Discovery of Web Communities Based on the Co-occurrence of References

Tsuyoshi Murata

Department of Computer Science, Gunma University
1-5-1 Tenjin-cho, Kiryu, Gunma, 376-8510 Japan
murata@cs.gunma-u.ac.jp

Abstract. This paper proposes a method of discovering Web communities. A complete bipartite graph $K_{ij}$ of Web pages can be regarded as a community sharing a common interest. Discovery of such community is expected to assist users’ information retrieval from the Web. The method proposed in this paper is based on the assumption that hyperlinks to related Web pages often co-occur. Relations of Web pages are detected by the co-occurrence of hyperlinks on the pages which are acquired from a search engine by backlink search. In order to find a new member of a Web community, all the hyperlinks contained in the acquired pages are extracted. Then a page which is pointed by the most frequent hyperlinks is regarded as a new member of the community. We have built a system which discovers complete bipartite graphs based on the method. Only from a few URLs of initial community members, the system succeeds in discovering several genres of Web communities without analyzing the contents of Web pages.

1 Introduction

According to an announcement of Inktomi and NEC Research Institute, the number of Web pages in the world surpasses one billion documents as of January 2000[8]. In order to find useful pages from such vast Web network, many users rely on search engines such as AltaVista or Yahoo. However, they often find difficulty because of ubiquitous synonymy (different words having the same meaning) and polysemic (the same word having multiple meanings). A system which discovers related Web pages is expected to assist users’ information retrieval from the Web.

In general, Web pages contain various forms of information such as sentences, images, and sounds. Understanding all of such contents and classifying the pages appropriately are not easy tasks even for humans. A large number of studies have been made for finding relations of Web pages based on linguistic information. Broder defines document similarity as the ratio of common subsequence of words[3]. Chang employs TFIDF as the criterion for document similarity[6]. Although these methods are widely applicable to ordinary documents, utilizing hyperlinks, which are the information peculiar to Web pages, is expected to help greatly for accurate classification of Web pages.

© Springer-Verlag Berlin Heidelberg 2000
This paper proposes a method of discovering communities of Web pages based on the co-occurrence of references. A system for discovering Web communities is developed based on the method. The input to the system is a few URLs of Web pages about specific topics such as movies or sports. The output of the system is the communities of Web pages sharing common interests with the input pages.

In this paper, a community is defined as a set of Web pages whose hyperlinks form a complete bipartite graph $K_{i,j}$, as mentioned in Kumar’s paper of Web Trawling[13]. In $K_{i,j}$ graph, each of $i$ pages contains hyperlinks to all of the $j$ pages. We will call the former $i$ pages as fans and the latter $j$ pages as centers in this paper. The procedure for community discovery is as follows; first, our system acquires fans which have hyperlinks to all the input centers by backlink search on a search engine. From the HTML files of the acquired fans, all the hyperlinks are extracted. Then a page which is pointed by the most frequent hyperlink is added as a new member of centers. By repeating these two steps, a Web community is searched without analyzing the contents of Web pages. Since the method utilizes backlink information, implicit relation between the pages which have no direct hyperlinks to each other can be detected. The system succeeds in discovering several genres of communities only from a few input URLs.

2 Related Work

Hyperlinks often give authority to the contents they point to. Several attempts have been made on link analysis[10] and Web visualization[9][14] since hyperlinks are expected to give important clues for finding relations among Web pages. However, this assumption is not always true because of the following reasons. First, hyperlinks between two pages in the same Web site very often serve a purely navigational function and typically do not represent a referral of authority[12]. Second, related Web sites frequently do not reference one another because they are rivals, they are on opposite sides of thorny social issues, or they simply do not aware of each other’s presence[13]. In order to find relations among Web pages, it is not enough to investigate hyperlinks from the pages; hyperlinks to the pages, which we call backlinks, are often more important. Although it is not possible to search all the backlinks that point to an arbitrary Web page, many of them can be obtained by backlink search on a search engine.


In order to measure the degree of relation between two URLs, this system