the analysis of the web graph as it is formulated by hyperlinks. It has been ten years since PageRank [5], the most cited ranking algorithm, has been introduced. Several research works, during this period, have attempted to improve PageRank's performance and incorporate as many information as possible in the web graph, resulting in numerous PageRank variations and a multitude of interesting ideas (e.g., Topic-sensitive pagerank [6], Trustrank [7], Spamrank [8], Page-reRank [9], biased PageRank [10]).

The primary aim in all the aforementioned works is to attach extra semantics to hyperlinks, by analyzing neighboring content, or other structural information (e.g. topic, negative or positive opinion, etc.). In addition to automatically extracted semantics, several hyperlink metadata formats have been proposed, which allow web content authors to annotate hyperlinks [11], [12], [13] and search engines to distinguish between links that provide positive and negative recommendations. However, none of these metadata formats has yet been widely employed, and as a result, there is still not a widely accepted method for distinguishing between positive and negative links.

In the case of blogs, several ranking algorithms have been suggested that exploit explicit (EigenRumor algorithm [14]) and/or implicit (BlogRank [15], [16]) hyperlinks between blogs. All these algorithms formulate a graph of blogs, based on hyperlinks and then apply PageRank or a variation of it in order to provide an overall ranking of blogs. However, all these algorithms provide a static measure of blog importance that does not reflect the temporal aspects accompanying the evolution of the blogosphere.

Several models that capture the *freshness* of links have been proposed with applications in web pages and hyperlinks [17], [18], scientific papers and bibliographic citations [19], [20]. All these works are based on the fact that PageRank and its variations favor old pages. In order to balance this, a link (or citation) weighting scheme is employed, which is based on the age of the web page (or paper). In a post-processing step the authority of a pointed node decays, based on the node's age and the incoming links age.

In the current work, we consider that ranking in the blogosphere is an iterative process. As a first step, we consider that links in the blogosphere act as recommendations to readers. In a second step, we exploit two special features of the blogosphere links: a) the difference between blogroll links, which denote a more permanent trust towards the blog being pointed, and post links, which represent a more transient reference to a blog, b) the timestamp information of a post, which can be employed as a timestamp for a hyperlink.

3 Background

This section illustrates the useful information that can be found in a blog and can be incorporated in the blog rating mechanism. In the following, we explain the details of each piece of information; we discuss its availability and its role in the iterative rating model.

3.1 Blog Structure

Although the blog structure is not standard, most blogs share the following structure: Each blog has a host URL and contains one or more posts, authored by the blog editors. Post information comprises an author, a body, a date and time of publishing and a URL of the full, individual article, called the permalink. A post optionally includes: comments of readers, category tags and links to referrers (trackback links).

The number of ***comments*** and ***trackbacks***, where available, can be retrieved by processing the contents of each post. Since this type of information is not standard for all blog servers, the numbers can be retrieved for a small portion of the blogosphere (research works report that less than 1% of posts offers trackbacks and comments information [15]). ***Topic*** information is available for more posts ([15] report a number close to 24%). Although the choice of topic is subjective to the author, through the combined analysis of topic and author information we may

obtain useful information from of the blogosphere, such as authors that link to other authors, linked–related topics etc.

The ***date*** and ***time*** that an entry was registered is another useful piece of information. Analysis of entries based on date and time, will reveal more or less recent blogs, more or less active blogs and authors, and topics with short or long lifecycle.

Finally, the ***blogroll***, the list of blogs that is usually placed in the sidebar of a blog, can be used as a list of recommendations by the blogger of other blogs. Blogroll is considered to be a fixed list of links that is updated infrequently. Blogrolls can be used to indicate the affiliated blogs of a certain blog.

3.2 Hyperlinks and the Blog Rating Model

The aim of the proposed blog rating model is to adaptively assign a score to every blog based on the recommendations from other blogs. Each blog contains: a blogroll, which is a set of hyperlinks to affiliated blogs, and one or more posts, published at different times that contain hyperlinks to the posts of other blogs. The model distinguishes between these two hyperlink types as depicted in Figure 1.

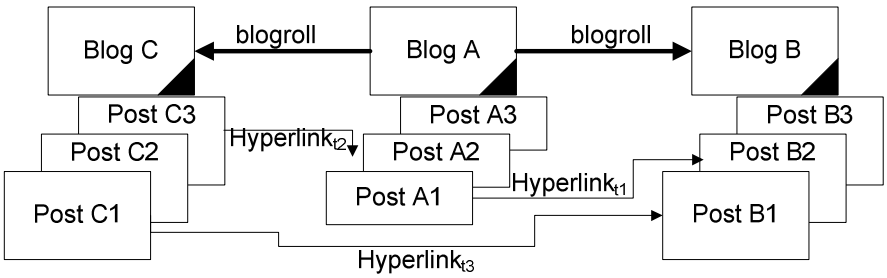

Fig. 1 Hyperlink types in the blog rating model

More specifically, a ***blogroll hyperlink*** is a link in the blogroll of blog A pointing to a blog B. It denotes that A gives a permanent recommendation for B and thus contributes a constant degree to the score of B. On the contrary, a ***post hyperlink*** from an individual post A1 of blog A to a post B2 in blog B denotes a temporary interest of A to the contents of B and consequently increases the score of B only for a short period of time after the post has been published.

The blogroll links of blog A increase the rating of all pointed blogs. Moreover, any new posts, which are added daily in blog A, contribute to the rating of the respective blogs they point to. As a consequence, the ***local rating*** assigned by a blog A to a blog B is the weighted sum of ratings assigned by blogroll and post hyperlinks respectively. This local rating information depicts the image of X for the part of the blogosphere pointed by X. This information is updated every time a new hyperlink appears, either in a post or in the blogroll of X. The rating assigned by a certain post hyperlink decreases as days pass and the post becomes old. By

monitoring a certain blog X for several periods, we are able to compute the ***accumulative local ratings*** assigned to all blogs pointed by blog X.

The local rating information of a blog A can be enhanced by the information provided by its affiliated blogs F_A (e.g., the blogs in its blogroll). The affiliated blogs are the trusted blogs of A and their opinion for the blogosphere is of interest to A. As a result, the ***collaborative local rating*** combines the direct experiences of the evaluator blog A for B with information regarding B gathered from the *N* affiliated witness blog sites. If we consider that in Figure 1 the blogs C and A collaborate, then the collaborative rating of blog B is a weighted sum of local ratings provided for B by A and C.

In a similar manner, we are able to compute a ***global rating*** for every blog, by aggregating the local rating information of all blogs. Every new blog Y that is added to the blogosphere receives a default minimum global rating. This score increases by the number of incoming blogroll or post hyperlinks and is an indication of the blog's credibility when it is used as a witness to other blogs.

In general, when we rate a service by combining multiple witnesses, we take into account the credibility of the witness and the freshness of information. In an analogous manner, when we combine local ratings from different blogs we must consider the freshness of ratings, which corresponds to a) the freshness of links and b) the freshness of prior rating (i.e., considering the time period during which the rating was estimated) and the credibility of each individual blog, which in context of this study for the global rating formation is depicted in the blog's global rating on a previous period. For example, the rating of Blog B, in figure 1, is subject to the local ratings from A to B and from C to B, weighted by the global ratings of A and C in the previous known period. The former is the sum of ratings assigned via the blogroll link and Hyperlink_{t1}, where as the latter is based only on Hyperlink_{t3}, which however is more recent than Hyperlink_{t1} (given that t3>t2>t1). If no more post links are added in the next period, the global rating of B decreases, since the freshness of Hyperlink_{t1} and Hyperlink_{t3} decreases. If no more post links are added for several periods, then the global rating of B is subject only to the rating assigned by the blogroll link of blog A.

4 Blog Site Rating System Formulation

Let us assume the presence of M Blog Sites *BS*s falling within the same category with respect to the topics covered and the interests shared. Let $BS = \{BS_1, BS_2, ...BS_M\}$ be the set of Blog Sites in the system. In subsection 4.1, the local blog site rating formation is formally described taking into account only first hand information (i.e. what the evaluator blog site considers about the target blog site), in subsection 4.2, the blog site local rating is collaboratively formed (the evaluator blog site takes into account the opinion of other affiliated blog sites concerning the blog site under evaluation) , while in subsection 4.3, a global value for a blog site is formed taking into account the view of all blog sites in the system.

4.1 Local Accumulative Blog Site Rating Formation

Concerning the local formation of the Blog Site BS_i rating, the Blog Site BS_j may rate BS_i at time period c in accordance with the following formula:

$$LABSR_{t_p=c}^{BS_j}(BS_i) = \sum_{\substack{k=c-n+1 \\ k>0}}^{c} w_{t_p=k} \cdot LBSR_{t_p=k}^{BS_j}(BS_i) \ . \quad (1)$$

where $LABSR_{t_p=c}^{BS_j}(BS_i)$ is the local accumulative BS_i rating estimated by BS_j at time period $t_p = c$, $LBSR_{t_p=k}^{BS_j}(BS_i)$ denotes the local rating the evaluator BS_j attributes to the target BS_i at time period $t_p = k$ and weight $w_{t_p=k}$ provides the relative significance of the $LBSR_{t_p=k}^{BS_j}(BS_i)$ factor estimated at time period k to the overall BS_i rating estimation by the evaluator BS_j .

Concerning the $LBSR_{t_p=k}^{BS_j}(BS_i)$ factor estimation, the evaluator BS_j may exploit the following formula:

$$LBSR_{t_p=k}^{BS_j}(BS_i) = w_{BR} \cdot BR_{t_p=k}^{BS_j}(BS_i) + w_{EP} \cdot EP_{t_p=k}^{BS_j}(BS_i) \ . \quad (2)$$

As may be observed from Equation 2, the local rating of the target BS_i is a weighted combination of two factors. The first factor contributing to the overall BS_i rating value (i.e., $BR_{t_p=k}^{BS_j}(BS_i)$) forms the blogroll related factor. This factor is introduced on the basis that the BS_j blogroll provides a list of friendly blog sites frequently accessed/read by the authors of BS_j. It has been assumed that $BR_{t_p=k}^{BS_j}(BS_i)$ lies within the [0,1] range, where a value close to 1 indicates that the target BS_i is a friendly blog site to the evaluator BS_j. In the context of this study, $BR_{t_p=k}^{BS_j}(BS_i)$ is modeled as a decision variable assuming values 1 or 0 depending on whether BS_i belongs to the blogroll of BS_j or not at time period k, respectively. Alternatively, BS_j could provide a rating of the friendly blog sites in the blogroll, which could be exploited in order to differentiate $BR_{t_p=k}^{BS_j}(BS_i)$ factor for the friendly blog sites comprised in the BS_j blogroll. This issue will be considered in a future version of this study.

The second factor contributing to the overall $LBSR_{t_p=k}^{BS_j}(BS_i)$ (i.e. $EP_{t_p=k}^{BS_j}(BS_i)$) depends on the fraction of BS_j posts pointing to BS_i at time period k. This factor has been assumed to lie within the [0,1] range and may be given by the following equation:

$$EP_{t_p=k}^{BS_j}(BS_i) = \frac{NoP_{t_p=k}^{BS_j}(BS_i)}{NoP_{t_p=k}^{BS_j}}. \tag{3}$$

where $NoP_{t_p=k}^{BS_j}(BS_i)$ denotes the number of posts created between time period $t_p = k-1$ and $t_p = k$ pointing to the target blog site BS_i and $NoP_{t_p=k}^{BS_j}$ denotes the total number of the evaluator BS_j posts created in between time period $k-1$ and k.

Weights w_{BR} and w_{EP} provide the relative significance of the anticipated blogroll related part and the posts related factor. It is assumed that weights w_{BR} and w_{EP} are normalized to add up to 1 (i.e., $w_{BR} + w_{EP} = 1$). From the aforementioned analysis, it is obvious that the $LBSR_{t_p=k}^{BS_j}(BS_i)$ factor lies within the [0,1] range.

Weights $w_{t_p=k}$ in equation (1) are normalized to add up to 1 ($\sum_{\substack{k=c-n+1 \\ k>0}}^{c} w_{t_p=k} = 1$) and may be given by the following equation:

$$w_{t_p=k} = \frac{w_k}{\sum_{l=1}^{n} w_l}. \tag{4}$$

where $w_k = \begin{Bmatrix} n-c+k, & c \geq n \\ k, & c < n \end{Bmatrix}$.

At this point it should be noted that the authors have assumed that the local rating estimation takes place at consecutive, equally distributed, time intervals. For the formation of the local accumulative BS rating at a time period c, the evaluator considers only the n more recent ratings formed. The value n determines the memory of the system. Small value for the n parameter means that the memory of the system is small, whereas large value considers a large memory for the system. Equation (4) in essence models the fact that more recent local BS ratings should weigh more in the overall BS rating evaluation.

4.2 Collaborative Local Blog Site Rating Formation

In order to estimate the rating of a target Blog Site BS_i, the evaluator Blog Site BS_j needs to contact a set WBS of N witness Blog Sites ($WBS \subseteq BS$) in order to get feedback reports on the usability of the BS_i. The set of the N witnesses is a subset of the $BS = \{BS_1, BS_2, ...BS_M\}$ set and can be the blog sites in the blog roll of BS_j. The target BS_i overall collaborative rating $CLBSR_{t_p=c}^{BS_j}(BS_i)$ may be estimated by the evaluator Blog Site BS_j at time period c in accordance with the following formula:

$$CLBSR_{t_p=c}^{BS_j}(BS_i) = w_{t_p=c}^{BS_j}(BS_j) \cdot LABSR_{t_{p=c}}^{BS_j}(BS_i) + \sum_{\substack{k=1 \\ k \neq i}}^{N} w_{t_p=c}^{BS_j}(BS_k) \cdot LABSR_{t_p=c}^{BS_k}(BS_i) \quad . \tag{5}$$

As may be observed from equation (5), the collaborative rating of the target BS_i is a weighted combination of two factors. The first factor contributing to the rating value is based on the direct experiences of the evaluator blog site BS_j, while the second factor depends on information regarding BS_i past behaviour gathered from the N witnesses blog sites.

Weight $w_{t_p=c}^{BS_j}(BS_x)$ provides the relative significance of the rating of the target blog site BS_i as formed by the blog site BS_x to the overall rating estimation by the evaluator BS_j. In general, $w_{t_p=c}^{BS_j}(BS_x)$ is a measure of the credibility of witness BS_x and may be a function of the local accumulative blog site rating attributed to each BS_x by the evaluator BS_j. It has been assumed that weights $w_{t_p=c}^{BS_j}(BS_x)$ are normalized to add up to 1 (i.e., $w_{t_p=c}^{BS_j}(BS_j) + \sum_{\substack{k=1 \\ k \neq i}}^{N} w_{t_p=c}^{BS_j}(BS_k) = 1$). Thus, weight $w_{t_p=c}^{BS_j}(BS_x)$ may be given by the following equation:

$$w_{t_p=c}^{BS_j}(BS_x) = \frac{LABSR_{t_p=c}^{BS_j}(BS_x)}{\sum_{BS_x \in WBS \cup BS_j} LABSR_{t_p=c}^{BS_j}(BS_x)} \quad . \tag{6}$$

where $LABSR_{t_p=c}^{BS_j}(BS_x)$ is the local accumulative blog site rating attributed to Blog Site BS_x by the evaluator BS_j. One may easily conclude that for the evaluator BS_j it stands $LABSR_{t_p=c}^{BS_j}(BS_j)=1$.

At this point it should be noted that, considering different blog sites, the duration of each time interval introduced in subsection 4.1 for the local accumulative blog site rating estimation may differ. This has the side-effect that it is not necessary all witness blog sites to have estimated their local accumulative rating concerning target blog site BS_i at the same time. Let us for example consider a blog site updating local accumulative blog site ratings per month and a blog site updating the related information per day. In order to introduce the time effect in our mechanism and model the fact that more recent ratings should weigh more in the overall collaborative blog site rating estimation, equation (5) should be rewritten as follows:

$$CLBSR_{time_c}^{BS_j}(BS_i)=w_{time_c}^{BS_j}(BS_j)\cdot LABSR_{time_c}^{BS_j}(BS_i)+ \sum_{\substack{k=1 \\ k\neq i}}^{N} w_{time_{d_k}}\cdot w_{time_c}^{BS_j}(BS_k)\cdot LABSR_{time_{d_k}}^{BS_k}(BS_i)\} \quad . \tag{7}$$

where $w_{time_{d_k}}$ is a decaying parameter given by the following equation:

$$w_{time_{d_k}}=1-\frac{time_c-time_{d_k}}{time_c} \quad . \tag{8}$$

In the context of this study, $w_{time_{d_k}}$ is modeled as a polynomial function. Other functions (for example exponential) could be defined as well. As may be observed from equation (7), the bigger the quantity $time_c-time_{d_k}$, the smaller the contribution of witness blog site BS_k rating provided to the overall collaborative target BS_i rating formation. At this point, we assume that when at $time_c$ the collaborative rating is estimated by the evaluator BS_j, its local accumulative ratings have also been updated.

4.3 Global Blog Site Rating Formation

In order to estimate the global rating of a target Blog Site BS_i, a specialized blog site search engine evaluator collects the feedback reports on the usability of the

BS_i from the M Blog Sites belonging to the set $BS = \{BS_1, BS_2, ...BSs_M\}$. The target BS_i overall collaborative global rating $GBSR^{BS_j}_{t_p=c}(BS_i)$ may be estimated by the evaluator Blog Site BS_j at time c in accordance with the following formula:

$$GBSR^{BS_j}_{time_c}(BS_i) = \sum_{\substack{k=1 \\ k \neq i}}^{M} w_{time_{d_k}} \cdot w^{BS_k} \cdot LABSR^{BS_k}_{time_{d_k}}(BS_i) \ . \quad (9)$$

As may be observed from equation (9), the global rating of the target BS_i is a weighted combination of the rating values provided by the blog sites BS_k, based on the direct experiences.

Weight w^{BS_k} provides the relative significance of the rating of the target blog site BS_i as formed by the blog site BS_k to the overall rating estimation by the evaluator blog site search engine. In general, w^{BS_k} is a measure of the credibility of blog site BS_k and may be a function of its prior global rating as estimated by the evaluator blog site search engine during the previous time. In the context of this study, weight w^{BS_k} is given by the following equation:

$$w^{BS_k} = GBSR^{BS_k}_{time_{c-1}}(BS_i) \cdot \frac{NoBS(BS_k)}{M} \ . \quad (10)$$

where $GBSR^{BS_k}_{time_{c-1}}(BS_i)$ denotes the prior global rating of blog site BS_k estimated at $time_{c-1}$ in accordance with equation (9), $NoBS(BS_k)$ denotes the number of blog sites pointing to BS_k at $time_c$ and M is the total number of blog sites in the system. The portion of the blog sites pointing to BS_k at time c has been introduced in equation (10) in order to enhance the credibility value of a witness blog site providing a rating for the blog site BS_i under evaluation. Finally, in analogy to subsection 4.2, parameter $w_{time_{d_k}}$ is the decaying factor given by equation (8), introduced in order to weigh down possible outdated evaluation ratings provided.

4.4 The Semantics of the Rating Model

In order to support the operation of the rating mechanism, we suggest the use of semantics in the description of post and blogroll hyperlinks. The semantic

information which will be attached to each hyperlink will allow bloggers to better describe their intentions behind creating the link, to prioritize affiliated blogs in the blogroll or even to provide topic information for the pointed posts. The rating mechanism can be adopted to update the local scores, and to employ them in providing collaborative and global scores. RDF is a popular format for describing metadata and it is used to support our rating model.

As mentioned in section 4.1, the local (accumulative or not) blog site ratings are solely based on the recommendations provided by the blog itself. As a result, only the RDF file associated with the current blog is required for storing local ratings. In general, the RDF file comprises URI and ratings for each blog in the blogroll and for each blog referenced in the posts. A software entity acting on behalf of each blog is responsible for reading the RDF file, recomputing the accumulative localRating and updating the RDF with the new ratings and the new date of update. In the following, we provide a fictional example of an RDF file for a blog (e.g., my.blog.co.uk) that contains two links in the blogroll (e.g., myother.blog.co.uk and blog.co.uk/agoodone) and two post with hyperlinks to an affiliated blog (e.g., blog.co.uk/agoodone) and a non-affiliated blog (e.g., anyblog.co.uk), respectively.

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:blog="http://www.blogtrust.fake/blog#">
<rdf:Description rdf:about="http://my.blog.co.uk/">
<blog:OutLinkTo>
        <rdf:Description rdf:about="http://myother.blog.co.uk/">
                <blog:blogroll blog:localRating="1" blog:dateUpdated="2008-1-1"/>
        </rdf:Description>
</blog:OutLinkTo>
<blog:OutLinkTo>
        <rdf:Description rdf:about="http://blog.co.uk/agoodone">
                <blog:blogroll blog:localRating="0.6" blog:dateUpdated="2009-5-1"/>
                <blog:postlink rdf:parseType="Resource">
                        <blog:localRating>0.9</blog:localRating>
                        <blog:dateUpdated>2008-11-1</blog:dateUpdated>
                        <blog:sourcepermalink>http://my.blog.co.uk/2008-11-
1</blog:sourcepermalink>
                        <blog:targetpermalink>http://blog.co.uk/agoodone/2008-10-
28</blog:targetpermalink>
                </blog:postlink>
        </rdf:Description>
</blog:OutLinkTo>
<blog:OutLinkTo>
        <rdf:Description rdf:about="http://anyblog.co.uk/">
                        <blog:postlink rdf:parseType="Resource">
                                <blog:localRating>0.8</blog:localRating>
                                <blog:dateUpdated>2008-12-15</blog:dateUpdated>
                                <blog:sourcepermalink>http://my.blog.co.uk/2008-12-
15</blog:sourcepermalink>
                                <blog:targetpermalink>http://anyblog.co.uk/2008-12-
12</blog:targetpermalink>
                        </blog:postlink>
        </rdf:Description>
</blog:OutLinkTo>
</rdf:Description>
</rdf:RDF>
```

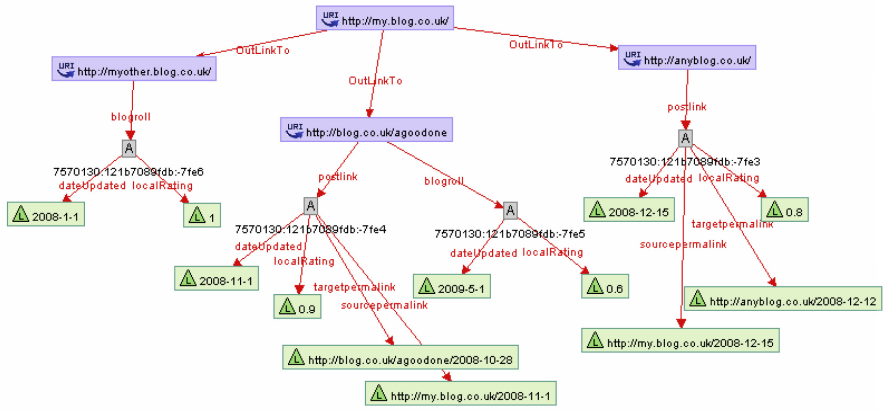

Fig. 2 The RDF structure for a blog

On the other side, the local collaborative blog site rating is subject to the blog's RDF, but also to the RDF files of all other blogs in its blogroll, assuming that the witness set is constituted by the blog sites comprised in the blogroll. Moreover, the collaborative process will take into account the rating of each external recommendation. The rating is available in the original RDF file and the external recommendations can be retrieved from the respective RDF files. In such case the rating mechanism for a blog should process the blog's current ratings and those provided by each of the affiliated blogs.

Finally, the rating mechanism must collect and process the RDF metadata files from all blogs in the set in order to calculate the global blog site rating, This process is repeated periodically, so as to keep the rating up to date.

In the computation of collaborative and global rating, the mechanism should take into account two factors that affect the transitivity of rating: a) the effect of ratings provided by the affiliated blogs depends upon their credibility (i.e., the picture the evaluator blog site has formed about them), b) the rating contributed by a certain postlink decreases day by day and c) the prior rating estimated during the previous time period decreases in order to weigh down possible outdated evaluation ratings provided. The latter factor is captured by the time decay factor of equation 8, whereas the former is captured by the respective weight $w_{t_p=c}^{BS_j}(BS_x)$ in equation 5. The *localRating* and *dateupdated* values of the postlink are employed to store these two factors.

5 Experimental Setup

In order to demonstrate the blog rating model, we performed experiments on a sample blog dataset provided by Nielsen BuzzMetrics, Inc. The dataset spans a period before and after an important event: the London bombings (4/7/2005 – 24/7/2005). Table 1 that follows summarizes the statistics of the dataset:

Table 1 Statistics of the sample blog set

Unique blogs number	Links to any blog	Links to blogs in the set	Links to news sites
1,545,205	2,138,381	331,068	498,834

It is obvious from Table 1 that the majority of the links points to blogs outside of the initial set and a large portion of the links points to news sites. The blogs that are outside of the initial set can probably be spam blogs (splogs), which are massively pointed by blogs in the set in an attempt to improve their ranking.

We perform three experiments on the same dataset: a) We find the most referenced blogs and news site for a single day using inlinks only, b) we rank sites according to the global rating using information for a single day and compare results with those of the first experiment, c) we apply the global rating model in the blogs using the posts of a single day, using different values for the memory factor and compare the position of spam blogs in the different sets of ranked results. As it is explained in the analysis of the results, our rating model penalizes the spam blogs, even for small values of the memory factor.

Results in Table 2 contain the top-20 blogs ranked using the number of incoming links as the rating factor. According to these results, the most popular sites on the first and the last day in the dataset comprise news sites and spam blogs (positions 13 to 20 on 4/7/2005 and 11 to 20 on 24/7/2005). Although news sites are acceptable in the top ranked results, the spam blogs should be penalized by the rating model.

Table 2 Most referenced sites in the dataset for the 4th and 24th of July 2005

Most referenced sites (4/7)			Most referenced sites (24/7)		
Rank	URL	Inlinks	Rank	URL	Inlinks
1	www.livejournal.com	9511	1	www.livejournal.com	3450
2	spaces.msn.com	2724	2	www.xanga.com	2502
3	www.xanga.com	2503	3	spaces.msn.com	1229
4	www.skaterz.info	1647	4	news.yahoo.com	1070
5	news.yahoo.com	1399	5	www.nytimes.com	1039
6	news.bbc.co.uk	1127	6	news.bbc.co.uk	841
7	www.nytimes.com	1092	7	pics.livejournal.com	535
8	www.cnn.com	563	8	www.washingtonpost.com	513
9	www.washingtonpost.com	560	9	biz.yahoo.com	443
10	pics.livejournal.com	530	10	www.bbc.co.uk	440
11	www.msnbc.msn.com	451	11	miss-usa-teen.blogspot.com	361
12	www.guardian.co.uk	389	12	nude-thumbnails.blogspot.com	361
13	fantasy-fest-nude.blogspot.com	376	13	nude-girls-thumbnails.blogspot.com	361
14	top-play-lolita.blogspot.com	376	14	asian-nude-thumbnails.blogspot.com	361
15	lolita-top-sites.blogspot.com	376	15	non-nude-teen-photos.blogspot.com	361
16	hardcore-lesbian-pictures.blogspot.com	376	16	amateur-teen-nude.blogspot.com	361
17	lesbian-kissing-pictures.blogspot.com	376	17	nude-amateur-photos.blogspot.com	361
18	naturist-teen-photos.blogspot.com	376	18	photos-amateur-gratuites.blogspot.com	361
19	funny-as-shit.blogspot.com	376	19	breast-pumps-reviews.blogspot.com	361
20	really-funny-shit.blogspot.com	376	20	young-naked-gay-boys.blogspot.com	361

The first step towards correcting this problem is to use our rating model instead of the number of inlinks. The local ratings are computed on a per blog basis. Thereafter, the global rating for all sites is estimated, using the accumulative

algorithm. A useful observation is that spam blogs usually receive a large number of inlinks the day they are created, but they further receive no inlinks, so it is expected that spam blogs will receive lower ratings by our model. The results in Table 3 show the ranking of blogs in the dataset according to their global rating in the 4th of July. Since this is the first day in our dataset, the global rating is computed using the inlinks of this specific day (memory equals to zero, m=0). It is obvious from the results in Table 3 that news sites have improved their ranking against all other blogs (including spam blogs).

Table 3 Top-20 ranked sites in 4/7 using global rating (m=0)

Most referenced sites		
Rank	URL	Rank in 4/7 using inlinks
1	www.livejournal.com	1
2	news.yahoo.com	5
3	news.bbc.co.uk	6
4	www.bbc.co.uk	120
5	www.nytimes.com	7
6	www.cnn.com	8
7	www.washingtonpost.com	9
8	biz.yahoo.com	118
9	pics.livejournal.com	10
10	www.msnbc.msn.com	11
11	www.guardian.co.uk	12
12	www.latimes.com	123
13	www.usatoday.com	238
14	livejournal.com	246
15	today.reuters.com	261
16	www.sfgate.com	122
17	www.boston.com	173
18	www.forbes.com	178
19	www.timesonline.co.uk	174
20	www.newsday.com	266

As mentioned before, spam blogs usually receive a large number of links in a single day, which explains their ranking in the results of Table 2. However, these incoming links have a single origin (another spam blog), which has been created for this reason. According to equation 3, the contribution of blogs that contain many links is small and consequently spam blogs of this type receive a small local rating. However, there are still spam blogs that receive fabricated inlinks from different origins in the same time. In order to penalize these links we examine the blogosphere for several days (i.e., 20 days) using our model with memory (i.e., local accumulative blog site rating formation considering 20 time periods – 20 days).

In Table 4, we present the top-20 ranked blog urls in the dataset (urls that contain the term 'blog') for the 24th of July, which is the last date in the set. The blogs are rated using the maximum possible memory in our dataset (m=20). The rightmost column of Table 4 contains the number of inlinks for each blog in the 24th of July and the middle column contains the position of this blog in the same date, ranked using only the number of inlinks. It is obvious that normal blogs rank higher when collective memory from the previous days is employed and surpass the spam blogs.

Table 4 Most highly ranked blogs in the 24th of July (global rating, m=20)

Most referenced blogs			
Rank	URL	Rank in 24/7 using inlinks	Inlinks in 24/7
1	radio.weblogs.com	199	88
2	blog.livedoor.jp	222	65
3	blogs.sun.com	318	26
4	postsecret.blogspot.com	348	22
5	blog.searchenginewatch.com	1204	4
6	www.blogathon.org	308	28
7	atrios.blogspot.com	302	30
8	blogs.salon.com	413	17
9	www.problogger.net	523	12
10	doc.weblogs.com	480	14
11	www.blogherald.com	519	12
12	profiles.blogdrive.com	253	47
13	powerlineblog.com	290	32
14	www.bloggingbaby.com	258	45
15	googleblog.blogspot.com	444	15
16	americaninlebanon.blogspot.com	1000	6
17	blogs.guardian.co.uk	1431	4
18	badhairblog.blogspot.com	496	13
19	hurryupharry.bloghouse.net	995	6
20	www.captainsquartersblog.com	267	40

A point of interest in the results is that professional blogs, such as Sun's blog or Google's blog are ranked high when global rating is employed considering accumulative local blog site rating, although they receive few links in a single day (low ranking using a single day's links). However, such blogs: receive links in a daily basis, are the single targets of the post each time (in contrast to spam blogs) and are pointed by highly rated blogs.

Table 5 The effect of memory size in spam blog global ranking

Memory	**Inlinks**	**1**	**3**	**5**	**7**
Global ranking for the first set of spam blogs first – last in the set	47 - 292	6298 - 12250	8207 - 19672	12157 - 21587	15019-23126
Best position for a spam blog	47	6298	7124	6699	7882

In a third set of experiments, we examine the ranking of the 53383 blogs of our blogosphere part in the 14th of July (the date was selected because it is in the middle of the period examined) using five different values for the system's memory: a) we consider that for memory equaling zero, only the inlinks created on the specific date affect rating, b) we take into account the postlinks provided at most *m* (m=1,3,5,7) days before the 14th of July. We manually examine the set of results to find the position of the first spam blog in the global ranking of blogs. As it is portrayed in Table 5, the first, from a set of spam blogs (ranging from the 49th to the 292nd position), falls below the 6298th position when the local ratings of the current day are employed in the calculation of global rating. It falls even lower for bigger values of *m*, although the change is smaller.

6 Conclusions

This work presented an iterative collaborative process to provide a global rating for a set of blogs using local rating information expressed via blogroll and post hyperlinks. The rating model is mathematically formulated, comprising local and local accumulative blog site rating formation (where the accumulative rating is calculated considering the local rating as estimated upon different consecutive time periods), collaborative local blog site formation (where the evaluator blog site exploits information gathered form other affilitated witnesses blog sites) and global rating formation, incorporating the view of all blog sited in the system. Our model exploits two special features of the blogosphere: a) the difference between blogroll links, which denote a more permanent trust towards the blog being pointed, and post links, which represent a more transient reference to a blog, b) the timestamp information of a post, which can be employed as a timestamp for a hyperlink. Additionally, a suggestion on the semantics that can be attached to each blog is also provided.

An initial experimental evaluation shows that the model performs well by punishing spam blogs that receive many links from a single source and favouring blogs that receive inlinks in a standard basis. The next steps of this work is to develop the architecture and the system entities that estimate and attach the rating information to blogs and that process local ratings in a periodic manner in order to update collaborative and global ratings. Future work additionally includes incorporation of possible postlink negative recommendations.

References

1. Blogpulse. Automated trend discovery system for blogs (2005), http://blogpulse.com/ (accessed May 2009)
2. Technorati, Blog tracking service (2005), http://technorati.com/ (accessed May 2009)
3. Mishne, G.: Information Access Challenges in the Blogspace. In: IIIA-2006: International Workshop on Intelligent Information Access, Helsinki, Finland (2006)
4. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Now Publishers (July 2008) ISBN 978-1-60198-150-9
5. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998)
6. Haveliwala, T.: Topic-sensitive PageRank. In: Proceedings of the Eleventh International World Wide Web Conference, Honolulu, Hawaii, May 2002, pp. 517–526 (2002)
7. Gyöngyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with TrustRank. In: Proceedings of the 30th International Conference on Very Large Data Bases (VLDB), Toronto, Canada, September 2004, pp. 271–279 (2004)
8. Benczur, A.A., Csalogany, K., Sarlos, T., Uher, M.: SpamRank - fully automatic link spam detection. In: Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web, AIRWeb (2005)

9. Massa, P., Hayes, C.: Page-rerank: using trusted links to re-rank authority. In: Proceedings of Web Intelligence Conference, France (September 2005)
10. Jeh, G., Widom, J.: Scaling personalized web search. In: Proceedings of the Twelfth International World Wide Web Conference, Budapest, Hungary, May 2003, pp. 271–279 (2003)
11. Technorati.com. VoteLinks, http://developer.technorati.com/wiki/VoteLinks
12. Technorati.com. XFN (Xhtml Friends Network), http://gmpg.org/xfn/
13. Varlamis, I., Vazirgiannis, M.: Web Document Searching. Using Enhanced Hyperlink Semantics Based on XML. In: Proceeding of the International. Database Eng. & Applications Symposium (IDEAS 2001), pp. 34–43 (2001)
14. Nakajima, S., Tatemura, J., Hino, Y., Hara, Y., Tanaka, K.: Discovering Important Bloggers based on Analyzing Blog Threads. In: 2nd Annual Workshop on the Blogging Ecosystem: Aggregation, Analysis and Dynamics, WWW 2005 (2005)
15. Kritikopoulos, A., Sideri, M., Varlamis, I.: BlogRank: ranking blogs based on connectivity and similarity features. In: Proceedings of the 2nd international Workshop on Advanced Architectures and Algorithms For internet Delivery and Applications, AAA-IDEA 2006, Pisa, Italy, October 10, vol. 198. ACM, New York (2006), http://doi.acm.org/10.1145/1190183.1190193
16. Adar, E., Zhang, L., Adamic, L., Lukose, R.: Implicit Structure and the Dynamics of Blogspace. In: Workshop on the Blogging Ecosystem: Aggregation, Analysis and Dynamics, WWW 2004 (2004)
17. Amitay, E., Carmel, D., Herscovici, M., Lempel, R., Soffer, A.: Trend Detection Through Temporal Link Analysis. Journal of the American Society for Information Science & Technology 55(14), 1270–1281 (2004)
18. Bar-Yossef, Z., Broder, A., Kumar, R., Tomkins, A.: Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay. In: Proceedings of the 13th International Conference on World Wide Web, pp. 328–337 (2004)
19. Berberich, K., Vazirgiannis, M., Weikum, G.: Time-Aware Authority Ranking. Internet Mathematics Journal 2(3) (2005)
20. Yu, P.S., Li, X., Liu, B.: On the Temporal Dimension of Search. In: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters, pp. 448–449. ACM Press, New York (2004)