Graph Mining Approaches for the Discovery of Web Communities

Tsuyoshi Murata[1,2]

[1] National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan
tmurata@nii.ac.jp
http://research.nii.ac.jp/~tmurata
[2] Japan Science and Technology Corporation
Yoyogi Community Building, 1-11-2, Yoyogi, Shibuya-ku,
Tokyo, 151-0053 Japan

Abstract. Finding related Web pages is important for assisting users' information retrieval from the Web. In general, related Web pages are densely connected with each other by hyperlinks, and graph mining approaches are applicable for discovering such clusters of related Web pages, which are called Web communities. Among the research of Web structure mining based on the graph structure of hyperlinks, discovery of Web communities is one of the important research topics. In this paper, recent approaches for the discovery of Web communities are introduced, and requirements for graph mining algorithms suitable for the discovery of Web communities are discussed.

1 Introduction

At the time the author is writing this paper, Google indexes more than 3 billion Web pages in the world. The goal of Web mining is to utilize this huge Web network. The Web can be regarded as a graph if we regard each Web page as a vertex and each hyperlink as an edge. Web structure mining is based on the graph structure of hyperlinks, and is one of the important research topics that graph mining algorithms are really required. There are several goals for Web structure mining, such as ranking important Web pages [4][8], discovery of Web communities [3][5], analysis of the Web graph from macroscopic point of view [2], and modeling and simulating the process of Web graph generation [1]. Among these, discovery of Web communities (clusters of related Web pages whose hyperlinks are densely connected) is important in order to assist users' information retrieval from the Web. However, applying graph mining algorithms to the Web is not simple since it is huge and is growing. There are some requirements for graph mining algorithms in order to handle Web data, such as partiality of input data and robustness for missing data.

This paper introduces some methods for the discovery of Web communities as an application of graph mining in order to clarify their characteristics. Requirements to graph mining algorithms for handling Web data are also discussed.

2 Methods for Discovering Web Communities

There are two main approaches for the discovery of Web communities; 1) search of fixed-size graph structure from Web snapshot data, and 2) decomposition of given Web graph into densely connected components. Both are explained below, followed by our own approach.

2.1 Search of fixed-size graph structure

For example, Web pages of aircraft enthusiasts often have hyperlinks to the companies of aircraft manufacturers. Hyperlinks of these pages (enthusiasts and companies) compose a bipartite graph and all of these pages are closely related. Kumar's trawling [5] is based on an assumption that Web pages constituting a bipartite graph are regarded as an indication of Web community sharing common interest. In his experiments, bipartite graph structures are enumerated by applying a priori algorithm. In addition to that, randomly selected samples are investigated by manual inspection. Its results show that most of the pages constituting a bipartite graph are actually closely related.

2.2 Decomposition of Web graph into densely connected components

In general, hyperlinks of related Web pages are densely connected with each other rather than others. Flake [3] applies maximum-flow minimum-cut theorem of network flow theory in order to discover densely connected components, which can be regarded as Web communities. His approach is often explained by the following metaphor: if edges are water pipes and vertices are pipe junctions, the maximum flow problem tells us how much water we can move from one junction to another, and the maximum flow is proved to be identical to minimum cut. Therefore, if you know the maximum flow between two points, you also know what edges you would have to remove to completely disconnect the same two points, which are called cut set. The approach accepts some Web pages as seeds of target Web community, and finds cut set that disconnect a component containing given seed pages.

2.3 Search of bipartite graphs based on data acquired from a search engine

A search engine can be regarded as a resource for Web data acquisition. The author proposed a method for discovering Web communities from the data acquired from a search engine [6][7]. Our method is similar to Kumar's one since both search bipartite graph structures. However, they are different in the following points:

1. Search of bipartite graphs from partial Web data without using Web snapshot data

Previous approaches of Web community discovery require relatively large-scale Web snapshot data. However, collecting Web data and maintaining them is not an

easy task. It is pointed out that the difference between Web snapshot data used for mining and actual Web data may cause the discovery of outdated Web communities [5]. Major search engines contain much updated Web data and they can be used for Web data acquisition in order to achieve relatively new Web communities. Some of the search engines allow users to access contained data, such as Google API.

2. Acquisition of backlinks from a search engine in order to follow hyperlink backward

Although most users use search engines in order to find Web pages about some keywords, a search engine enables us to follow hyperlinks backward. By attaching some option (such as "link:") to input URL, Web pages that contain hyperlinks to input URLs can be searched, which are called backlinks. Since hyperlinks to related Web pages often co-occur, backlink search enables us to find related Web pages.

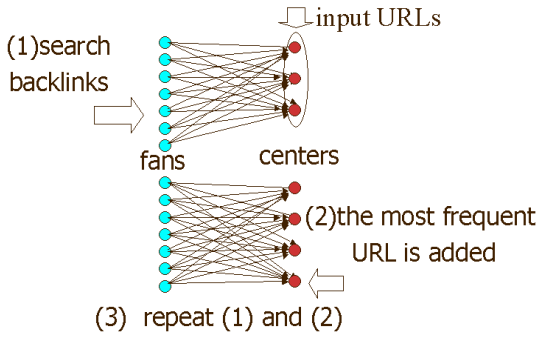

Fig. 1. Outline of our method for discovering Web communities

Fig.1 shows the outline of our method for Web community discovery. Our goal is to discover a bipartite graph containing some given URLs. At first, some URLs regarding specific topic (such as baseball or Macintosh) are given as initial centers, and fans which co-refer all of the centers are searched by backlink search on a search engine (step 1). HTML files of the searched fans are acquired through the internet, and all the hyperlinks contained in the files are extracted. The hyperlinks are sorted in the order of frequency. Since hyperlinks to related Web pages often co-occur, the top-ranking hyperlink of the sorted result is expected to point to a page whose contents are closely related to the contents of centers. Therefore, the URL of the page is added as a new member of centers (step 2). By using newly generated centers, the above steps are repeated in order to find more centers (step 3).

Although this method is quite simple, it succeeds in discovering many related Web pages. Experimental results show that 19.8 related centers are actually discovered from given 5 seed URLs on average [7].

3 Web Community Discovery as an Application of Graph Mining

As an application of graph mining algorithms, the following requirements should be considered for the discovery of Web communities:

1. Partiality of input data: Nobody can collect data of the whole Web. Algorithms for the discovery of Web communities need to handle partial Web data. Suitable strategies for collecting Web data have to be considered.
2. Quantities of input data: On the contrary to the above, Web data are still huge even though they are partial. Capabilities for handling large-scale data are required for Web community discovery methods.
3. Qualities of input data: Depending on the network conditions, some of the Web pages may not be accessible. Robustness for missing or noisy data is necessary for Web community discovery.
4. Various structure of Web communities: Although Kumar regards a bipartite graph as a characteristic structure for Web communities, there might be other characteristic graph structures. Search algorithms for specific graph structure, such as clique or bipartite graph, are important. However, they are not enough for discovering real complicated Web communities.
5. Post processing of discovered Web communities: When fixed graph structure is searched from given Web data, many overlapping graphs will be found. Post processing of discovered Web communities such as clustering or labeling is necessary to assist users' understanding.
6. Interactive discovery of Web communities: Discovered Web communities are not always satisfactory to users since there are several criteria for "relatedness" among Web pages. It is preferable if users can control the strategies for searching Web communities by giving examples or negative examples.

References

1. Barabasi, A.-L., "LINKED – The New Science of Networks", Perseus Publishing, 2002.
2. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: "Graph Structure in the Web: Experiments and models", Proc. of the 9th WWW Conference, pp.309-320, 2000.
3. G. W. Flake, S. Lawrence, C. L. Giles, F. M. Coetzee: "Self-Organization and Identification of Web Communities", IEEE Computer, Vol.35, No.3, pp.66-71, 2002.
4. J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins: "The Web as a Graph: Measurements, Models, and Methods", Proc. of COCOON'99, LNCS 1627, pp.1-17, 1999.
5. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins: "Trawling the Web for Emerging Cyber-Communities", Proc. of the 8th WWW Conference, 1999.
6. T. Murata: "Discovery of Web Communities Based on the Co-occurrence of References", Proc. of DS2000, LNAI 1967, pp.65-75, Springer, 2000.
7. T. Murata: "Finding Related Web Pages Based on Connectivity Information from a Search Engine", Poster Proc. of 10th WWW conference, pp.18-19, 2001.
8. L. Page, S. Brin, R. Motwani, T. Winograd.: "The PageRank Citation Ranking: Bringing Order to the Web", Online manuscript, http://www-db.stanford.edu/~backrub/pagerank sub.ps, 1998.