Finding Related Web Pages Based on Connectivity Information from a Search Engine

Tsuyoshi Murata
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo 101-8430, JAPAN
+81-3-4212-2555
tmurata@nii.ac.jp

ABSTRACT

This paper proposes a method for finding related Web pages based on connectivity information of hyperlinks. As claimed by Kumar, a complete bipartite graph of Web pages can be regarded as a Web community sharing a common interest. However, preparing Web snapshot data for the search of such communities is not an easy task since the Web is huge and is growing. In our method, connectivity information is acquired from a search engine by backlink search. A system based on the method succeeds in finding several genres of Web communities only from a few input URLs without analyzing the contents of Web pages.

Keywords

Web community, hyperlink, complete bipartite graph, search engine.

1. INTRODUCTION

It often happens that a user is already familiar with some Web pages of specific topic and needs to find more pages about the topic. A system for finding related Web pages is expected to assist users' information retrieval from the Web. This paper proposes a method for finding related Web pages, which we call Web community [4], based on the structure of hyperlinks. In our method, the input is a few URLs of Web pages about specific topic such as movies or sports, and the output is the communities of Web pages sharing common interests with the input URLs. A system based on the method succeeds in finding several genres of communities without analyzing the contents of Web pages.

2. A Method for Finding Web Communities

The goal of our method is to find a complete bipartite graph of Web pages which contain input URLs. A complete bipartite graph K*i,j* is composed of a set of *i* pages and a set of *j* pages: each of the *i* pages has hyperlinks pointing to all of the *j* pages. In the following explanation, *fans* refer to the set of *i* pages and *centers* refer to the set of *j* pages. The overall procedure consists of the following three steps.

1. Search of fans using a search engine
2. Addition of a new URL to centers
3. Sort of centers in the order of frequency

2.1 Search of fans using a search engine

In our method, input URLs are accepted as initial centers, and fans which co-refer all of the centers are searched. As shown in Figure 1, fans are searched from the centers by backlink search on a search engine. In general, popular centers may have too many backlinks. In such cases, a fixed number of high-ranking URLs are selected as fans.

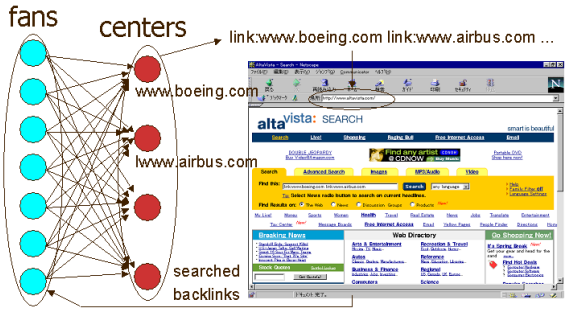

Figure 1. Search of fans using a search engine

2.2 Addition of a new URL to centers

The next step is to add a new URL to centers based on the hyperlinks of acquired fans. The fans' HTML files 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 is expected to point to a page whose contents are closely related to the centers. As shown in Figure 2, the URL of the page is added as a new member of centers.

In general, the number of fans decreases according as the number of centers increases. The above two steps are repeatedly applied until there are few fans which refer all the members of centers. Acquired centers are regarded as a Web community.

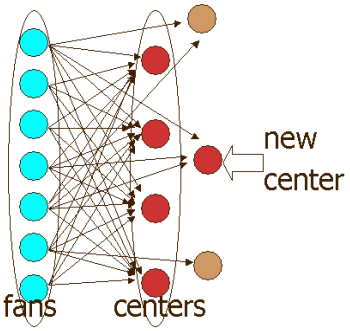

Figure 2. Addition of a new URL to centers

2.3 Sort of centers in the order of frequency

It is desirable for a user to rank the URLs of communities in the order of relevance to input URLs. In order to achieve such ranking, communities for every pair of input URLs are found. For example, if five URLs are provided, ${}_5C_2 = 10$ communities are found from the pairs of 1st & 2nd URLs, 1st & 3rd URLs, … and 4th & 5th URLs. Then the centers of all the communities are sorted in the order of frequency. The sorted result is expected to reflect the rank of relevance to input URLs because highly ranked URLs co-occur many times with the input URLs.

3. Experiments

Based on the above method, a system for finding Web communities is developed. It is desirable that input URLs are popular ones that many others refer to. As the input to the system, URLs of 100hot.com (http://www.100hot.com/) are used in our experiments. 100hot.com is the site of ranking one hundred famous Web pages for several genres. In order to find communities for 33 genres (which are not sponsored by companies concerned), top five URLs of each genre are provided to the system as inputs. Our system finds a community for every pair of the input URLs and outputs the centers of all the communities sorted in the order of frequency.

In order to evaluate the quality of the system's output URLs, the ranking of 100hot.com are regarded as the collection of "correct answers". In another words, if a URL is listed in the 100hot.com ranking of corresponding genre, it is regarded as a "correct answer", otherwise it is regarded as an "incorrect answer". Since there are many relevant URLs which are not listed in 100hot.com site, this evaluation criterion is rather too severe for the system. However, we dare to employ this criterion since it clarifies the power of our system. As a search engine for backlink search, AltaVista is used.

genre	total	correct	1Q	2Q	3Q	4Q
Art	45	2	2	0	0	0
Books	117	12	10	2	0	0
Events	172	6	4	0	0	2
Music	184	28	6	11	8	3
Finance	130	42	21	10	4	7
Jobs	73	27	15	6	6	0
Loans	142	12	10	0	0	2
College	196	47	20	13	7	7
Kids	172	42	30	3	9	0
Gambling	53	16	6	5	4	1
Movies	89	8	4	0	1	3
Games	137	36	17	8	10	1
Family	93	3	2	1	0	0
Food	94	7	4	1	0	2
Gardening	148	6	2	1	1	2
Pets	146	18	9	1	0	8
Cars	92	40	13	16	6	5
Chat	96	4	0	3	0	1
Dating	62	10	5	4	0	0
Spirits	171	21	4	4	13	0
Travel	124	38	12	10	16	0
Magazines	74	12	6	1	1	4
Newspape	167	40	11	23	3	3
Auction	158	18	9	6	3	0
Flowers	141	15	8	1	0	6
Shopping	162	21	6	6	9	0
Health	130	14	10	1	1	2
Sports	95	26	14	7	4	1
Developer	123	9	4	2	3	0
Hardware	164	29	17	9	2	1
Mac OS	143	29	19	2	7	1
Unix	95	8	5	2	1	0
Windows	130	9	5	1	2	1
average	124.8	19.8	9.4	4.8	3.7	1.9

Table 1. Results of the experiments

The results of the experiments are shown in Table 1. The first column of the table shows genres. The second and third column of the table (total, correct) show the total number of acquired centers for corresponding genre, and the number of "correct answers" among them respectively. From fourth to seventh column (1Q, 2Q, 3Q, 4Q) show the number of "correct answers" in each quarter of ordered list of output URLs. For example, "1Q" shows the number of "correct answers" which are located in the first quarter of the list of output URLs.

Table 1 shows that the system performs very well for many genres. The system finds 19.8 correct answers on average only from five input URLs. As a detailed example, the result of genre Kids is shown below. The following five URLs are given to the system as inputs: www.pbs.org, www.headbone.com, www.bolt.com, www.yahooligans.com, www.discovery.com .

The top 10 of output URLs are as follows: www.cyberkids.com, www.ctw.org, www.exploratorium.edu, www.si.edu, www.bonus.com, www.kids-space.org, www.discovery.com, www.youruleschool.com, www.planetzoom.com, www.kidscom.com. All of these URLs except www.planetzoom.com (9th) are listed in genre Kids of 100hot.com. If you watch the contents of each URL, you will agree that the 9th URL is also a site for kids although it is not listed in 100hot.com. These results show that the system has abilities of finding many URLs which are related to input URLs.

4. Visualizing the Structure of Communities

Besides the above-mentioned system, the author has developed a system [3] for visualizing URLs. Based solely on hyperlink structure, the system generates a graph in which related URLs are located close to each other. Its online demonstration is available at http://www.cs.gunma-u.ac.jp/~tmurata/. Figure 3 shows a part of visualized financial community that is found in the experiments of section 3. At the center of this star graph, stock exchanges such as www.nyse.com and www.amex.com are located. This graph reflects the structure of the community since stock exchanges often play the central role in finance.

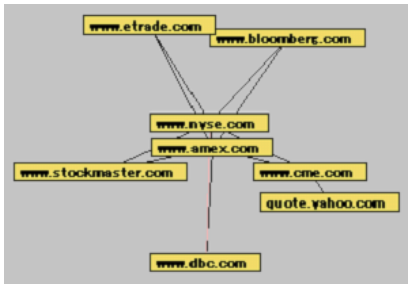

Figure 3. Visualizing the structure of financial community

5. Related Work

In order to find related pages or to rank pages based on the structure of hyperlinks, several researches have been made such as Clever project [1], Kumar's Web Trawling [2], and PageRank algorithm [5]. It is true that these three approaches are effective, but they require large-scale data of HTML files. Since the Web is huge and is growing, preparing such data is not a simple task. Our system acquires relatively new connectivity information from a search engine during the process of finding communities. This enables our system to find Web communities that are not outdated.

6. Conclusion

This paper describes a method for finding communities of related Web pages based on the structure of hyperlinks. We would like to know how far we could go with hyperlink information alone. More studies on characteristic structures of communities will make our system more powerful.

7. References

[1] Clever Project, "Hypersearching the Web", Scientific American, http://www.sciam.com/1999/0699issue/0699raghavan.html, 1999.

[2] R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, "Trawling the web for emerging cyber-communities", Proc. of WWW8, 1999.

[3] T. Murata, "Machine Discovery Based on the Co-occurrence of References in a Search Engine", Proc. of DS99, Lecture Notes in Artificial Intelligence 1721, pp.220-229, 1999.

[4] T. Murata, "Discovery of Web Communities Based on the Co-occurrence of References", Proc. of DS2000, Lecture Notes in Artificial Intelligence 1967, pp.65-75, 2000.

[5] L. Page, "The PageRank Citation Ranking: Bringing Order to the Web", http://google.stanford.edu/~backrub/pagerank sub.ps.