Discovering Weblog Communities

A Content- and Topology-Based Approach

Jeroen Bulters
ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
The Netherlands
jbulters@science.uva.nl

Maarten de Rijke
ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
The Netherlands
mdr@science.uva.nl

Abstract

Weblogs have become a leading form of self-publication on the web. Personal weblogs are often considered to represent a person, and the links between webogs can naturally be given a social interaction. Against this background, finding a community around a given weblog—i.e., identifying a set of weblogs that forms a natural group together with the starting point, because of content or social reasons—is a very natural task. Traditional methods for community finding methods focus almost exclusively on topology analysis. In this paper we present a novel method for discovering weblog communities that incorporates both topology analysis and content analysis. We evaluate our method in a small-scale user study, analyze the contributions of the various components of our approach, and compare it against a state-of-the-art topology-based community finding algorithm.

1. Introduction

In recent years weblogs have become a dominant form of self publication on the internet. The number of weblogs tracked by Technorati has been doubling every 5 months and it is often claimed that a new weblog is created every second. The vast and evolving nature of the blogosphere offers interesting challenges from the point of view of *information access*.

In this paper, we focus on the following access task: given a weblog (or blogger), return a set of other weblogs that form a community together with the starting blog. Traditional community extraction methods rely almost exclusively on an analysis of link topology around a given starting point, thereby effectively ignoring the immense amount of information given by the weblogger in his posts. For example, in the experimental evaluation in this paper one of the weblogs—*appelejan*—was assessed as having 18 members in its community; however, a state-of-the-art topology based algorithm yielded only three members of the community due to the fact that members in the community did not always link back to each other or to other members of the community.

We present a novel community finding method that incorporates both topology- and content-analysis. In addition to a detailed description of the core algorithm, we provide the outcomes of a small-scale user study aimed at understanding the algorithm's effectiveness and at comparing it with an existing state-of-the-art solution.

ICWSM 2007 Boulder, CO USA

We believe that our work is of interest to two types of end users: (1) the algorithm we propose lays the ground work for a tool that can used by individual bloggers as an exploratory search tool, and (2) our algorithm can be extended to a tool for advertisers and marketeers, for whom a global view of likes, dislikes, and interests of groups of bloggers matters.

The remainder of this paper is organized as follows. We start with a brief description of related work in Section 2. Then, in Section 3, we present our algorithm for discovering weblog communities. We follow with a description of an experimental evaluation of the algorithm in Section 4. We report on the results in Section 5 and conclude in Section 6.

2. Related work

The fact that a weblog is a web-based publication gives us the opportunity to apply traditional web-mining techniques to weblogs. A lot of work has been done on the identification of clustered websites; see e.g., [2]. Although weblogs are just websites, weblogs are often considered to "represent" a person while a website represents a subject [5]. Websites can be characterized in terms of the strong distinction between authority-type and hub-type pages [4]; authority-type pages are considered to have substantially more outgoing links than incoming links while hub-type pages have a—more-or-less—equal number of incoming and outgoing links. The analogy between authorities and subjects, and hubs and people is easily made. While websites can be related to two types of pages, weblogs are considered to "identify" a person — who can have many different interests (subjects) — and can thus only be related in an intuitive way with the hub-type pages of Kleinberg's HITS algorithm. Kumar et al. [5] present a topology-based algorithm for community extraction which they later use in so called Burst-Analysis. This algorithm is our baseline.

Lin et al. [7] focus on extracting communities based on two key insights: (a) communities form due to individual blogger actions that are mutually observable; (b) the semantics of the hyperlink structure are different from traditional web analysis problems. Their topology-based approach involves developing computational models for mutual awareness that incorporate the specific action type, frequency and time of occurrence.

Merelo-Guervos et al. [8] map a weblog hosting site using Kohonen's self-organizing map and discover interesting community features; they provide a comparison between their methods and other community-discovering algorithms. Like us, they use a mixture of topology- and content-analysis.

3. The main algorithm

In this section we introduce our algorithm for community discovery. It builds on three core ingredients:

- *content analysis*: blogs in the same community discuss related issues;
- *co-citation*: blogs in the same community link to similar resources; and
- *reciprocity*: blogs in the same community link to each other.

Below we describe the algorithms in a fair amount of detail; an experimental evaluation is provided in the next section.

3.1 Community discovery

For presentation purposes, our community discovery algorithm is split into two separate parts:

- Algorithm 1 shows the initialisation and discovery phases of the algorithm. Basically, Algorithm 1 is a simple iterative process which evaluates all weblogs pushed onto an intermediate "*discover stack*."
- Algorithm 2 is called on line 15 of Algorithm 1. On line 14 of Algorithm 1 we retrieve all relevant documents from the index by querying it with all terms occuring in the bodies of the posts present on the weblog currently being evaluated (`current`).

Algorithm 1 Extract community surrounding a certain starting point

Require: StartingPoint ← One weblog.
Ensure: All weblogs are in the index and links can be accessed.

```
 1: discoverStack.push StartingPoint
 2: resultSet ← ∅
 3: threshold ← ϵ
 4: thresholdCalculated ← false
 5: while discoverStack not Empty do
 6:    current ← discoverStack.pop
 7:    currentLinks ← current.linkSet
 8:    resultSet.add current
 9:    if current equals distance seperator then
10:       distance ← distance + 1
11:       discoverStack.push distance seperator
12:       next iteration
13:    end if
14:    relatedDocs ← Query(terms from current)
15:    Expand(current)
16: end while
```

After all related weblogs have been selected by Algorithm 1, they are passed to Algorithm 2 and a link strength is calculated for each related weblog. If this link strength exceeds a certain threshold value (defined as the average of the link strengths from the very first expansion step on line 18 in Algorithm 2), it is considered to be a part of the community. The link strength is based on aspects discussed in subsection 3.2 below.

Algorithm 2 The expansion step of the algorithm, follow all links and decide if the reached weblogs belong to the community

```
 1: for all weblog links from current as l do
 2:    relatedWeblog ← l.target
 3:    relatedLinks ← relatedWeblog.linkSet
 4:    linkIntersection ← currentLinks ∩ relatedLinks
 5:    cocit ← linkIntersection.size / currentLinks.size
 6:    if relatedLinks contains link to current then
 7:       recip ← reciprocity bonus
 8:    else
 9:       recip ← 0.0
10:    end if
11:    relevance ← relatedDocs.score for relatedWeblog
12:    linkStrength ← w_relevance · relevance + w_reciprocity ·
       recip + w_cocitation · cocit
13:    correction ← Correction based on distance
14:    if linkStrength − correction > threshold then
15:       discoverStack.push relatedWeblog
16:    end if
17:    if thresholdCalculated = false then
18:       threshold ← Average linkStrength
19:       thresholdCalculated ← true
20:    end if
21: end for
```

3.2 Link strength

The strength of a link between two weblogs is based on three factors—relevance, co-citation and reciprocity—and corrected with a distance penalty.

The relevance score for a certain link is extracted from the index by the query on line 14 in Algorithm 1. This query extracts the top n documents (in the experimental evaluation in Section 4 below we chose 20,000) from the index and their related relevance score.

The co-citation score is a fraction of the number of common resources both weblogs link to and the number of resources the referencing weblog links to. Motivations for this scheme are given by Brin and Page [1].

By adding a bonus to the link strength based on mutual acquiantance (link from A to B and from B to A) this kind of relations are rewarded. The value of this bonus is not easy to determine. One could argue that the value of this reward should be a number determined by the number of outgoing, incoming and reciprocal links. In the experiment a fixed value of 0.5 is used based on the belief that reciprocal links represent a strong link between two weblogs (or people).

As we diverge farther from the starting point (by number of links followed) it becomes less and less obvious that the weblog is related to the starting point. Therefore, a penalty (ρ) is substracted from the link strength based on the number of steps (`distance`) needed to reach the current weblog from the starting point. This correction is calculated as follows:

$$\rho = \left(1 - \frac{1}{distance}\right) \cdot linkstrength \tag{1}$$

By using this correction, weblogs close to the original starting point receive a smaller correction as opposed to weblogs "farther away" from the starting point which receive a larger correction. The reasoning behind this is that weblogs that are farther away are probably less known by the starting point (weblog/weblogger) and therefore need to have a stronger link

with their referrer to "prove themselves worthy."

3.3 Threshold

When the final link strength between two weblogs has been calculated, a conclusion on membership of the weblogs to a community can be drawn (Algorithm 2, line 14). This conclusion is reached based on a minimum threshold, which is calculated using the average link strength of the first expansion step, thereby effectively copying linking behaviour of the weblogger (the starting point). In other words, the starting point's linking pattern and content-based similarity with its direct neighbours is used as a definition for the rest of the community:

$$\sigma = \left(\sum_{i=1}^{n} linkstrength(A, B_i) \right) \cdot \frac{1}{n}, \quad (2)$$

where n is the number of direct neighbors of the starting point A.

3.4 Weights

When combining the three ingredients—relevance, co-citation and reciprocity—on line 12 in Algorithm 2, weights are needed for each of them. In order to find "optimal" settings for these weights an exhaustive parameter search is performed in our experimental evaluation in Section 4 below.

4. Experimental evaluation

In this section we provide details of an experimental evaluation of the algorithm introduced in Section 3. We start by describing our data set and data preparation efforts, our baseline, and our metrics. The results of the experiments are presented in Section 5.

4.1 Research questions

Our experimental evaluation is meant to address the following questions: Does incorporating a content-based approach into the exctraction of weblog communities help as compared against a topology-based approaches? And: What is the contribution of the various content- and topology-based components of our *Main algorithm*?

4.2 Data set

In order to test our algorithm a sufficiently large dataset was needed. The largest Dutch weblog service `web-log.nl` was kind enough to provide us with a data set. In total, `http://web-log.nl` hosts 317,104 weblogs with 6,450,291 posts. The suppplied data set contained all weblog posts from a period of one month (January 2006). During this month, 34,122 weblogs were updated with at least one post. In total, 367,129 posts were present in the data set making the average number of posts during this period 10.75 posts per weblog.

4.3 Data preparation

All data used in the experiment was first prepared for indexing. All links (URI's) were extracted from the bodies of the blog posts and saved in plain text (YAML) files. After link extracting, the log posts were cleaned (removing all links, html markup etc.) and indexed using Lucene [3].

4.4 Test topics

We randomly selected 22 weblogs, and then explored all weblogs linked to by these initial startings points and decided—per weblog—whether the potentially related weblog should be included in the community of the starting point or not. This process was repeated with the list of newly added weblogs until either no new weblogs were added to this list or we found that the subject of the weblog diverged to far from the original subject. This resulted in 22 lists of weblogs, each spanning a community surrounding one single weblog; on average 14 other weblogs were included in these communities.

Due to limited evaluation resources we were not able to have a second assessor examine these results.

4.5 Metrics

To measure the quality of the output of community finding algorithms, we simply adopt set intersection with the ground truth described above; this allows us to compute precision and recall values, as well the F_1-measure.

4.6 Baseline

As our baseline, we use Kumar et al. [5]'s topology-based algorithm for community extraction. This algorithm works in two steps: pruning and expansion. The pruning stage determines a number of seeds for the community from the complete graph spanned by all weblogs. These seeds are so-called K_3 cliques. A K_3 clique is defined by a weblog of which two links point to weblogs that both point to the other two members of the potential seed. If a node does not belong to a K_3 clique it is removed (pruned) from the graph toghether with all associated edges. After all weblogs are evaluated in this way the pruning step is repeated (Kumar et al. suggest three iterations).

After the pruning stage has been completed all seeds are passed on to the expansion stage of the algorithm in which the seeds are grown into community signatures. For each weblog in a seed (now a potential community), all links are followed to the potential new community members. A new member is added to the community if the number of links pointing back to the community exceeds a certain threshold t_k.

For the purposes of the experiments in this paper, only the seed containing the starting point used during the assessment of communities is considered and passed on to the expansion stage.

5. Results

After running the experiments for the content based algorithm presented in this article, 66 communities were generated for each topic present in the test set (one for each setting of the feature weights); these were compared against the ground truth described in Subsection 4.4.

Method	F_1
Baseline	0.5293
Relevance	0.7791
Co-citation	0.7427
Reciprocity	0.8549
Main algorithm	0.8660

Table 1: *Results of the experimental evaluation*

We compared the baseline community finding method against four other methods: one based on only on relevance, one based only on co-citation, one based only on reciprocity, and, finally, one based on all three (the *Main algorithm*). In Table 1 we list the average F_1-scores (averaged over all starting

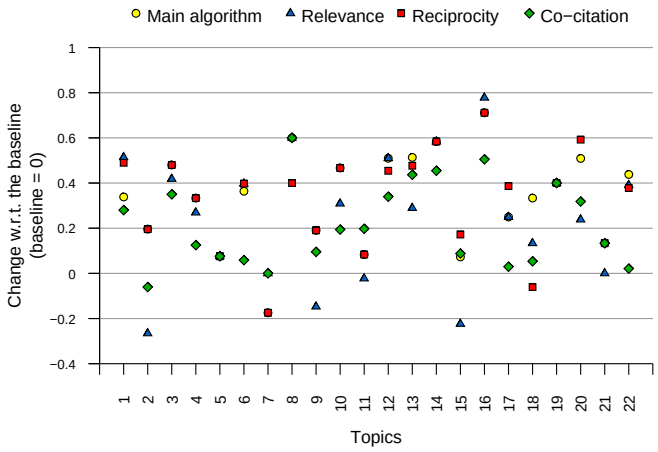

Fig. 1: *Difference in F_1 score per topic as compared against the baseline*

points) per community finding method. A few observations can be made. First of all, *Reciprocity* alone yields very high results. Yet, the best results are archieved when giving reciprocity a lower weight than the other two factors (0.5 for relevance, 0.3 for co-citation and 0.2 for reciprocity).

We turn to significance testing. In Table 2 we provide pairwise comparisons between all 5 community finding methods considered in this paper; we use the Wilcoxon Matched-Pairs Signed-Ranks test to test for significant differences. All methods (other than *Baseline*) significantly outperform the baseline. Also, as could be expected from the numbers listed in Table 1, there is no significant difference between *Reprocity* and the *Main algorithm*, and between *Relevance* and *Co-citation*.

	Relevance	Co-citation	Reciprocity	Main algorithm
Baseline	0.0012*	0.0001*	0.0001*	0.0001*
Relevance		0.4591	0.0023*	0.0250*
Co-citation			0.0074*	0.0023*
Reciprocity				1

Table 2: *Comparison of community finding methods; * denotes a significant difference in F_1-scores ($p < .05$)*

Next, we analyze the results by topic. In Figure 1 we plot a per topic comparison, where we display the differences (per topic) of the various methods as compared against the baseline. Given the results listed in Table 2, the picture is "as expected:" on most topics the *Baseline* loses out to the other methods, while *Reciprocity* and *Main algorithm* outperform the other methods on most topics.

6. Conclusions

We have presented a novel method for inferring weblog communities. The main research question we aimed to answer was whether incorporating content analysis into the extraction of weblog communities could improve over purely topological techniques was answered positively, even though one particular topology-based approach (*Reciprocity*) was not signficantly different from the combined content- and topology-based approach. We also found that each of the three core components of our main algorithm (relevance, co-citation, and reciprocity) contributed towards its overall effectivess.

Extensions to the algorithm presented here suggestion themselves. For a start, various aspects of the data in the data set were not exploited. E.g., comments on weblog posts were present in the data set but were ignored. Also, temporal information from the data set was not incorporated into the indexing and content-analysis proces. If this would have been taken into account, more recent postings could have been considered "more important" as they more accurately reflect a weblogger's current interests. The idea can be implemented through Lucene's boost function or using time-based language models [6].

Finally, more comparisons with communities assessed by additional assessors should be performed so as to be able to draw more solid conclusions. Also, the algorithm presented in this article should be compared to more community discovery algorithms than the algorithm described by Kumar et al. [5].

Acknowledgments

We would like to thank Ilse Media B.V. for providing us with the dataset; without the dataset this research would not have been posssible. Maarten de Rijke was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.065.120, 612-13-001, 612.000.106, 612.066.-302, 612.069.006, 640.001.501, 640.002.501, and by the E.U. IST programme of the 6th FP for RTD under project Multi-MATCH contract IST-033104.

References

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998.

[2] S. Chakrabarti. *Mining the Web*. Morgan Kaufmann, 2002.

[3] E. Hatcher and O. Gospodneti'c. *Lucene in Action*. Manning Publications, 2004. ISBN 1932394281.

[4] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *SODA '98: Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pages 668–677. Society for Industrial and Applied Mathematics, 1998.

[5] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. *World Wide Web*, 8 (2):159–178, June 2005.

[6] X. Li and W. B. Croft. Time-based language models. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 469–475, New York, NY, USA, 2003. ACM Press.

[7] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. Tseng. Discovery of blog communities based on mutual awareness. In *Proceedings WWE 2006*, 2006.

[8] J.-J. Merelo-Guervos, B. Prieto, F. Rateb, and F. Tricas. Mapping weblog communities, 2003. CoRR cs.NE/0312047.