Community Discovery and Analysis in Blogspace

Ying Zhou, Joseph Davis
School of Information Technologies, The University of Sydney, Australia
{zhouy, jdavis}@it.usyd.edu.au

ABSTRACT

Weblog has quickly evolved into a new information and knowledge dissemination channel. Yet it is not easy to discover weblog communities through keyword search. The main contribution of this paper is the study of weblog communities from the perspective of social network analysis. We proposed a new way of collecting and preparing data for weblog community discovery. The data collection stage focuses on gaining knowledge of the strength of social ties between weblogs. The strength of social ties and the clustering feature of social network guided the discovery of weblog communities.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]:

Clustering and Retrieval models.

General Terms

Algorithms, Experimentation,

Keywords

Weblog, Community, Social Network, Social tie

1. INTRODUCTION

Weblogging has been the latest Internet killer application, with thousands of blogs added to the Internet everyday. For a simple definition, weblogs are web pages with several dated entries usually arranged in reverse chronological order[3]. Each entry of a blog has its own "*permalink*"(Permanent URL address). In most cases, all entriess of a particular weblog are written by a single author. Sometimes, a weblog can have a few co-authors. The authors of blog entries are called bloggers.

The rapid growth in weblogs promotes a new, dynamic and loosely organized on-line community activity. Such communities are usually formed by bloggers listing each other on the side bar of their weblogs and especially by commenting on and citing each others's entries. The prominent feature of those communities is the speed and accuracy of capturing the latest discussion and information on an array of t opics ranging from politics to specific technologies. More and more Internet users are keen on locating and following the discussion of topics of their own interest from particular communities. Yet it is very difficult to discover such online communities, especially those that discuss less popular topics. Currently, people get to know those communities either through word of mouth, or through limited categorization or search services provided by well known websites such as New York Times which maintains a list of popular blogs under different categories. There are also a few specialized search engines such as Google's BlogSearch, bloglines.com, daypop.com and blogdex.com. Those specialized blog search services perform page level retrieval and are similar from traditional web keywords based search except that they have more focused content.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

Internet users perform web search and weblog search for entirely different purposes. They use web search to obtain information. For instance, we may enter keyword like "Voronio partition" in Google to get the explanation or definition of this term. This is most likely an one-off action. Only on rare occasion, we might bookmark the result page for further reference. Weblog search, on the other hand is meant to be the first step of a sustainable publish/subscribe process. For instance, if we type "web services" in bloglines.com, our purpose is to find the community of bloggers that actively write on "web services" and keep tracking what they are talking about for a certain period of time. We are not looking for one-off information.

The contribution of this project is the study of weblog communities from the perspective of social network analysis. We proposed a new way of collecting and preparing data for community search which emphasizes gaining information on the strength of social ties between weblogs. The social tie strength data are used as key knowledge in discovering closely knitted communities from a large set of interwoven weblogs.

2. METHODOLOGY

2.1 Blog Communities and Social Networks

A blog community is the network of bloggers who write on similar topics and from time to time read, write and comment on each others' posts. A blog community can be viewed as a special type of social network which is defined as a set of people or groups connected with each other under a particular relationship[4]. Examples of typical social network include the friendship network of high school students and scientific co-authorship network of academics. In social network terminology, the people or groups are called "actors" and the connections are called "ties". In a blog community context, the bloggers who write weblogs are the actors and the hyperlinks between weblogs are ties.

Bloggers generally have *read* and *respond* relations with each other. These are indicated by hyperlinks between weblogs. The blogroll lists the *read* relations while hyperlinks appear in the entry bodies or entry comments represent the *respond* relations. Each hyperlink is a communication instance between two bloggers. The number of instances would indicate the strength of tie between bloggers. Bloggers in a community should have strong ties with other members[2].

2.2 Blogspace acquisition

The starting point of weblog community identification is a connected blogspace with relatively similar topic. Since weblog rather than individual web page is the unit of analysis,a special weblog crawler is developed to construct the blogspaces.

The weblog crawler takes a weblog URL as seed and incrementally adds ties between weblogs in the collection. Let us consider the simplest case which finds all tie instances from the seed to other weblogs. The simple algorithm has two steps:

1. Construct a large candidate set of hyperlinks find in all pages of the seed weblog.
2. Remove hyperlinks pointing back to the seed (including home page, entry pages, or archive pages) and hyperlinks not pointing to weblogs. The remaining set consists of hyperlinks pointing to other weblogs. These hyperlinks may appear in the blogroll, entry body or comment section. Each hyperlink in the remaining set is considered as an instance of some certain tie.

The basic algorithm will generate a star shaped graph with seed in the center and all edges coming out from it. To construct the blogspace, we can repeat the basic algorithm for each distinct weblogs appear in the target end of the edges discovered so far. Table 1 gives the crawling algorithm. In implementation, step 1 and step 2 are executed in pipeline style for efficiency and a variable *depth* is used to control the size of the blogspace.

Table 1 The main weblog crawler algorithm

```
1   Url = seed url; depth  = 0;
2   method crawl (Url, depth)
3     if (depth < maxDepth)
4       for all hyper links link in Url
5         if  link belongs to the same weblog
6               crawl (link, depth)
7         else if link points a weblog entry
8               find the home page of link as link.home
9               add a record Url.home and link.home
10              crawl(link.home, depth + 1)
11        end if
12      end for
13    end if
14  end method.
```

2.3 Blog community extraction

The blogspace constructed usually contains several thousands or more unique weblogs. We expected to discover community structure from it. To construct the directed graph, we first removed the multiple lines between two vertices and assign the line value as the number of lines. This is interpreted as the strength of tie between weblogs.

We adopt the island partitioning algorithm developed by Vladimir Batagelj[1] to extract blog communities. Here island is defined as a connected small sub-network of size [*min*, *max*] with stronger internal line weight than line weight to vertices outside the sub-network. Island partition is a hierarchical clustering method. The algorithm first orders all ties in decreasing order of their strength. It then, in the sorted order, merges ties to form sub-network, based on common vertex of sub-networks. All sub-networks form a hierarchical structure, with possibly a common root. Desirable sub-networks were chosen from the hierarchy based on the size range. It was set to [5, 50] in the experiment.

3. EXPERIMENTS

As a case study, we took Savas Parastadist's weblog (**savas.parastatidist.name)** as seed to get an interconnected blogspace. It contains around 3800 unique weblogs and over 33000 lines. Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/) is used to perform the partition and the visualization.

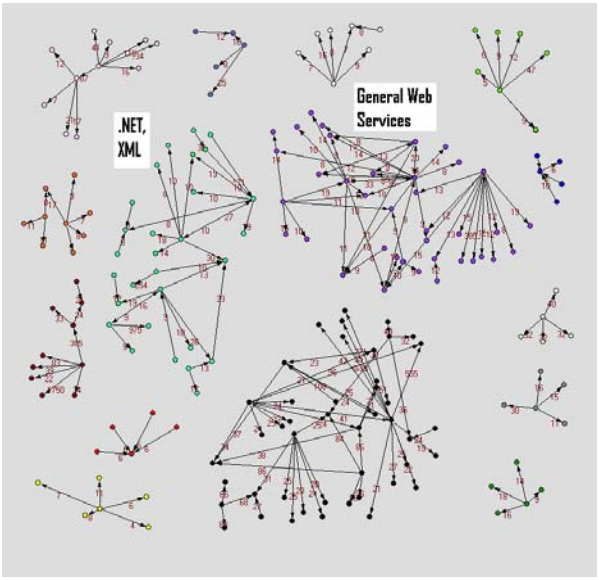

Figure 1 Web Services blog communities

15 communities were identified from the blogspace, each with its own theme. Figure 1 is a visual display of the communities and members. We only illustrate the communities of "general web services" and community with a focus on ".NET and XML" technology. Majority of the community members we discovered are not able to be located through current blog search engines. We used the keyword "web services",".NET","ASP.NET" in Google's BlogSearch, bloglines.com, blogdex.net and daypop.com and checked the top 20 results returned. Except for blogdex.net, which contains some web services weblogs, none of the rest has those members returned.

4. CONCLUSIONS

In this paper, we studied the elements of social ties between weblogs and use the tie strength data as key knowledge to discover closely related communities from large and loosely connected blogspace. There are many applications for the methods we proposed. For instance it can be used to prepare and search local copy of web pages to build the blogspace. There are also opportunities for future work to improve the clustering techniques we used here.

5. REFERENCE

[1] Batagelj V. Analysis of large networks – Islands Presented at Dagstuhl seminar 03361: *Algorithmic Aspects of Large and Complex Networks* Dagstuhl, 2003

[2] Granovetter M. The Strength of weak ties: network theory revisited. *Sociological Theory*, Vol. 1 (1983), 201-233

[3] Kumar R., Raghavan P., and Tomkins A. Trawling the Web for emerging cyber-communities. In *Proc. 8thth WWW Conference, 1999*

[4] Wasserman S. and Faust K., *Social Network Analysis*, Cambridge University Press, Cambridge, 1994