Detecting and Analyzing Blog Community

Tomoya TANIGUCHI
University of Tokyo,
Hongo 7-3-1,Bunkyo-ku,
Tokyo, JPN
tani@miv.t.u-tokyo.ac.jp

Yutaka MATSUO
National Inst. of AIST,
Aomi 2-41-6,Koto-ku,
Tokyo, JPN
y.matsuo@carc.aist.go.jp

Koiti HASIDA
National Inst. of AIST,
Aomi 2-41-6,Koto-ku,
Tokyo, JPN
hasida.k@aist.go.jp

Mitsuru ISHIZUKA
University of Tokyo,
Hongo 7-3-1,Bunkyo-ku
Tokyo, JPN
ishizuka@miv.t.u-tokyo.ac.jp

Abstract

Recently, many web-log (blog) hosting services have begun in Japan. Blog users have continued to multiply, producing new Internet communities. This study extracts and analyzes blog communities based on the blog link structure. Blog networks can be constructed as follows: a blog (all entries of one blog) is considered as a node. A link between two blogs is considered as an edge. We attempt two experiments to two areas: scoring blog pages by PageRank and clustering blog communities by betweenness clustering for the blog networks on “baseball” and “Winny”.

1 Introduction

In 1999, softwares packages that facilitate the construction of web-logs (blogs), such as “Blogger.com” and “Pita”, have gained in popularity. Since then, blog users have increased all over the world [Blood, 2002]. In Japan, the number of blog users has increased rapidly especially since 2003. Many blog-hosting services started their services in 2003: “Cocolog”, ”Livedoor Blog", “JUGEM”, “Excite Blog”, “Doblog”, etc.

Blogs are a novel medium for individuals to publish their opinions. Blogs have several characteristics: First, a blog usually allows a user to perform a trackback search. Trackbacks show connection of other blogs in both directions. Thereby, people can easily understand discussion among blogs. Secondly, a blog can send an RDF Site Summary (RSS). It is easy to aggregate such RSSs and show a user new and interesting changes. Thirdly, it is easy to write an entry (article) of a blog, just like writing a message to a message board.

One author usually writes a blog. We infer that a blog is “representative of one individual”. Cases exist where multiple authors write one blog, but this study ignores such cases. Therefore, relationships between blogs correspond to the relationships between people.

Links and trackbacks can be considered as a social/interaction relationship between blog authors. For example, some blogs contain discussion of baseball and some refer to political information. Baseball blogs often links and trackbacks to other articles related to baseball, and politics blogs usually have links and trackbacks to other political blogs. Therefore, we can identify the topical communities of blogs if we analyze the link and trackback structure.

Blogs usually have several semantic web technologies, such as RSS and RDF. Blogs might bring a promising breakthrough for the semantic web technologies to prevail: blog can be a killer application of semantic web. Analyzing blogs and extracting communities is important in semantic web as well as in current Web because there are always we can know only by accumulating information. Community can be extracted by information that is accumulated from small pieces of information.

So far, there are a lot of researches about detecting communities in web pages from link structures. For example, bipartite subgraphs are extracted as hubs and authorities. Link co-occurrence can also detect communities.

The aim of our research is to apply these link-based analyses to Blogs and show the applicability of this approach. First, we try to find authoritative blogs by PageRank algorithm [Page et al., 1997]. Second we apply graph-based clustering technique to the blog link structure and try to detect communities [Michelle et al., 2001].

In this research, the target blogs are those that are hosted by famous blog hosting services. (Top nine services that cover more than 80% Japanese blogs as of June 30, 2004.) We exclude Movable Type and other types of blogs due to implementation problem.

2 Extracting Link Structures

We collect links from blogs using a crawler that we have made. In this section, we show how to extract links from blogs.

2.1 Target of Extracting

One blog page has some parts, such as sides, comments, trackback, and entries (articles). In the entries, there are links to other blogs. We extract such links in entries. An author writes his opinion in entries with reference to other blogs. Therefore, if we collect the links to other blogs, we can have a link structure of the target blogs.

One problem when we employ simple enumeration of links to other blogs is: a blog that has many entries and many links will have a large outdegree in the blog graph.

Usually blogs have many outgoing links, so we pick up only top three frequent links to other blogs.

2.2 How to Collect Links

We collect Links by following ways: First, we define 10 blogs as *seed blogs*. Next we extract links from seed blogs, and we collect blogs that are linked from the seed blogs. We repeat these process five times.

Extracting links from a blog is rather complicated. Because RSS can show recent entries, but we do not want to limit the range in recent entries. So we crawl blog entries one by one. The following shows this procedure.

```
this_address ← Latest entry's address;
while( Dose this_address exist? ){
    Extract_next_address( Source(this_address) );
    this_address ← next_address;
}
```

where Source(x) denotes the source of x (x is URI), and Extract_next_address(y) returns a URI of previous entry in Source(y).

Obtained blogs depends heavily on seed blogs. We select two sets of blogs as seed blogs: baseball blogs and Winny blogs. Baseball blogs are collected by putting a query "blog" and "baseball" to a search engine and get ten qualified blogs. Similarly, Winny blogs are collected. Winny is a famous Japanese P2P software and bring social concerns because the developer (a university staff) is arrested for helping to make illegal copies of copyrighted files.

3 How to analyze Link Structures

3.1 Rating Blogs by PageRank

PageRank is a method for rating web pages: Assume the number of pages is n. Vector $\boldsymbol{R}(1\times n)$ whose elements(r_1–r_n) denotes PageRanks of these pages, and matrix $\boldsymbol{M}(n\times n)$ denotes a link information between these pages. Thus vectors that satisfy the following formula show PageRank (**c** is constant).

$$\boldsymbol{R} = \mathbf{c}\boldsymbol{M}\boldsymbol{R}$$

3.2 Betweenness Clustering

Betweenness Clustering [Michelle et al., 2001] is a method for detecting densely connected subgraphs from a graph. Betweenness of an edge show degree that this edge is included by all shortest routes between all pairs of nodes. Betweenness Clustering proceed clustering by removing higher betweenness links from link structure.

4 Results

We collect 80 blogs and 242 links by "Baseball" seed blogs. Result of Removing 100 links by Betweenness Clustering of these blogs is Figure 1.

Figure 1: Betweenness Clustering of "Baseball" Blogs.

In two clusters, most of their members have different topic each other, but are hosted by same blog hosting services. While in other three clusters, most of their members have similar topic, such as the same baseball team, but are hosted by different blog hosting services. Though in other two cluster, most of their members have similar topic, and same blog hosting services.

In top ten of PageRank of "Baseball", 1, 2, 3, 4, 5, 9th blogs are not related with baseball, but 6, 7, 8, 10th blogs are related with baseball. This result is considered as topic-distillation problem that also appears in community mining from a Web page link structure.

We collect 73 blogs and 212 links by "Winny" seed blogs. By betweenness clustering, two clusters consist of blogs that are hosted by the same hosting services. However, other clusters don't seem to consist any meaningful communities. Because Winny is a temporary social topic, there might not be any stable communities about Winny.

5 Conclusion

In this research, we apply Web structure mining approach to blogs. Our conclusions are: PageRank can not efficiently determine important blogs for a given topic. The results might be better if we consider the effect of dangling links. Second, Betweenness clustering is effective to extract stable communities. To experiment on more blogs and more topics is our future work.

Reference

[Blood et al., 2002] Rebecca Blood. We've Got Blog: How Weblogs are Changing Our Culture. Perseus Publishing, 2002.

[Page et al., 1997] Page Lawrence, Brin Sergey, Motwani Rajeev, and Winograd Terry.
The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Library working paper 1997-0072, 1997.

[Michelle et al., 2001] Michelle Girvan and M. E. Newman. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99, 7821-7826, 2002.