The Learning Organization

Quality assessment in the blog space

Markus Schaal, Guven Fidan, Roland M. Müller, Orhan Dagli,

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0969-6474.htm

Quality assessment in the blog space

Markus Schaal
Department of Computer Engineering, Bilkent University, Ankara, Turkey
Guven Fidan
AGMLAB Information Technologies, METU Technopolis, Ankara, Turkey
Roland M. Müller
Department of Information Systems & Change Management, University of Twente, Enschede, The Netherlands, and
Orhan Dagli
Department of Computer Engineering, Bilkent University, Ankara, Turkey

Abstract

Purpose – The purpose of this paper is the presentation of a new method for blog quality assessment. The method uses the temporal sequence of link creation events between blogs as an implicit source for the collective tacit knowledge of blog authors about blog quality.

Design/methodology/approach – The blog data are processed by the novel method for the assessment of blog quality. The results are compared to Google Page Rank with respect to the Gold Standard, the BlogRazzi Bookmark Rank.

Findings – The method is similar or better than Google Page Rank with respect to the chosen Gold Standard.

Originality/value – The major contribution of this paper is the introduction of a novel method for blog quality assessment. Even though its superiority to other and more established methods cannot be proven in the context of this limited study, it enriches the toolset available for blog quality assessment and may become important for a deeper understanding of organizational learning.

Keywords Information management, Quality assessment, Communication technologies

Paper type Technical paper

1. Introduction

With the emergence of Web 2.0 applications, where information is not only disseminated from trusted sources across the net, but also anonymously published, syndicated, evaluated, selected, recombined, and edited, information quality assessment becomes crucial. This is particularly true for the blog space, which emerges as a popular means for knowledge sharing.

Quality is a result of a quality creation process and thus can be best observed by a temporal analysis of the events leading to the emergence and evolution of the content. Based on a general model for an event-based assessment of information quality in a shared knowledge space, we will describe a formalization of the blog space and an algorithm for the assessment of quality within this model.

2. Previous research

PageRank (Brin and Page, 1998) and, based on a similar intuition, online page importance computation (OPIC; Abiteboul *et al.*, 2003) provide measurements for the

The Learning Organization
Vol. 17 No. 6, 2010
pp. 529-536

0969-6474
DOI 10.1108/09696471011082385

importance of a web page based on the link structure among web sites. The underlying idea is simple: important pages link to other important pages.

In PageRank, the importance, namely authority of a page depends on both the number of incoming links to the page and the importance of the pages which give those links. Google describes the concept with a non-egalitarian voting mechanism, where a link from one page is interpreted as a vote for this page. However, links from important pages are weighted higher than links from unimportant pages. The calculated PageRank value represents the probability of arriving at the particular page by clicking randomly links. Therefore, the PageRank algorithm can be understo as a Markov Chaing with states (pages), and equally probable transitions (links) between states.

OPIC – contrary to PageRank – is as an online algorithm, which does not require offline computation. In off-line algorithms like PageRank, there is a need to compute the importance of the pages by iterated computations, which can take a lot of time and space. In OPIC, each page has an initial score. While crawling, each page distributes its current score equally to the pages that it links. The importance is calculated by looking at the distribution logs for a page, it means the more a page gets credit from other pages, the more important it is. The off-line ranking algorithms depend on fast sparse matrix multiplication, however, OPIC does not need a full matrix to calculate the importance, it can start while matrix and web graph are still being created.

The proposed trust update model starts with a similar idea like PageRank and OPIC: a trusted blogger is likely to set a link to another trusted blog or resource, where trust is closely related to quality. As OPIC, trust update is an online algorithm, which has a lot of advantages for processing frequently updating sources like blogs or Twitter streams.

We do not assume that a link already propagates trust if there is no initial trust. In order to achieve that, we will distinguish two qualities per blog content trust and link trust. By learning trust throughout time, we can employ information about initial trust and eliminate spam bloggers by their inability to accumulate trust over time.

The notion of trust and belief is crucial for our analysis. Being investigated for multi-agent systems (Sabater and Sierra, 2005), and slowly applied to the social web (Golbeck and Hendler, 2006). The bridge between trust and information quality is about to be discovered by innovative applications (Schaal, 2006). Recommender systems and collaborative filtering (Adomavicius and Tuzhilin, 2005) are other approaches for the aggregation and mining of collective reputation, but mostly neglecting the notion of trust among the people.

3. The model

We will distinguish between the trust update model and the blogosphere model.

3.1. Blogosphere

The blogosphere consists of a set of blogs containing blog items (or blog entries). Each blog entry has a creation time and a set of links to other blog entries. We consider the creation of blog entries as events and analyze their appearance in the sequence of time. Our model of the blog space (or blogosphere) is shown in Figure 1. For simplicity, we associate each blog with its owner.

3.2. Trust update model

We model the trust in a user's feedback quality (FQ) and his content creation quality separately per user u as a function of time denoted by fqu(t) (feedback quality of user u)

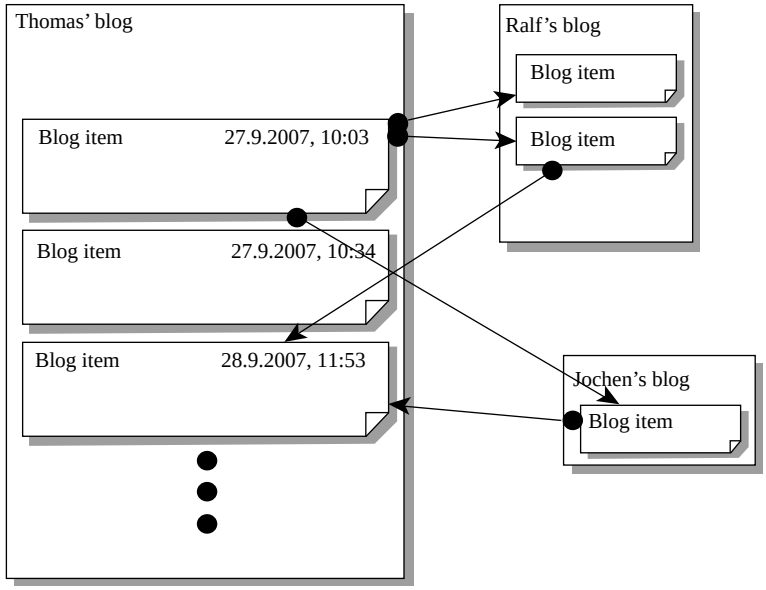

Figure 1.
Temporal Blogspace Model

and cqu(t) (content quality of user u), respectively. By this, we provide a model of trust/quality for a web user that is not limited to a single trust value. Specifically, we will learn about the quality of a web user as a source of feedback in addition to his quality as a content creator/contributor. For automatically learning these qualities, we employ a sequence of Bayesian updates throughout time while feedback about other users qualities is received.

The separation of FQ and CQ is motivated by the intuition, that spam bloggers may link to targets of high quality even though they do not provide high quality (high FQ and low CQ) while on the other hand non-spam bloggers may link to spam targets even though they provide good content (low FQ and high CQ). Without a separation of qualities, these groups could not be distinguished.

With a fine-tuned model for quality updates, we expect to distinguish quality from spam by the following differences between a social network of low quality (e.g. a linkfarm) and a social network of high quality:

- The high quality network does not link to the low quality network, while the contrary is generally not true.
- The high quality network invests more energy over a long time period than the low quality network.

3.3. Mapping blogosphere events to trust updates

We interpret the creation of links in a blog item as positive feedback about the targets. For our update model, we will sort all blog items according to their creation time and process them in temporal sequence. For each link between blog entries, we will update the link quality of the source and the content quality of the target according to the Bayesian model is shown in Figure 2.

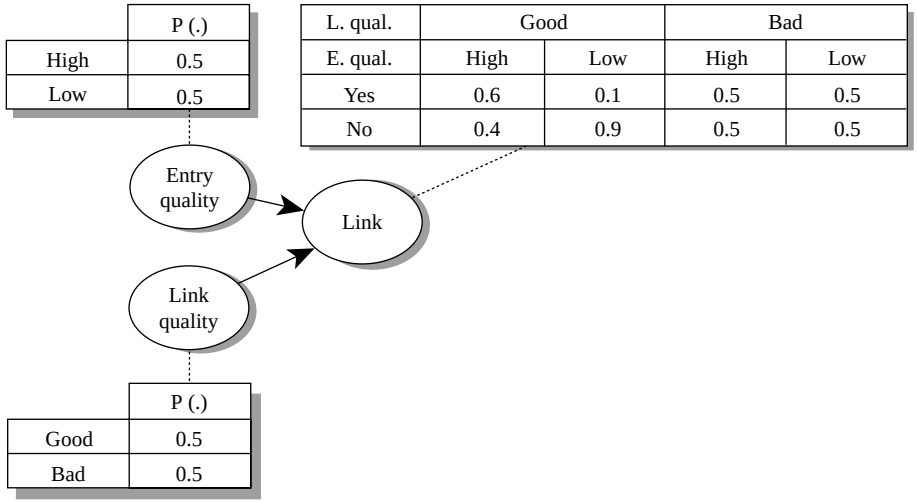

Figure 2.
Update Model for Content and Link Quality

The marginal distribution of the CQ (of the target blog entry) and the FQ of the source user (link quality) is given here only for illustration: 0.5 is the initial default value. Let i be the source (blog item) and j be the target (blog item) of a particular link. Let u and v be two blogs (or blog owners) each with a blog entry: blog entry i from blog u and blog entry j from blog v. Then, during computation of the trust updates, link quality fqu(t) and content creation quality cqv(t) will be used as a-priori of the marginal distributions, with $t = t_{link} - \varepsilon$, where t_{link} is the time of the link creation.

The conditional probabilities used for link probability (to the right of Figure 2) are initial expert estimates and need to be improved later according to a proper model. For a given link, the quality values fqu(t_{link}) and cqv(t_{link}) are the a posteriori values of the Bayesian network for the given link evidence. Owing to our mapping from links (in the blogspace) to positive feedback (in the trust update model), we will not have negative feedback.

4. Data

4.1. Blog data

The data are collected from the turkish blogosphere by AGMLAB. The data set consists of 2.3 Mio. links from turkish blogs. We used a data set with the following pieces of information per link (multiple links into the same target are grouped):

- *Target*. This can be a blog item or any other resource specified by an URL.
- *From blog entry*. The source blog item's URL.
- *From blog*. The URL of the source blog, i.e. the blog to which the item belongs.
- *To blog*. The URL of the target blog, i.e. the blog to which the target belongs. This applies only if the target is actually a blog entry. The data field remains empty otherwise.
- *Date*. The creation date of the source blog item, i.e. the date of the positive feedback implied by the link from the source blog item to the target.

A sample of the raw data is shown below.

Target:

www.one.org/news/2007/08/26

Inlinks:

(1) fromBlogEntry: www.two.org/node/850
fromBlog: www.two.org/
toBlog: www.one.org/
date: Sun Aug 26 14:46:59 EEST 2007

(2) fromBlogEntry: www.three.org/node/3
fromBlog: www.three.org/
toBlog: www.one.org/
date: Sun Aug 26 18:23:12 EEST 2007

Here, www.one.org blog has an entry with file path news/2007/08/26 and there is two incoming links from blogs www.two.org and www.three.org

4.2. Blog Razzi gold standard

For having a gold standard, we used the rankings provided by Blog Razzi, a web site dedicated to the ranking and evaluation of the blogspace. Blog Razzi creates four different scores for each blog and ranks the blogs accordingly:

(1) *User rating*. Users can rate each blog with one to five stars. The average of these ratings is taken.

(2) *Bookmark score*. Number of users that have a bookmark to a particular blog in Blog Razzi.

(3) *Blog Razzi score*. An internal score computed by Blog Razzi.

(4) *Comment score*. The number of comments users left about a particular blog.

All these scores can be used for blog ranking and we will compare our results with the various Blog Razzi rankings.

5. Experiments

The data set of temporal blog links (self-references removed) has been used to:

- Compute both CQ and FQ with our trust updates method.
- Compute Google Page Rank (GPR).
- Compute CQ-bias, FQ-bias, and GPR-bias, by starting with biased data sets. In the case of our trust updates method, high-ranked blogs received initially some additional trust. In the case of Google Page Rank, high-ranked blogs received initially some additional likelihood of page visit. A blog was considerd high-ranked, if it was on the second, fourth, sixth, etc. position in the top 100 Blog Razzi bookmark ranking (BM), i.e. 50 high-ranked blogs were used.
- Juxtapose CQ, FQ, CQ-bias, FQ-bias, and BM versus GPR, GPR-bias, and BM.

We started our experiments with the following hypotheses, which will be tested in the analysis part of the paper:

H1. Trust updates (CQ) outperform Google Page Rank (GPR) with respect to Blog Razzi bookmark score (BM), i.e. the difference between trust updates and bookmark score is less than the difference between Google Page Rank and Bookmark Score.

H2. Biased data increases the performance of trust updates more than the performance of Google Page Rank.

The first hypothesis is motivated by the expectation, that trust updates reflect users quality assessments (e.g. Bookmark Score) more accurately than the likelihood of random arrival implemented by Google Page Rank. The second hypothesis is motivated by the nature of trust updates – including reinforcement.

6. Analysis

For comparison of different quality measures, we did not consider absolute values. Instead we investigated the rankings of the inner join (blogs/users evaluated by a pair of different methods) and compared them according to the following ranking distance (Lempel and Moran 2005).

Let v, w be N-dimensional real vectors (representing rankings). The ranking distance d_r between v and w is defined as follows:

$$d_r(v,w) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} c_{v,w}(i,j)$$

with:

$$c_{v,w}(i,j) = \begin{cases} 1 & (v(i) \le v(j) \wedge w(i) > w(j)) \vee \\ & (v(i) < v(j) \wedge w(i) \ge w(j)) \\ 0 & \text{otherwise} \end{cases}$$

for example:

$$d_r([2,4,6,8],[2,9,5,3]) = \frac{3}{16}$$

Note, that this is not exactly the same formula as reported by Lempel and Moran 2005), since the original formula created weird results if applied to vectors with many duplicate values.

The following Table I gives the comparison between the Blog Razzi Bookmark Ranking (no. BM) and all experiments as well as the comparison of Google Page Rank (GPR and GPR-bias) with trust updates.

Table I. Ranking Distances

	Trust upd.		Trust upd. bias		
	CQ	FQ	CQ	FQ	No. BM
No. BM	0.27	0.36	0.26	0.36	–
GPR	0.42	0.27	0.42	0.27	0.43
GPR-bias	0.38	0.30	0.38	0.30	0.42

Note, the ranking distances with FQ are not relevant. Our aim is the assessment of the CQ of the blogs.

With regard to our hypotheses, the following observations can be made:

- The first hypothesis is clearly supported, with ranking distance of 0.27 between CQ and bookmark score for trust updates as opposed to 0.43 for the distance between GPR and bookmark score.
- The second hypothesis is not supported, with hardly any difference between biased and unbiased data for neither Google Page Rank nor trust updates.

7. Conclusion

We presented an online method for quality assessment in the blog space, which considers the feedback quality as an implicit signal of the human-edited links among blogs. We demonstrated our new method trust updates by a preliminary case study in the turkish blog space, and compared our results with Google Page Rank. As gold standard we used the bookmark rank aggregated as a collective quality measure by the collaborative Blog Razzi web site. We found our method nearer to our gold standard than Google Page Rank and interpret this as a confirmation for the feasibility of our approach.

We are looking forward to extend our approach towards being more generic for quality processing in the Web 2.0. In particular:

- We want to investigate more and novel quality dimensions for online user and CQ assessment.
- We want to incorporate known interaction patterns of quality emergence and social interaction into our assessment methods.
- We want to develop a test bed for quality assessment in the Web 2.0.

References

Abiteboul, S., Preda, M. and Cobena, G. (2003), *Adaptive On-line Page Importance Computation*, ACM Press, New York, NY, pp. 280-90.

Adomavicius, G. and Tuzhilin, A. (2005), "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17 No. 6, pp. 734-49.

Brin, S. and Page, L. (1998), "The anatomy of a large-scale hypertextual web search engine", *Computer Networks and ISDN Systems*, Vol. 30 Nos 1-7, pp. 107-17.

Golbeck, J. and Hendler, J. (2006), "Inferring binary trust relationships in web-based social networks", *ACM Trans. Inter. Tech.*, Vol. 6 No. 4, pp. 497-529.

Lempel, R. and Moran, S. (2005), "Rank-stability and rank-similarity of link-based web ranking algorithms in authority-connected graphs", *Information Retrieval*, Vol. 8 No. 2, pp. 245-64.

Sabater, J. and Sierra, C. (2005), "Review on computational trust and reputation models", *Artificial Intelligence Review*, Vol. 24 No. 1, pp. 33-60.

Schaal, M. (2006), "A bayesian approach for small information trust updates", *Proceedings of IeCCS*.

Further reading

Hinze, A. and Voisard, A. (2002), "A parameterized algebra for event notification services", *Proceedings of the 9th International Symposium on Temporal Representation and Reasoning, Manchester, UK*.

Corresponding author
Markus Schaal can be contacted at: schaal@acm.org