A Hyperlink-Induced Method for Extracting Web Communities of Personal Interests

Toyohisa Nakada

School of Knowledge Science,
Japan Advanced Institute of Science and Technology
March 2002

Keywords: Internet / WWW intelligence, Information gathering, KDD

1. Objective

The extraction of beneficial information from World Wide Web has recently received much attention from researchers. One attractive branch of this research area is discovery of Web communities. A Web community is a group of Web sites that share common interests.

A Web community can generally be split into two parts. One is a group of Web sites that share common interests. The other is a set of URLs that link to the Web community's URL. The former is called "Web community" or "centers" and the latter is called "Hubs" or "Fans". The basic principle in our studies is such "Fans" can be applied to studying human relationship. We suppose that a personal homepage shows a part of the person's interests. Therefore, a group of personal homepages that link to one Web community should be a group of people that have common interests.

The purposes of the work presented in this paper are:

(1) To extract Web communities of personal interests from specific domains (ex. school's, ISP's, company's www site) using hyperlink analyses;
(2) To summarize common interests by observing such extracted Web communities.

Finding Web communities of personal interests can be considered as a new and significant problem in the area of discovering Web communities.

2. Method for Extracting Web Communities of Personal Interests

The method proposed in this paper is based on the hypothesis that a Web community

implies interests of persons each of them has his/her Web site containing at least one URL linking to the URLs that are contained by the Web community. The following is the outline of the method:

(1) gathers hyperlinks from personal homepages in a specific domain.
(2) sorts the hyperlinks in the order of OScore that is proposed in this paper and gives high score if the URL is a well-known one in a specific domain and not in general.
(3) gets one URL as a seed of a Web community from the hyperlinks, and gathers URLs that are similar to the seed from the hyperlinks, and then created groups of URLs are Web communities. (We also proposed similarity measure between two URLs in this paper that is defined by Jaccard coefficient.)

Additionally, we developed a visualization system based on spring model that is well-known technique for drawing general undirected graphs to explore obtained Web communities.

3. Experiments and Evaluations

According to the goal of extracting Web communities of personal interests, the following are questions we would like to examine in order to evaluate the method.

Question 1 Whether one can understand what is the topic of the Web community obtained by the proposed method?

Question 2 Whether obtained Web communities likely to be valid from the viewpoint of implying personal interests?

Experiments have been carried in five domains to answer these questions, and we used the questionnaire to evaluate the result obtained by our method.

Following is an interpretation of the experiment results in terms of two questions.

Regarding the question 1 it seems to be quite reasonable to consider 80% discovered Web communities are understandable. According to many related other works on discovering of Web communities, the result we obtained could be considered to be significant.

Regarding the question 2 it is satisfactory to attain that from 40% to 50% Web communities can explain some interest. We tried a new problem of summarizing common interests by observing obtained Web communities in a specific domain. In such a situation, the result shows that our method was also a valuable one.