Important Weblog Identification and Hot Story Summarization

Yi Wu

Intel Corporation
2200 Mission College Blvd
Santa Clara, CA 95054
yi.y.wu@intel.com

Belle L. Tseng

NEC Laboratories America
10080 N. Wolfe Road, SW3-350
Cupertino, CA 95014 USA
belle@sv.nec-labs.com

Abstract

In this paper, we propose the architecture for a weblog data mining system. Our objective is to allow users to interactively understand the blogspace by providing a system framework for retrieving relevant weblogs and obtaining highlighted information. We focus on two important technical components in the system. The first is weblog ranking. We introduce weighted link-based weblog ranking, which ranks the popularity of weblogs according to their entry semantic content and time delay of citation. Furthermore, our weblog ranking algorithms provide the flexibility to rank weblogs not only based on their different roles in the society, but also based on end-users' different ranking interests. The second component is hot story summarization. A hot story is the discussion that attracts various weblogs' attention. Influential bloggers are useful in identifying hot conversations because these bloggers are likely to be the leader in such conversations. We propose a method based on first discovering weblogs that take important roles in the society, and then extracting hot story from these important weblogs.

1. Introduction

Recently, weblogs (or blogs) have become prominent social media on the Internet that enable users to quickly and easily publish content including highly personal thoughts. A blog is typically a web site that consists of dated entries in reverse chronological order written and maintained by a user (blogger) using a specialized tool. A blog entry can have hyperlinks to web pages or other blog entries, resulting in multiple hyperlinks between different blogs. The information structure of blogs and links is sometimes referred to as the blogspace. Figure 1 shows a typical link structure in blogspace.

Blogs have created a fast growing social network on the Internet. The blogspace can be exploited for identifying opinion formation, maintaining online communities, supporting knowledge management within large global collaborative environments, monitoring reactions to public events and is seen as the upcoming alternative to the mass media. Compared to typical social networks in the real world, blogspace has the following special characteristics.

- The link structure in blogspace implies the influence of bloggers. One can become much more easily known to others. For instance, a blog that provides interesting entries can attract audiences and get citation links from others. In this manner, a "celebrity" within an interest community can emerge dynamically. On the other hand, since anyone can publish and link to any blog, there are many unimportant or even harmful entries such as spam.
- The blogspace is a dynamic social network. Different from regular webpages, which are mostly stable in content, the content in blogspace is updated frequently, reflecting the new trend in a society. As a result, the popularity and quality of blogs change dynamically.
- There are many variations on the weblog content. Some of the weblogs might only discuss focused topic. Some of the weblogs might be interested in diverse content. The popularity and quality of weblogs change according to their semantic content.

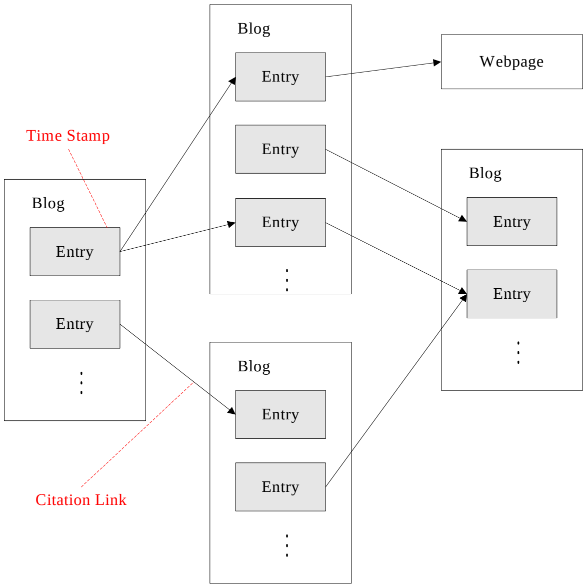


Figure 1: Typical Link Structure in Blogspace

In this paper, our focus is the architecture of weblog data mining. More specifically, given a specific query topic and a certain time frame, we want to (1) search related blogs/entries, (2) identify good information sources in a community (i.e., who are good at making summary? who are influential bloggers whose voice is echoed by others?), and (3) summarize and highlight information (i.e., what are the hot stories? what are the major topic categories? what are the hot phrases in each category?).

We first propose the architecture for weblog mining. Our system allows the user to specify a query and retrieve highlighted weblog information of the topic, through which the user can explore the details of weblog content.

We then focus on two important components in our weblog mining system. One is to identify important weblogs and the other is to summarize hot stories. We introduce weighted link-based weblog ranking in order to overcome the weakness of traditional linked-based ranking approaches (Details will be discussed in Section 4). Our weblog ranking algorithms can rank weblogs not only based on their different roles in the society, but also based on end-users' different ranking criteria. To analyze the important conversation in the blogspace, we introduce hot story summarization by incorporating popularity (given by ranking score) and content association (relevant to a query topic).

The paper is organized as follows. Section 2 reviews related work on blogspace. In Section 3, we propose our system architecture for weblog data mining. Section 4 proposes weblog ranking algorithms. Section 5 proposes hot story summarization to help a user understand the information dissemination of blogs on a specified topic. Data collection for our system setup and preliminary results on the data are shown in Section 6. Section 7 summarizes our conclusion and future works.

2. Related Work

Various weblog search engines have been developed to maintain lists of the most popular bloggers, such as Daypop[1], blogdex[2], Technorati[3], BlogStreet[4], BlogPulse[5] and etc. Technorati allows a blogger to identify his/her blog within a neighborhood of other bloggers. Blogpulse publishes daily lists of key persons, key phrases, and key paragraphs to a public web site. Beyond a search index, Blogpulse implemented trend search, which graphs the normalized trend line over time for a search query and provides a way to estimate the relative buzz by word of mouth for some given topics over time. However, most of these search engines are based on explicit citation counts to rank weblogs. But the quality of citation and the relevance of citation to the query topic are not taken into consideration.

The limited quantitative research on blogs has primarily focused on information diffusion in the blogspace. Kumar et al. studied the "burstiness" of blogspace [6]. They extracted blog communities and investigated the bursts of activity with the communities based on analyzing the evolving link structure. Gruhl et al. studied the diffusion of information through Blogspace [7]. They tried to characterize and model information diffusion in two levels. One is macroscopic characterization of topic propagation with topics generated by outside events; the other is microscopic characterization of propagation from individual to individual. Adar et al. studied the implicit structure and dynamics of blogspace [8]. They examined the explicit link structure and the implicit link structure between blogs for finding blogs that are sources of information diffusion. However, their purpose was not to acquire important weblog content.

Recently, researchers start to look into the community structure in blogspace. Tseng et al. proposed tomographic clustering to explore multiple communities of interests in blogspace and mountain-view visualization to provide a landscape of blog communities in terms of popularity and connectivity [9].

To capture hot conversational threads from blogs, Nakajima et al. proposed a method of discovering bloggers who take an important role in conversations [10]. However, they only studied the weblog hyperlink structure, without analyzing the semantic content. As a result, they couldn't discover whether the detected story satisfies users' interest and whether topic drift happened in the conversational threads.

In our paper, we focus on providing the architecture of a blog mining system. Our objective is to allow users to interactively understand the blogspace by providing a system framework for retrieving relevant blogs and obtaining highlighted information.

3. System Overview

We propose a system framework that will allow a user to specify a query, retrieve relevant blogs, and extract representative information. Figure 2 illustrates the blog mining system architecture.

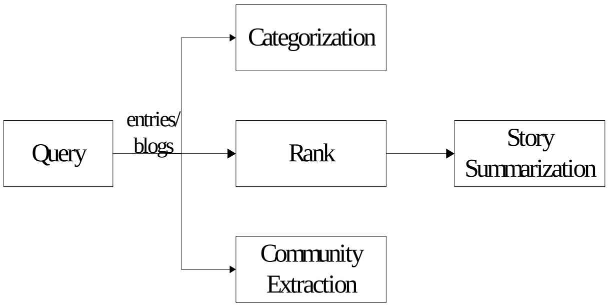


Figure 2: System Framework Block Diagram

The first module is *blog query*, which allows the user to specify a topic of interest and returns semantically relevant entries from the blogspace. From these relevant entries, we can further retrieve relevant blogs who are the owners of these entries. The *categorization* module is to extract the popular categories and hot phrases as discovered by the relevant entries, which has been presented in [14]. The we*blog rank* module ranks the relevant entries or weblogs for the query topic, as will be described in Section 4. The *story summarization* module is to extract hot discussion threads of the query topic, and will be described in Section 5. Finally, the *community extraction* module performs clustering to discover communities of connected and relevant blogs corresponding to a user query. A mountain view visualization is provided to explore different communities of interest in blogspace. The mountain views are generated using a tomographic clustering algorithm on the blog social network. The mountain view shows mountains of communities consisting of connected blogs. Peaks and valleys of the mountain view depict representative blogs as community authorities and community connectors, respectively. Details of community extraction have been introduced in [9]. In the following sections, we will emphasize on two main modules in the system: we*blog rank* and *story summarization*.

4. Weblog Rank

To locate valuable web contents on the Internet, search engines are used to retrieve a sequential rank ordering of important pages with respect to the user's request. Our goal in this section is to identify important weblogs in the blogspace (i.e., blogs taking important roles in a community and trusted by their community members).

4.1 Link-based Webpage Ranking

In the literature, link analysis algorithms have shown their successes in measuring the importance of webpages. Among them, PageRank [11] and HITS [12] are two widely recognized ranking techniques. Representing the structure, the entire web can be modeled as a directed graph:

$$G =< V,E >,$$

where V = {1, 2, · · · , n} is the collection of nodes, each of which represents a page; and E = {< i, j > |i, j ∈V } is the collection of edges, each of which represents a hyperlink. For example, < i, j > means a hyperlink from page i to page j.

The PageRank algorithm assigns a numeric property, called PageRank, to each page to represent its importance. This algorithm simulates a random walk process in the web graph. Suppose there is a surfer on an arbitrary page of the Web. At each step, he/she will jump to one of the destination pages via the hyperlinks on the current page with probability 1-d, or to another page in the whole graph with probability d. This process can also be formulated in an iterative manner. Let PR = (PR(1), PR(2), · · · , PR(n))^T denote the PageRank scores of the web graph. The iteration process can be formulated as follows:

$$PR_{k+1}(i) = (1-d) + d\sum_{j\in IN_i}\frac{PR_k(j)}{|OUT_j|} \qquad (1)$$

where k represents the iteration step, d is a constant set as 0.85 in Pagerank algorithm, IN_i is a set of pages citing page i, and $|OUT_j|$ is the number of outgoing link from page j.

Pagerank has been used in the Google search engine and shown promising performance. However, Pagerank only generates one score for each page in terms of the authority influence. For our blog mining system, our objective is to identify not only influential bloggers whose voice is echoed by others in a community, but also influential bloggers who are good at making summaries.

Different from PageRank, HITS assigns two scores to each page, called authority score and hub score. Hubs and authorities exhibit a mutually reinforcing relationship. If a page links to many pages with high authority score, it will obtain a high hub score. On the other hand, if a page is linked by many pages with high hub score, it will obtain a high authority score symmetrically. We can obtain two scores for each page in an iterative manner.

Let A = (A(1), A(2), · · · , A(n))^T and H = (H(1), H(2), · · · , H(n))^T denote the authority and the hub scores of the web graph respectively. Without regard to normalization, the iteration process can be formulated as follows:

$$A_{k+1}(i) = \sum_{j\in IN_i} H_k(j) \qquad (2)$$

$$H_{k+1}(i) = \sum_{j\in OUT_i} A_k(j) \qquad (3)$$

Link analysis has shown great potential in improving the performance of web search. All these link analysis algorithms are based on two assumptions. First, each webpage is treated as a single node in the web graph, and the link relationship is studied on the page level. Second, the links are treated as identically important. If there exists a link from page A to page B, then the author of the first page A finds the second page B valuable. Thus the importance of a page can be propagated to those pages it links to. Furthermore, the pages that are co-cited by a certain page are likely treated with the same importance by that page.

However, for weblog analysis, those aforementioned link-based ranking algorithms cannot be employed directly. The reasons are as follow. First, a weblog page contains multiple entries with different semantics and hence the weblog page might not be considered as the atomic node. Second, the hyperlink does not have the same meaning, but organized as multiple semantics. In this regard, hyperlinks should also be treated non-identically.

4.2 Weighted Link-based Weblog ranking

In this section, we will introduce weighted link-based weblog ranking in order to overcome the weakness of traditional linked-based ranking approaches. The novelty of our weighted link-based weblog ranking lies in two parts. First, we take an entry instead of a weblog page as an atomic node, because a weblog might cover varied topics, but an entry mostly contains focused content. We then derive the weblog properties by accumulating the effects at entry level. As a result, the popularity and quality of weblogs change according to their entry semantics. Second, we apply different weights on each hyper-link between entries. The weight considers two factors: content semantic similarity between entries, and time delay of citation.

4.2.1 Entry Ranking

In the blogspace, we treat each entry as an atomic node and study the entry-to-entry link relationships. Let entry graph EG be denoted as EG = <E, EL>, where E = $\{e_i\}$ is the set of entries and EL = $\{(e_i, e_j)$ is set of entry links. For example, (e_i, e_j) means entry e_i cites entry e_j.

Given an entry graph EG, the link graph EG^L of the entry graph EG is defined as follows:

$$EG^L = \{ \begin{matrix} w(i,j) & if\ (e_i, e_j) \in EL \\ 0 & otherwise \end{matrix} \qquad (4)$$

where *w(i,j)* represents the weight of link from entry e_i to e_i.

We employ HITS as the basic ranking algorithm to rank entries because we are interested in two scores for each blog, one is to represent the importance of a blog in terms of its ability of making good summary (hub score) and the other is to represent the importance of a blog in terms of its influence (authority score). We modify the original HITS ranking to deal with the weight of each link. The iterative process of the weighted HITS ranking can thus be formulated as follows:

$$A_{k+1}(i) = \sum_{j \in IN_i} H_k(j) \times w(j,i) \qquad (5)$$

$$H_{k+1}(i) = \sum_{j \in OUT_i} A_k(j) \times w(i,j) \qquad (6)$$

To define the link weight of *w(i,j),* we aim to combine three components: *content, time* and *link* together. We study the content and time property of each link in order to deal with special characteristics of the blogspace: dynamics and divergence of popularity and quality of blogs.

Figure 3 shows an example of four entries in terms of the query “google”. Entry b, entry c and entry d point to entry a. Traditionally, the weights of each link are assigned as equal. However, if we look at the property of citation link, entry b and entry a talk about more similar topic regarding search engine, and the citation happens promptly after the entry a got published; entry c is talking about google stock price; the citation from entry d happens several days later. Intuitively, we would like to assign higher endorsement for the link from b to a, because the citation comes from related content and the response is very quick.

We apply different weights on each citation link between entries. The weight considers two factors: content semantic similarity between entries, and time delay of citation. The more similar content of two entries and the more recent citation, the weight of the link is larger. The weight between entry e_i and e_j is thus defined as follows:

$$w(i,j) = sem(i,j) * exp(-k\Delta t(i,j)) \qquad (7)$$

where *sem(i,j)* is the semantic similarity between e_i and e_j, $\Delta t(i,j)$ is the time difference between e_i and e_j and *k* is the constant of a decay function. The semantic similarity *sem(i,j)* is calculated as the cosine similarity of term vectors. These term vectors are weighted using the standard TFIDF scheme [13]. The similarity value ranges from 0 to 1. As a result of this step, we obtain two scores for each entry, one is the hub score and the other is authority score.

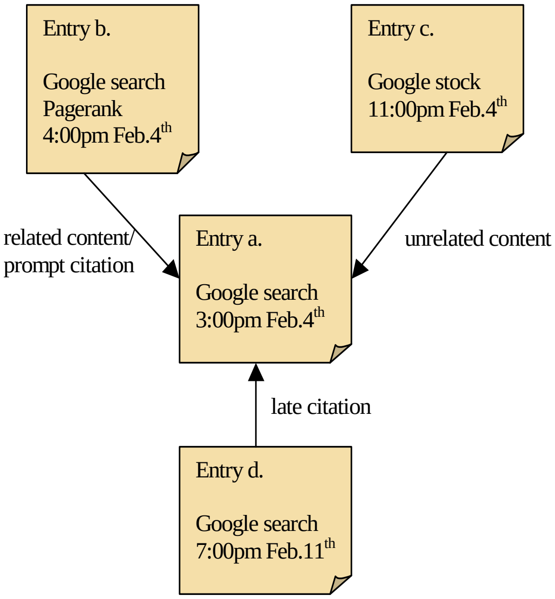


Figure 3: Unequal Importance of Links

4.2.2 Weblog Ranking

Once we calculate the entry ranking scores, weblog ranking is based on entry ranking. We consider a set of blogs B = $\{b_i\}$. Each blog b_i owns a set of entries E_i. There could be multiple entry-to-entry links from blog b_i to blog b_j, which we denote as entry link

$$EL_{ij} = \{(e_k, e_l) \mid e_k \in E_i, e_l \in E_j, (e_k, e_l) \in EL\} \qquad (8)$$

Those entry-to-entry links generate the social network of weblogs B. For each entry, we have hub score and authority score. The same for each blog, we also can obtain two scores, hub score and authority score, representing two *roles* of that blog in the society (blogs who are good at making summary have high hub scores; blogs whose voice is echoed by others have high authority scores).

Furthermore, users might be interested in the different ranking lists in terms of different ranking criteria. To provide the flexibility, we proposed multiple weblog ranking algorithms not only based on their different roles in the society, but also based on end-users’ different ranking criteria.

(1) Ranking Based on Average Entry Quality

The average quality of entries that a weblog owns represents the quality of the weblog. This ranking algorithm calculates the authority score ($A(b_i)$) and hub score ($H(b_i)$) of a weblog b_i as the average scores (authority or hub) of entries belonging to b_i as shown in the below.

$$A(b_i) = \frac{1}{|E_i|} \sum_{e_j \in E_i} A(e_j) \qquad (9)$$

$$H(b_i) = \frac{1}{|E_i|} \sum_{e_j \in E_i} H(e_j) \qquad (10)$$

where $|E_i|$ is the total number of entries blog b_i owns, $A(e_j)$ is the hub score of the entry e_j in blog b_i, and $H(e_j)$ is the hub score of the entry e_j in blog b_i. Because there exists a high variance in the total number of entries that a blog can own, we take the normalization.

(2) Ranking Based on Citation Frequency

The number and the quality of incoming citations represent the respectful influence of a weblog in the community (the role as an authority). The number and the quality of outgoing citations represent the capability of a weblog to identify influential weblogs (the role as a hub).

This ranking algorithm calculates the authority score of a weblog b_i based on incoming citation number and the hub score of entries citing b_i as below.

$$A(b_i) = \sum_{e_j \in IN(b_i)} H(e_j). \qquad (11)$$

where $IN(b_i)$ is the set of entries citing blog b_i and $H(e_j)$ is the hub score of the entry e_j which has cited blog b_i. The more frequently a blog is cited by entries with high hub scores, the more respectful influence the blog tends to have in the community of interest.

The hub score of a weblog b_i is calculated based on outgoing citation number and the authority score of entries being cited by b_i as below.

$$H(b_i) = \sum_{e_j \in OUT(b_i)} A(e_j) \qquad (12)$$

where $OUT(b_i)$ is the set of entries blog b_i cites and $A(e_j)$ is the hub score of the entry e_j which has been cited by blog b_i. The more frequently a blog cites entries with high authority scores, the more capability that the blog tends to be a good summarizer in the community of interest.

(3) Ranking Based on Citation Diversity

Instead of accumulating the effect of each entry reference between two blogs, this ranking algorithm takes the average. The authority score of a blog is calculated as below.

$$A(b_i) = \sum_{b_j \in IN(b_i)} \left\{ \frac{1}{|EL_{ij}|} \sum\nolimits_{(e_k, e_l) \in EL_{ij}} H(e_k) \right\} (13)$$

where $IN(b_i)$ is a set of of blogs citing blog b_i, EL_{ij} is the set of entry links from blog b_j to blog b_i. The more frequently a blog is cited by entries from multiple blogs, the more respectful influence the blog tends to have in the community of interest.

The underlying reason of this ranking algorithm is that as we have observed, there was a high variance in the total number of links between blog b_i and blog b_j, variance of $|EL_{ij}|$ is large. Blog b_i might have a large number of incoming links, and they are all from the same blog b_j. The reason could be biased opinions or blog spamming. Ranking based on citation diversity can alleviate the side effects by favoring the situation that one blog gets citations from diverse blogs. For example, if blog b_1 gets 10 citations from five different blogs and blog b_2 gets 10 citations from one blog (denoted as b_3), we think b_1 might be more influential, because very likely, b_2 and b_3 are friends or even the same person.

Following the same idea, the hub score of a blog is calculated as below.

$$H(b_i) = \sum_{b_j \in OUT(b_i)} \left\{ \frac{1}{|EL_{ij}|} \sum\nolimits_{(e_k, e_l) \in EL_{ij}} A(e_l) \right\} (14)$$

5. Story Summarization

Given a set of entries related to a topic, we are interested in extracting important story threads, and see how the story is initiated, how it developed, and how it ended. Our objective is to provide summarized hot stories to the users.

A hot story is the discussion that attracts various weblogs' attention. Influential bloggers are useful in identifying hot conversations because these bloggers are likely to be the leader in the conversations. To capture potentially hot stories of a certain topic on the weblog, we propose a method based on first discovering entries that take important roles in this topic, and then extracting story threads from these important entries.

Section 4.2.1 defines the importance of an entry in terms of authority and hub scores. An entry is important on the basis of his or her role in the blogspace, i.e., a set of blog entries get highly recommended by other entries (authority), or a set of blog entries providing good summaries of the discussion for a specific topic (hub).

Let E_{ta} be the top-authority ranking set of entries with threshold t (i.e., $E_{ta} = \{e|e \in E, A(e) >= t\}$). Let E_{th} be the top-hub ranking set of entries with threshold t (i.e., E_{th}

={e|e∈E, H(e)>=t}). Once we have identified the important entries E_{ta} and E_{th}, we can extract hot stories based on the following definition.

Definition. *A hot story is defined in two ways: (1) A hot story is cited or ended by entries with high hub scores, because a good hub tends to be a good story summarizer, and (2) A hot story is initiated by entries with high authority scores, because a good authority tends to be a good influential leader of a story.*

Figure 4 shows an example of hot story ending with a "hub" entry. In the figure, each node represents an entry and the one labeled with "H" is the "hub" entry. We start from this "hub" entry and trace back all the entries it pointed to, which constructs a hot story summarized by the hub entry.

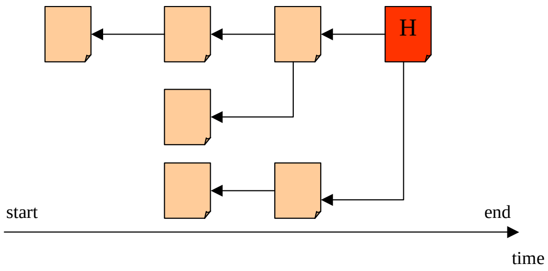


Figure 4: Hot Story Ending with a "Hub" Entry

Figure 5 shows an example of hot story initiated by an "authority" entry. In the figure, each node represents an entry and the one labeled with "A" is the "authority" entry. We start from this "authority" entry and track all the entries pointing to it, which constructs a hot story initiated by the authority entry.

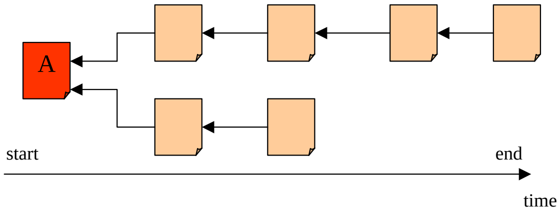


Figure 5: Hot Story Initiated by an "Authority" Entry

We also define story scores for hot story ranking. Assume that each story contains N entries, then the story score is defined as the product of the score of story root (a hub or an authority) and the average relevance of entries to the query, as shown in Equation 15.

$$S(story) = S(e_{root}) \times \frac{\sum_{i=1}^{N} relevance(e_i, query)}{N} \quad (15)$$

where relevance(e_i, query) is the relevance score of entry e_i to the query, which is calculated using TFIDF [13], and S(e_{root}) is the score of a story root entry (a hub entry or an authority entry).

6. Our System

6.1 Data Corpus Collection

For our weblog mining system, we have developed a focused crawler to collect blog data from the Internet. The crawling process is divided into four components, (1) initial seeds, (2) blog discovery, (3) entry extraction, and (4) seed expansion.

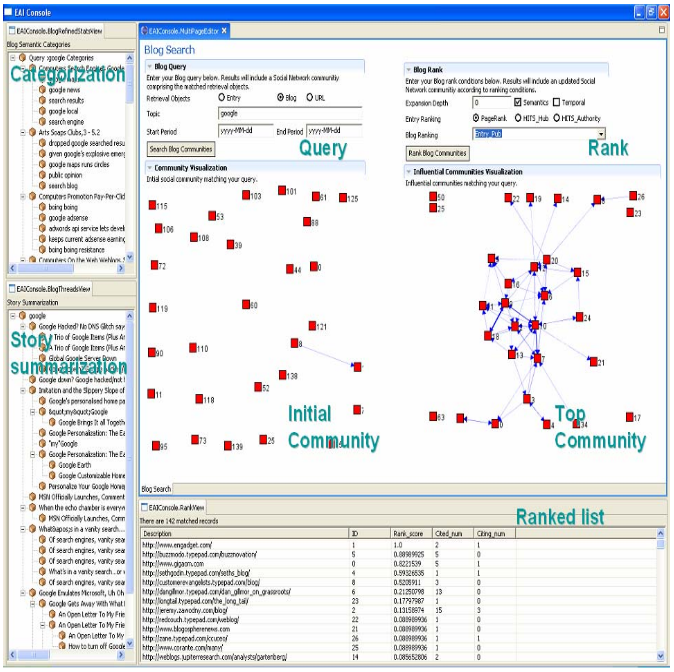
We started from 100 blogs that are most famous for technical discussion. From the initial set of seeds, the crawler retrieves RSS files of the seed blogs and pages referred to by the RSS files. Next for blog discovery, when a crawled page has an HTML link tag that refers to an RSS (this common feature of recent blog tools is called "RSS auto discovery"), the crawler checks whether this RSS represents a blog. If an RSS file satisfies the following conditions, the web site referred to by the RSS as "channel" is recognized as a blog if: (1) The RSS contains items referring to pages in the same host and (2) Each page referred to by the RSS has an HTML link tag that refers back to the RSS.

In entry extraction, the crawler needs to extract entry data from web pages since RSS file does not always contain the entire entry content. The crawler extracts the content of an entry from the corresponding web page using an extraction pattern described with XPath expressions.

Subsequently, seed expansion is incorporated to capture possibly relevant contents. From entry pages of the seed blogs, the crawler crawls hyperlinks for N hops (currently N is set to 1), and discovers blogs. From the collected blogs, highly reference blogs in the set are chosen for new seeds. From June 2005 to September 2005, we have collected 40,284 weblogs and 192,391 entries. Of course, the size of blogosphere on the Internet is growing every day. But we believe that our collected weblogs are representative enough for our experimental purpose.

6.2 System Interface

For a query, users are interested in finding the top blogs that cover their topic of interest. We built our blog mining system interface to allow users to input their queries and retrieve important representative weblogs and hot story summaries. For the input query, users can enter (1) keywords, (2) the period of retrieval, and (3) a selection for either entry or blog ranking outputs. Figure 6 illustrates the blog search interface where we chose the keyword "google", the period spanning from July 1 to Aug 2, 2005 and query object as blogs. Subsequently, the relevant blogs can be ranked according to different blog ranking options as described earlier. The generated list of blogs in ranking order with their corresponding IDs and impact scores are depicted on the bottom view of Figure 6. The blog social networks corresponding to the top ranked blog scores are shown as "top community". The major categories and hot phrases in each category are also displayed in the top-left view of Figure 6. Hot stories are displayed in the bottom-left view of Figure 6, where they were rank ordered according to their story scores.



EAI Console

EAIConsole.MultPageEditor

Blog Search

Blog Query

Enter your Blog query below. Results will include a Social Network communitiy comprising the matched retrieval objects.

Retrieval Objects ○ Entry ● Blog ○ URL

Topic: google

Start Period: yyyy-MM-dd End Period: yyyy-MM-dd

Search Blog Communities

Community Visualization

Intial social community matching your query.

Blog Rank

Enter your Blog rank conditions below. Results will include an updated Social Network communitiy according to ranking conditions.

Expansion Depth: 0 ☑ Semantics □ Temporal

Entry Ranking ● PageRank ○ HITS_Hub ○ HITS_Authority

Blog Ranking: Entry_Pub

Rank Blog Communities

Influential Communities Visualization

Influential communities matching your query.

Blog Search

EAIConsole.RankView

There are 142 matched records

Description	ID	Rank_score	Cited_num	Citing_num
http://www.engadget.com/	1	1.0	2	1
http://buzzmodo.typepad.com/buzznovation/	5	0.88989925	5	0
http://www.gigaom.com	0	0.8221539	5	1
http://sethgodin.typepad.com/seths_blog/	4	0.59326535	1	1
http://customerevangelists.typepad.com/blog/	8	0.5205911	3	0
http://dangillmor.typepad.com/dan_gillmor_on_grassroots/	6	0.21250798	13	0
http://longtail.typepad.com/the_long_tail/	23	0.17797987	1	0
http://jeremy.zawodny.com/blog/	2	0.13158974	15	3
http://redcouch.typepad.com/weblog/	22	0.088989936	1	0
http://www.blogospherenews.com	21	0.088989936	1	0
http://zane.typepad.com/ccuceo/	26	0.088989936	1	1
http://www.corante.com/many/	25	0.088989936	1	0
http://weblogs.jupiterresearch.com/analysts/gartenberg/	14	0.085652806	2	0

Categorization
Story summarization
Query
Rank
Initial Community
Top Community
Ranked list
EAIConsole.BlogRefinedStatsView
Blog Semantic Categories
Query :google Categories
Computers Search Engines Google
google news
search results
google local
search engine
Arts Soaps Clubs,3 - 5.2
dropped google searched resu
given google's explosive emerg
google maps runs circles
public opinion
search blog
Computers Promotion Pay-Per-Clic
boing boing
google adsense
adwords api service lets devel
keeps current adsense earning
boing boing resistance
Computers On the Web Weblogs
EAIConsole.BlogThreadsView
Story Summarization
google
Google Hacked? No DNS Glitch says
A Trio of Google Items (Plus Ar
A Trio of Google Items (Plus Ar
Global Google Server Down
Google down? Google hacked/not
Imitation and the Slippery Slope of
Google's personalised home pa
&quot;my&quot;Google
Google Brings It all Togeth
Google Personalization: The Ea
"my"Google
Google Personalization: The Ea
Google Earth
Google Customizable Home
Personalize Your Google Home
MSN Officially Launches, Comment
When the echo chamber is everyw
MSN Officially Launches, Comm
What&apos;s in a vanity search...
Of search engines, vanity sear
Of search engines, vanity sear
Of search engines, vanity sear
What's in a vanity search...or
Of search engines, vanity sear
Google Emulates Microsoft, Uh Oh
Google Gets Away With What I
An Open Letter To My Frie
An Open Letter To My Frie
An Open Letter To My
How to turn off Google
115
103
101
61
125
53
88
106
108
39
72
44
0
119
60
121
90
110
8
138
52
11
118
7
95
73
139
25
50
25
22
19
14
26
23
20
15
16
24
13
21
63
0
4
34
3
17

Figure 6: System Interface for User Query Input, Resulting Blog Rankings, Community Social Network, and Hot Story Summarization

7. Conclusions

Blogs provide an opportunity for people to share important information in a community. In our paper, we introduce our weblog mining system. Given a specific query topic and a certain time frame, our system will search related blogs/entries, identify good informational sources in the community, and summarize important information. This paper focuses on two important components in our weblog mining system. One is to identify important representative weblogs and the other is to summarize hot stories. We introduce weighted link-based weblog ranking, which ranks the popularity of weblogs according to their entry semantic content and time delay of citation. Our weblog ranking algorithms can rank weblogs based on weblogs' different roles in the society or based on end-users' different ranking criteria. To extract the important conversation in the blogspace, we introduced hot story summarization by incorporating popularity (given by ranking score) and content association (relevant to query topic).

In our future work, we will study the semantics of anchor text around a hyperlink to reflect more link properties. The existing link analysis methods treat all hyperlinks in the same sense as recommendation. But in reality, the links might convey criticism instead of endorsement. A weblog page can contain multiple outgoing citation links. Some of the links support the cited pages while some of the other links criticize the cited pages. By differentiating endorsement link and criticism links, we can obtain more accurate weblog ranking.

References

[1] http://www.daypop.com
[2] http://www.blogdex.com
[3] http://www.technorati.com
[4] http://www.blogstreet.com
[5] http://www.blogpulse.com
[6] R. Kumar, J. Novak, P. Raghavan, A. Tomkins. *On the Bursty Evolution of Blogspace*. The Twelfth International World Wide Web Conference (2003).
[7] D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins. *Information Diffusion Through Blogspace*. The Thirteenth International World Wide Web Conference (2004).
[8] E. Adar, L. Zhang: *Implicit Structure and Dynamics of Blogspace*. WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2004).
[9] B. L. Tseng, J. Tatemura and Y. Wu. *Tomographic Clustering To Visualize Blog Communities as Mountain Views.* WWW2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2005).
[10] S. Nakajima, J. Tatemura and Y. Hino. *Discovering Important Bloggers Based on a Blog Thread Analysis*. WWW2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2005).
[11]L. Page, S. Brin, R. Motwani and T. Winograd. *The pagerank citation ranking: Bringing order to the web.* Technical report, Stanford Digital Library Technologies Project, 1998.
[12] J. M. Kleinberg. *Authoritative sources in a hyperlinked environment*. Journal of ACM, 46(5):604-632, 1999.
[13] C. D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA, 1999.
[14] P. Appan, B. Tseng and H. Sundaram. *Weblog entry categorization based on query context* TR-no: AME-TR-2005-16, University of Arizona.