JAIST Repository
https://dspace.jaist.ac.jp/

Title	個人の興味を映し出すWeb Communityの抽出方法の提案
Author(s)	中田, 豊久
Citation	
Issue Date	2002-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/367
Rights	
Description	Supervisor:Ho Tu Bao, 知識科学研究科, 修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

修　士　論　文

個人の興味を映し出す
Web Community 抽出手法の提案

指導教官　　Ho Tu Bao　教授

北陸先端科学技術大学院大学
知識科学研究科知識システム基礎学専攻

050059　中田　豊久

審査委員：　Ho Tu Bao　教授（主査）
石崎　雅人　助教授
中森　義輝　教授
佐藤　賢二　助教授

2002 年 2 月

Copyright © 2002 by Toyohisa Nakada

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

By Toyohisa Nakada

A thesis submitted to
School of Knowledge Science,
Japan Advanced Institute of Science and Technology
in partial fulfillment of the requirements
for the degree of
Master of Knowledge Science
Graduate Program in Knowledge Science

Written under the direction of
Professor Tu Bao Ho

February 13, 2002

Copyright © 2002 by Toyohisa Nakada

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter explains our basic ideas and describes structure of this paper.

1.1 Web Communities

The extraction of beneficial information from World Wide Web has recently received much attention from researchers. Web has huge information and it is useful resources for computer exploitation. Meanwhile, data mining is a growing research area that stands on artificial intelligence, statistics, and so on. One purpose in data mining is dealing with huge datasets. Therefore, it is natural to apply data mining to Web data, called "Web Mining".

Web Mining consists of three parts. One is Web Contents Mining that is to extract beneficial information from Web contents (a text, a style, and so on). The second is Web Usage Mining that mainly analyzes access log on Web server in order for us to understand the user behavior or to improve the Web site toward a well-designed Web site. The third is Web Structure Mining that mainly analyzes Web structure in order to discover useful information or to get well designed Web robots, and so on. This area includes researches of discovery of Web communities. A Web community is a group of Web sites that share common interests. The major purpose of finding Web communities is to find new beneficial Web pages from the Internet.

1.2 Application of Web Communities

The major purpose of finding Web communities and link analyses is to find new

beneficial Web pages from the Internet. For example, HITS [Kleinberg 99], which is an influential work in this area, outputs a set of URLs related to keywords given to the system, or PageRank [Page 98] is a ranking system in Google [Google]. On the other hand, it has been suggested that Web data can imply human relations. For example, REFERRAL [Kauts 97] shows networks of researchers by using co-occurrence citations in the document on the Web.

It is reasonable to suppose that Web communities can serve to study human relationships. 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 following are the major benefit of extracting such groups.

First, groups that have common interests can be used for human recommender systems. When you have something you need to find out about, one of the effective ways to approach the problem is to find people who share the same interest with you.

Second, it is important to manage knowledge assets from the viewpoint of knowledge management [Nonaka 01]. It is not right to assume that finding groups that have common interests means finding knowledge assets. But we think distributions of interest hint a part of knowledge distribution because having interests is the first step toward acquisition of knowledge.

1.3 Objective of Our Research

The following are the purposes of this study.

- To extract Web communities of personal interests from specific domains (ex. school's, ISP's, company's www site) using hyperlink analyses.
- 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.

We began our study by creating the method for extracting Web communities in the specific domain. And we used a questionnaire to evaluate our method.

1.4 Structure of this paper

This paper consists of 6 chapters. Chapter 1 (this chapter) is describing introduction. We introduced our research and the research area, and made clear our objectives. Chapter 2 that is "Web Mining" describes our research area in detail. Chapter 3 that is "A Method for Extracting Web Communities of Personal Interest" explains our method. Chapter 4 that is "Implementation" explains how we implement our method. Chapter 5 that is "Experiment and Evaluation" explains the experiment in order to evaluate our method and shows that results. Chapter 5 that is "Conclusion" is conclusion of this paper.

Chapter 2

Web Mining

This chapter explains a research area of Web Mining and the difference between our research and existing researches on this topic.

2.1 Taxonomy of Web Mining

Web Mining is defined as discovery and analysis of useful information from the World Wide Web. Its taxonomy, introduced in some papers, makes clear the position of discovery of Web Communities.

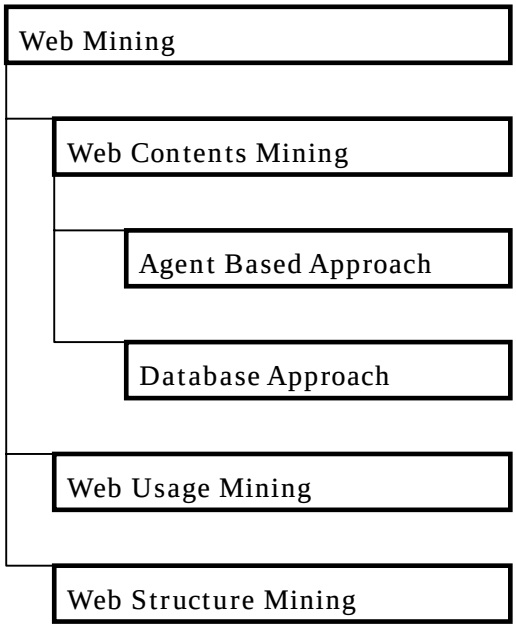


Figure 2.1 Taxonomy of Web Mining

Web Mining consists of three parts. One is Web Contents Mining that consists of two parts: Agent Based Approach and Database approach. Agent based approach

is to collect and organize information that meets each user's need or interests by using an agent technology. Database approach is to translate half structured data to structured data. Web contents are a kind of half structured data, while database is structured data. In general, structured data can be analyzed more easily than half structured data.

The second is Web Usage Mining that mainly analyzes access log on Web server in order to allow us to understand the user behavior or to improve the Web site toward a well-designed Web site. Some commercial software uses the word "Web Mining" to mean Web Usage Mining.

The third is Web Structure Mining that mainly analyzes the Web structure in order to discover useful information or get well designed Web robots, and so on. This area includes researches of discovery of Web communities.

In our understanding, our research belong to the both of Web Contents Mining and Web Structure Mining, because our method that is explained later uses a hyperlink as contents of personal homepages and network density of a bipartite graph [Stanley 94] in order to calculate similarity between two URLs. The former belongs to Web Contents Mining, and the latter belongs to Web Structure Mining.

2.2 Related Works

The first one is not any kind of research, but it says that groups that share common interests based on a specific keyword can be found easily. Recently some search engines can search with a given keyword in a restricted domain. For example, a domain could be "JAIST" (Japan Advanced Institute of Science and Technology to which we belong), a keyword could be "java", the searched results with such a restriction can be called as a group based on an interest in "java" within JAIST. But it is impossible to answer the following question. What kinds

of groups that share common interests are there in JAIST? That means no keywords are given to guide the search. We want to be able to answer such a question.

There is a system called REFERRAL [Kauts 97] that shows the relation between persons using co-occurrence of names on documents in the Internet. For example, there are two persons A and B. If there are lots of documents that describe both A and B, the relation between A and B is strong. This research is similar to our research in the purpose to extract human attributes implied by the Internet, but this system can show relations of only famous people. It is impossible to show relations of people in general whose names do not appear many times in the Internet.

HITS [Kleinberg 99] is the one of famous algorithms for extracting Web communities. HITS defined "authority page" as linked page, and "hub page" as a page that link to authority page. The basic principle in that paper is that a good authority is a page pointed to by many good hubs and a good hub is a page that points to many good authority pages. HITS creates Web communities using an iterative algorithm that calculates authority and hub scores.

PageRank [Page 98] is another well-known algorithm implemented as a system for ranking Web pages in Google [Google b]. PageRank interprets a link from page A to page B as a vote, by page A, for page B. For each page, the probability distribution that users visit is calculated as the sum of the probabilities of pages that link to the page. The propagation of probabilities is iterated until they converge.

Web Trawling [Kumar 00] is to extract all Web communities in the whole Internet. It defined Web community as completely bipartite graph [Stanley 94] that is a graph in which nodes can be partitioned into two subsets, and all lines

are between pairs of nodes belonging to different subsets. The result of this system states that 5 million Web communities were found when one Web community was defined as 3×3 bipartite graph.

Murata proposed another method for extracting Web communities [Murata 00a]. The method is based on the assumption that hyperlinks to related Web pages often co-occur. The method requires a few URLs as input called "centers". Next, the method searches sites that have a hyperlink to centers using a search engine. Those sites are called "fans". After that, the method adds sites that are linked by all of fans and that is not contained by the "centers". The process for constructing Web communities is iterated with when there is no center to add.

Chapter 3

A Method for Extracting Web Communities of Personal Interests

This chapter explains our method for extracting Web communities of personal interests.

3.1 Basic Ideas

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. Our method extracts Web communities from a specific domain in order to understand what are the major interests in the domain.

Because of HITS and other methods require keywords as input to the systems, it is difficult to apply to extracting Web communities that imply personal interests from restricted domains. Recently some search engines can search with a given keyword in a restricted domain. The searched results with restricted scope can be called a Web community related with input keywords. But it is impossible to answer the question: what kinds of Web communities are there in the domain? We want to be able to answer such a question.

The purpose of Web Trawling [Kumar 00] is to extract all Web communities in the whole Internet. It would be the same purpose between our study and Web Trawling, if the scope for searching were changed from the whole Internet to restricted domain. However, when applying Web Trawling algorithm to

extracting restricted domain, we could not get good results. Web Trawling defined a Web community as a completely bipartite graph [Stanley 94]. The approach is seen to be of value to extracting Web communities from huge datasets such as the whole Internet, because an error can be pruned by using some pruning algorithm; an error is a Web community consists of coincidental URLs that do not have single topic. Because our target deals with small dataset compared with the whole Internet, the result we got by using a completely bipartite graph contained a lot of errors that should not be ignored.

From these arguments, we developed a new method for extracting Web communities from personal homepages in a restricted domain. The method extracts hyperlinks in a restricted domain and constructs Web communities based on similarity between two URLs.

3.2 Similarity Between Two URLs

Before we explain our method, we need to define a similarity measure between two URLs. We used this measure introduced in [Kauts 97], [Murata 00] in order to construct Web communities.

Definition 1. *Similarity* between URLi and URLj is defined by Jaccard coefficient [Salton 89]

$$Similarity\ (URLi, URLj) = \frac{The\ number\ of\ pages\ that\ link\ to\ URLi\ and\ URLj}{The\ number\ of\ pages\ that\ link\ to\ URLi + The\ number\ of\ pages\ that\ link\ to\ URLj} \quad (3.1)$$

Figure 3.1 shows an illustration of definition 1.

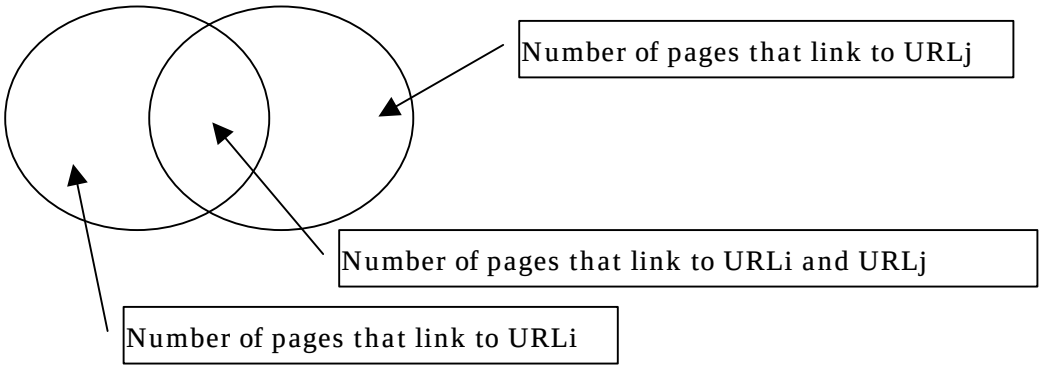


Figure 3.1 Similarity between two URLs

We would like to show another representation of definition 1. It is network density of a bipartite graph that is graph in which nodes can be partitioned into two subsets, and all lines are between pairs of nodes belonging to different subsets. The following is an illustration of a bipartite graph.

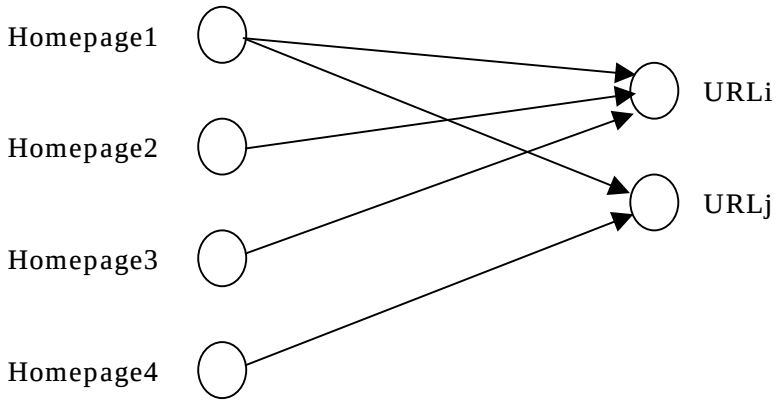


Figure 3.2 Network density of a bipartite graph

Figure 3.2 shows 62.5% density of a bipartite graph since the number of arrows is 5 and the number of possible arrows is $4 \times 2 = 8$, then the density is $5/8 = 62.5\%$. In passing, Similarity(URLi,URLj) on Figure 3.2 is $1/(3+2) = 0.2$.

In addition, if the network density is 100%, that is called a completely bipartite graph. Web Trawling [Kumar 00] defined a Web community as such a completely

bipartite graph and extract all Web communities from the Internet.

We used Similarity(URLi,URLj) for the following reasons.

- We can measure similarity without some analyses of Web contents.
- The use of search engines facilitates for us collecting the number of backlinks.

On the other hand, the following are the problems about Similarity(URLi,URLj).

- If the numbers of backlinks of URLi and URLj are imbalance, Similarity(URLi,URLj) always outputs small number that represents dissimilar.

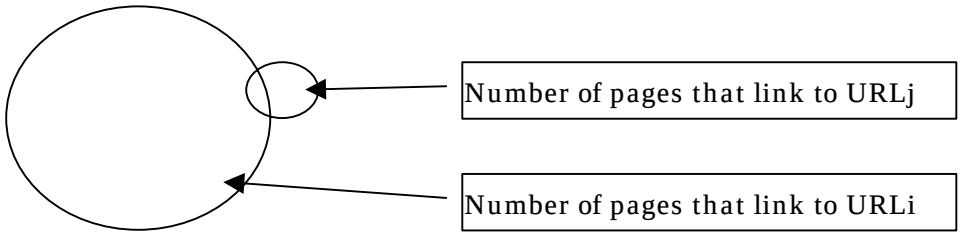


Figure 3.3 Unbalanced numbers of backlinks

We tried a new measure given in the following formula in order to avoid this problem.

Similarity(URLi,URLj)=Number of pages that link to URLi and URLj / Number of smaller one among number of pages that link to URLi and URLj

But we could not get better result than definition 1.

- The network load is exponential increases. The number of requests to a search engine for getting the number of backlinks is the following:

$$request = n + \binom{n}{2} \quad (3.2)$$

where n is number of URLs

We thought another similarity measure for low traffic. First, we translate an URL to a vector with arbitrary number of elements that represent backlinks. Second, we define a similarity measure as the angle between two vectors. In this measure, although the network traffic per one request increases, the

number of requests remains *n*. But it is difficult to determine the number of dimension of the vector. If the dimension is small, grouping will be rough. If the dimensionis large, URLs that have backlinks less than the dimension are judged similarly to each other, since the vectors have a lot of elements that represent no-backlink.

- In the case of small number of backlinks, the similarity measure is often imprecise. If the number of backlinks is too small, a coincidental co-occurrence of hyperlinks often occurs. This problem made us determine following two strategies. One is that we could construct Web communities without using search engines, however we had to use search engines, since we could not get enough number of backlinks without using search engines. Therefore, we had to use search engines. However, when a search engine returned small number of backlinks to the system, we encountered this problem. Therefore, it is a second decision that we adopted some threshold of the number of backlinks in order to avoid this problem.

Although definition 1 had some problems as we explained so far, we adopted definition 1 to construct Web communities.

3.3 Construction of Web Communities of Personal Interests

Let us begin with an explanation of our method by defining two terms. One is "site" that is a group of URLs. The other is "type" of hyperlink.

Definition 2. A *site* is a set of Web pages in which we distinguish one root page and the other pages. The character sequence of the other pages' URL must start with the character sequence of the root page's URL.

For example, site "A" is created with the root page that is "http://www.jaist.ac.jp/~t-nakada/". Next, if our method finds the page http://www.jaist.ac.jp/~t-nakada/myself.html, the page will belong to site "A". On the other hand, if our method finds the http://www.yahoo.co.jp/ page, the page will not belong to site "A". The following figure illustrates its example.

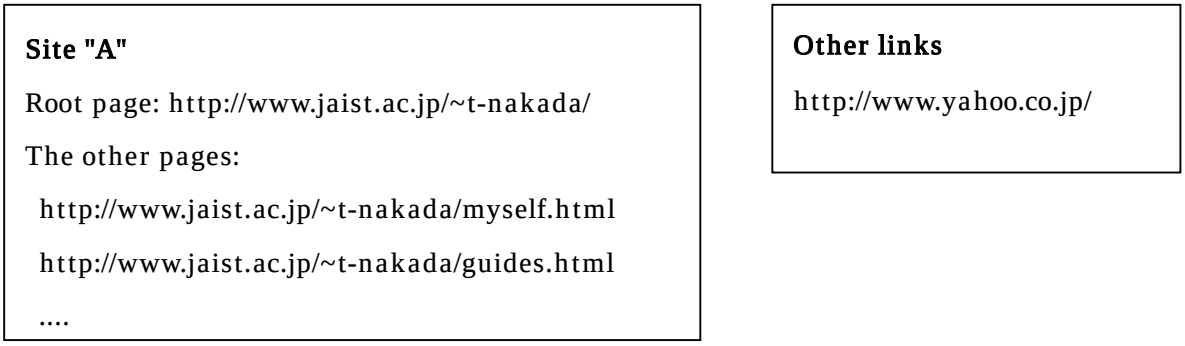


Figure 3.4 Definition of "site"

Definition 3. An *in-link* is a hyperlink that links to the same site's page. A *out-link* is a hyperlink that links to different site's pages.

An in-link is usually used to navigate visitors in a site. We would like to focus on out-links, because in-links does not imply some interests.

Our method gathers Web pages from a specific domain. Our method regards personal homepage as a site defined by definition 2, and then extracts out-links. Next our method translates an URL into another that contains only a protocol and a hostname. For example, our method translates "http://www.jaist.ac.jp/~t-nakada/index.html" into "http://www.jaist.ac.jp/". This is due to the function of a search engines. Most search engines without Google return a result list that contains an input character sequence in its URL of link. For example, when search engine searches with the input character sequence "http://www.jaist.ac.jp/", its result contains the page that links to "http://www.jaist.ac.jp/~t-nakada/". We want to distinguish "http://www.jaist.ac.jp/~t-nakada/" from

"http://www.jaist.ac.jp/", because they are not the same sites. The function of a search engine did not permit us to distinguish the pages that link to those URLs. Although we cannot solve this problem, we adopted out-links gathered by the same hostname at least in order to avoid unfair judgment between two previous URLs. Unfair judgment means that the page that links to "http://www.jaist.ac.jp/~t-nakada/" judges the page that links to "http://www.jaist.ac.jp/". In case of Google, the problem does not appear, but Google does not permit us to send automated queries [Google a].

We would like to propose the score for sorting out-links.

Definition 4. *OScore* of an URLi with respect to its out-links is defined as

$$OScore(i) = \frac{\sum_{j=1}^{n}\left(\frac{1}{out-link(j)}\right)}{backlink(i)} \tag{3.3}$$

where *OScore(i)* denotes the score of URLi, *out-link(j)* denotes the number of out-links in a personal homepage, n is the number of personal homepages, and *backlink(i)* is the number of pages that link to URLi in the whole Internet.

The numerator of OScore represents the sum of normalized score of each personal homepage that links to its out-link. It is to say that how many pages that link to URLi and that do not depend on the size of personal homepages. On the other hand, the denominator of OScore represents how well-known URLi is. Therefore, OScore gives high score if the URL is a well-known one in a specific domain and not in general. The maximum value of OScore is 1.0, but OScore usually ranges from 1×10^{-10} to 0.01. We would like to sort out-links by using the OScore.

We would like to propose a definition of Web community here.

Definition 5. A *Web community* of personal interests is a set of URLs in which one URL is selected as a seed of the Web community and other URLs that have some similarity to the seed.

The purpose of our study is to investigate what is major interest of people in a specific domain. Therefore, we would like to try to extract a Web community that has many backlinks in the specific domain, since such a Web community serves to understand outline of interests. We developed an algorithm for constructing such Web communities. The algorithm is based on the similarity measure in definition 1.

Before we show the algorithm, we would like to show four parameters in the algorithm.

Table 3.1 Four parameters in our algorithm

Name	Description
α: threshold for deleting an URL from out-links	The URLs that have the number of backlinks smaller than this threshold will be removed, because precision of the similarity of such URLs is often low.
β: maximum of a Web communities to be found	When you set β, the method creates at most β of Web communities.
γ: threshold for a seed of a Web community	If an URL has the number of backlinks > γ, it can be used as a seed of a Web community.
σ: threshold for similarity measure	If Similarity(URLi,URLj) > σ, the algorithm will judge that these two URLs are similar.

Table 3.2 Algorithm for Constructing Web communities

```
GET out-links from personal homepages in a specific domain
GET the number of backlinks by using a search engine
REMOVE URLs that have the number of backlinks smaller than α
WHILE the number of created communities < β
    SET URL_i as a seed of a Web community to the first URL in out-links that has
    linked sites more than γ
    REMOVE URL_i from out-links
    CREATE community_i and ADD URL_i to it
    FOR j=1 to end of out-links
        SET URL_j to out-links(j)
        IF Similarity(URL_i,URL_j) is greater than σ THEN
            ADD URL_j to community_i
            REMOVE URL_j from out-links
        ENDIF
    ENDFOR
ENDWHILE
```

The algorithm given in Table 3.2 first gets a URLi as a seed of a Web community from the list of out-links that is sorted by OScore. Next, it gathers URLs that are similar to URLi from out-links. A URL that is already the member of one of created Web communities will not be used in the process lately. This process is iterated until the number of Web communities is greater than the β—the threshold for maximum of Web communities to be found—or there are no Web communities in out-links.

The following example explains the algorithm.

Table 3.3 Example data of out-links

Rank of OScore	URL	Similarity to URL2	Similarity to URL3	Similarity to URL4
1	URL1	0.2	0.05	0.02
2	URL2	-	0.3	0.02
3	URL3	-	-	0.04
4	URL4	-	-	-

Let us begin to examine the example by explaining Table 3.3. The first column of the table shows ranking numbers based on OScore. The second column shows URLs that is one of out-links. The last three columns show value of Similarity(URLi,URLj) between its URL and the other respectively.

First, the algorithm gets URL1 from Table 3.1 and checks similarity to the other URL (URL2,URL3,URL4). If the threshold for similarity σ is 0.1, URL1 and URL2 will forms a Web community, since similarities of URL1-URL3 and URL1-URL4 are smaller than the threshold.

Second, the algorithm gets URL3 from Table 3.1, since URL2 is already removed from the list for constructing the previous Web community. The algorithm constructs Web community whose member is only URL3, because similarity between URL3 and URL4 is smaller than the threshold.

Finally, the algorithm constructs Web community whose member is only URL4. The following is a result list of Web communities.

Table 3.4 A result list of Web communities

Web Communities
URL1, URL2
URL3
URL4

We should point out one problem in the algorithm. When we explain the above example, the problem of the algorithm is to ignore the similarity between URL2 and URL3. The factor affecting it could be the following: before the algorithm found the similarity between URL2 and URL3, the algorithm already constructed Web community with URL2, then URL2 was deleted in out-links. If we want to avoid the problem, we have to calculate the similarity of all of each pairs. It means we have to encounter the computational problem such as equation (3.2). We would like to show the number of request to the search engine, and compare it with equation (3.2).

$$n + (n-1) + (n-2-nm) + (n-3-2nm) + ... + (n-nc-(nc-1)nm) \qquad (3.4)$$
$$= (nc+1)n - \sum_{i=1}^{nc} (i + i \times nm - nm)$$

where n :the number of out-links

nc : the number of communities

nm : the average of the number of URLs in a community

The following is the calculation for the number of request to search engine using experimental data that will be shown later.

Information about experimental data

Specific Domain: JAIST school of knowledge science

The number of personal homepages: 294

The number of out-links: 5426

Parameters

α : threshold for deleting an URL from out-links: 10

β : maximum of a Web communities to be found: 50

γ : threshold for seed of a Web community: 2

σ : threshold for similarity measure: 0.07

The number of out-links that have more than α linked site: 3188

Number of request according to equation (3.2) ($request = n + \binom{n}{2}$)

Request = 3188+ ${}_{3188}C_2$ = 5,083,266

Number of request according to equation (3.4) ($(nc+1)n - \sum_{i=1}^{nc}(i + i\alpha - \alpha)$)

Request = (50+1)$\times$3188- $\sum_{i=1}^{50}(i + i \times 1.7 - 1.7)$ 159,231

(The experiment that we will explain later made us assume that α is 1.7)

We roughly assumed the time of one request is about 0.06 on the ground of experiments. Accordingly the following are the times for accessing search engine.

Case equation (3.2) 5,083,266$\times$0.06[s] 3.5 days

Case equation (3.4) 159,231$\times$0.06[s] 2.7 hours

Although it is roughly calculation, we could get the time for constructing Web communities. We thought 3.5 days is too long. Moreover, equation (3.2) increase exponentially, if n that is the number of out-links increase. In summary, our algorithm considered computational cost rather than precision.

Our algorithm often outputs a Web community that has only one member. But someone may argue that such Web community is not a kind of community. Although it may be true, we think that such Web community is needed in the light of our purpose that is to extract Web community that has many backlinks.

The following summarized figure explains the whole process of our method.

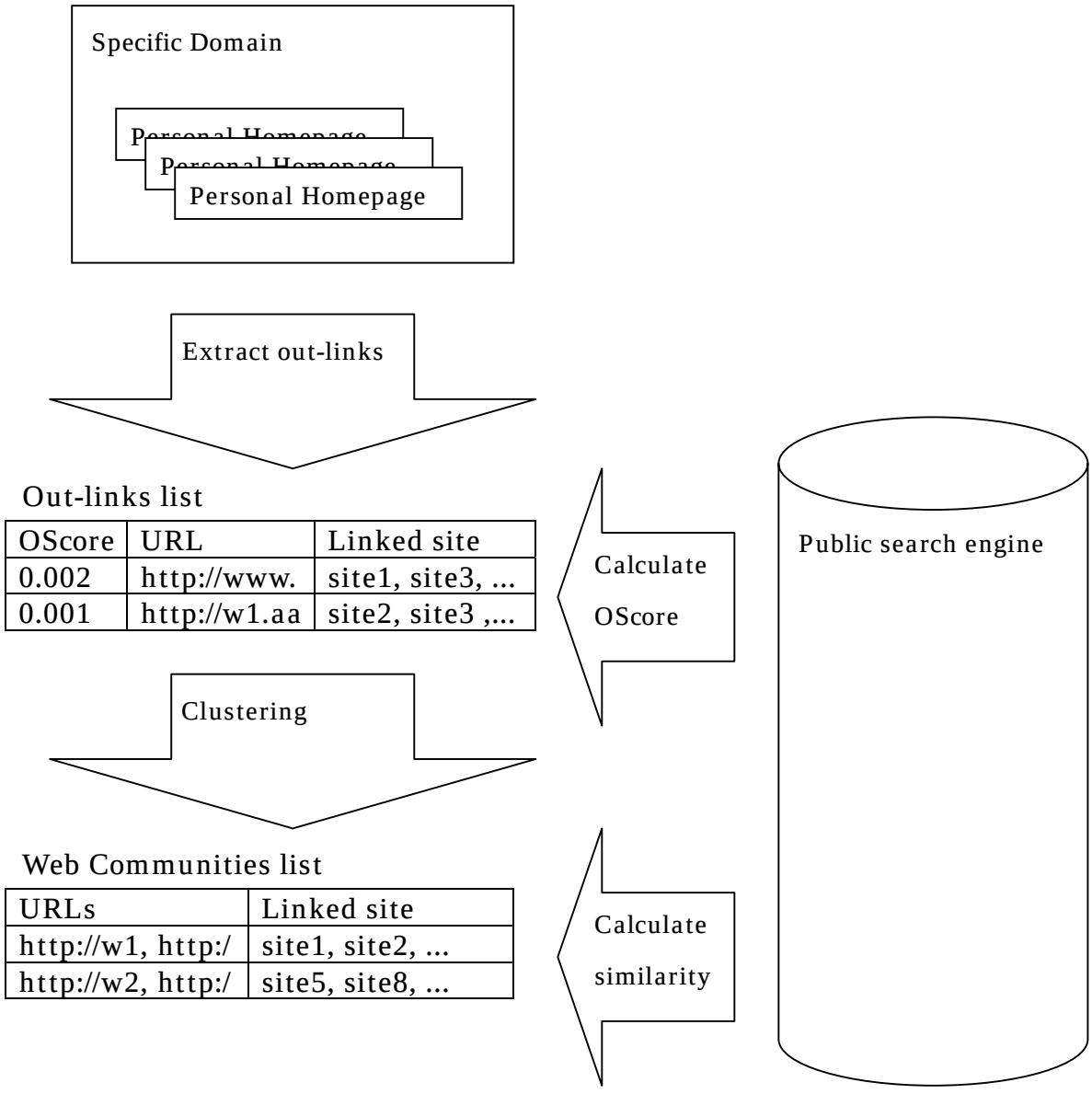


Figure 3.5 An outline of whole process of our method

3.4 Visualizations of Web Communities of Personal Interests

We developed a visualization system for exploring Web communities found by our method. The following is the outline of the system.

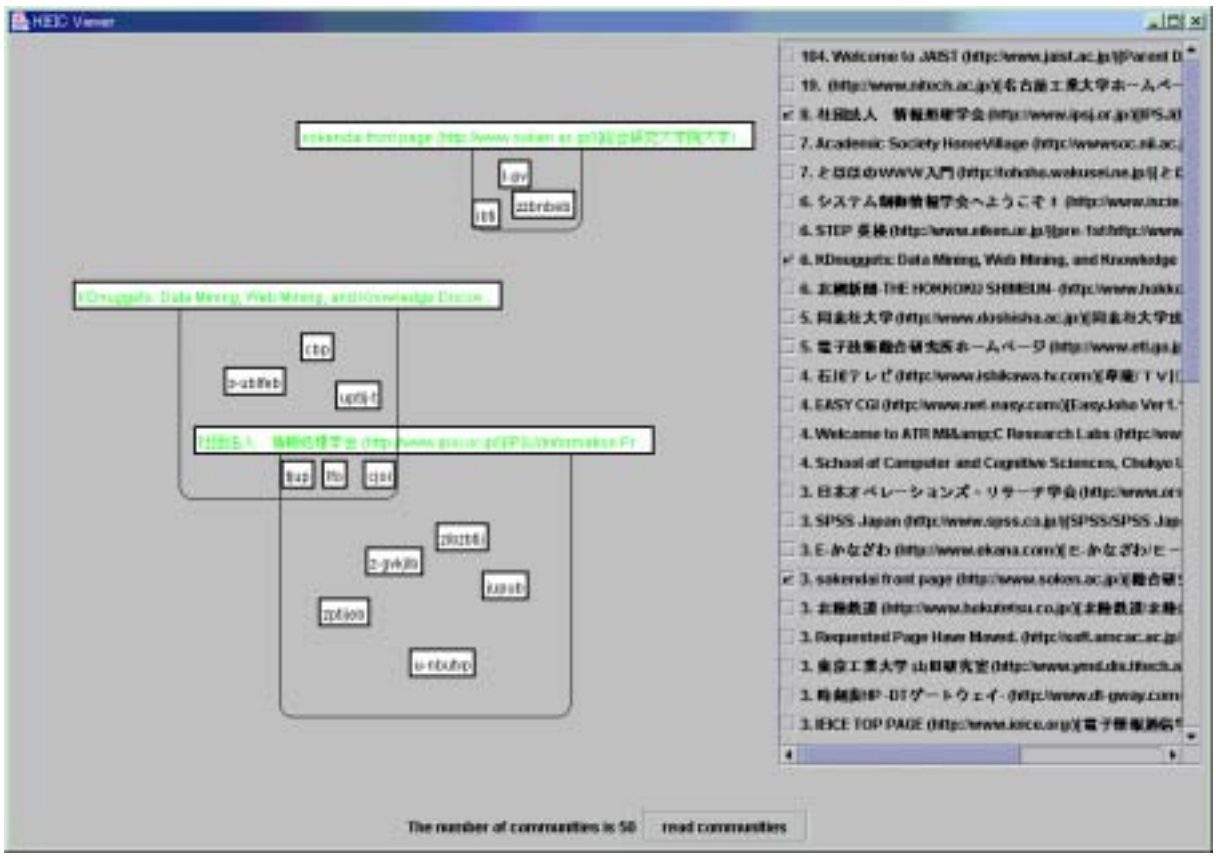


Figure 3.6 A visualization system for exploring Web communities

In Figure 3.6 there are 50 Web communities in the right panel and 3 Web communities from them in the left panel. If the user set true to the checkbox in right panel, the corresponding Web community will be shown in the left panel.

The following figure shows how a Web community is expressed in the left panel.

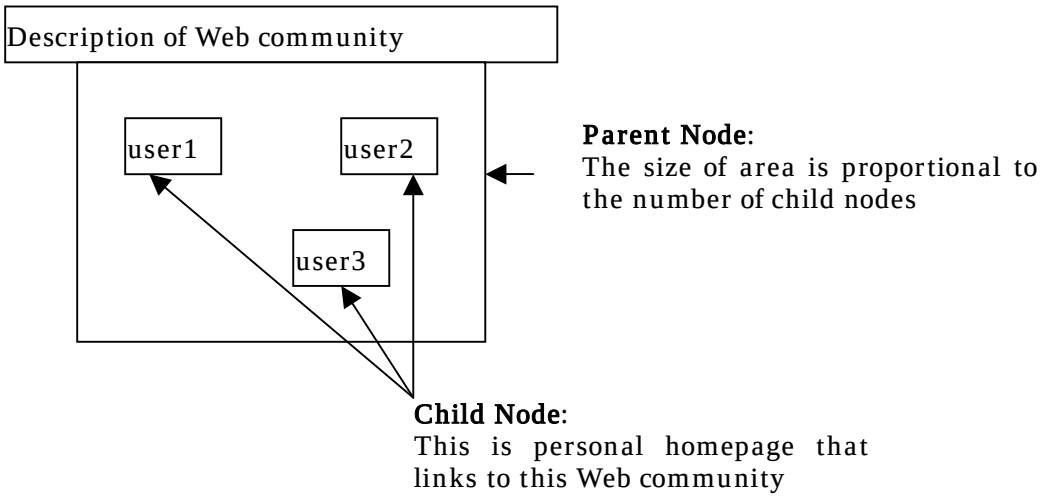


Figure 3.7 An explanation about expression of Web community

In Figure 3.7, "Description of Web community" means explanations about the Web community. The following is its format.

Title of URL1 (URL1) [The letter of guide 11/ The letter of guide 12/...]
Title of URL2 (URL2) [The letter of guide 21/ The letter of guide 22/...]
...
Title of URLn (URLn) [The letter of guide n1/ The letter of guide n2/...]

One line explains one URL in a Web community, this line consists of title of its page, its URL, and the letter of guide that is a character sequence and if you click on it, you will move to new page indicated by its URL.

We would like to explain in more detail with the following figure.

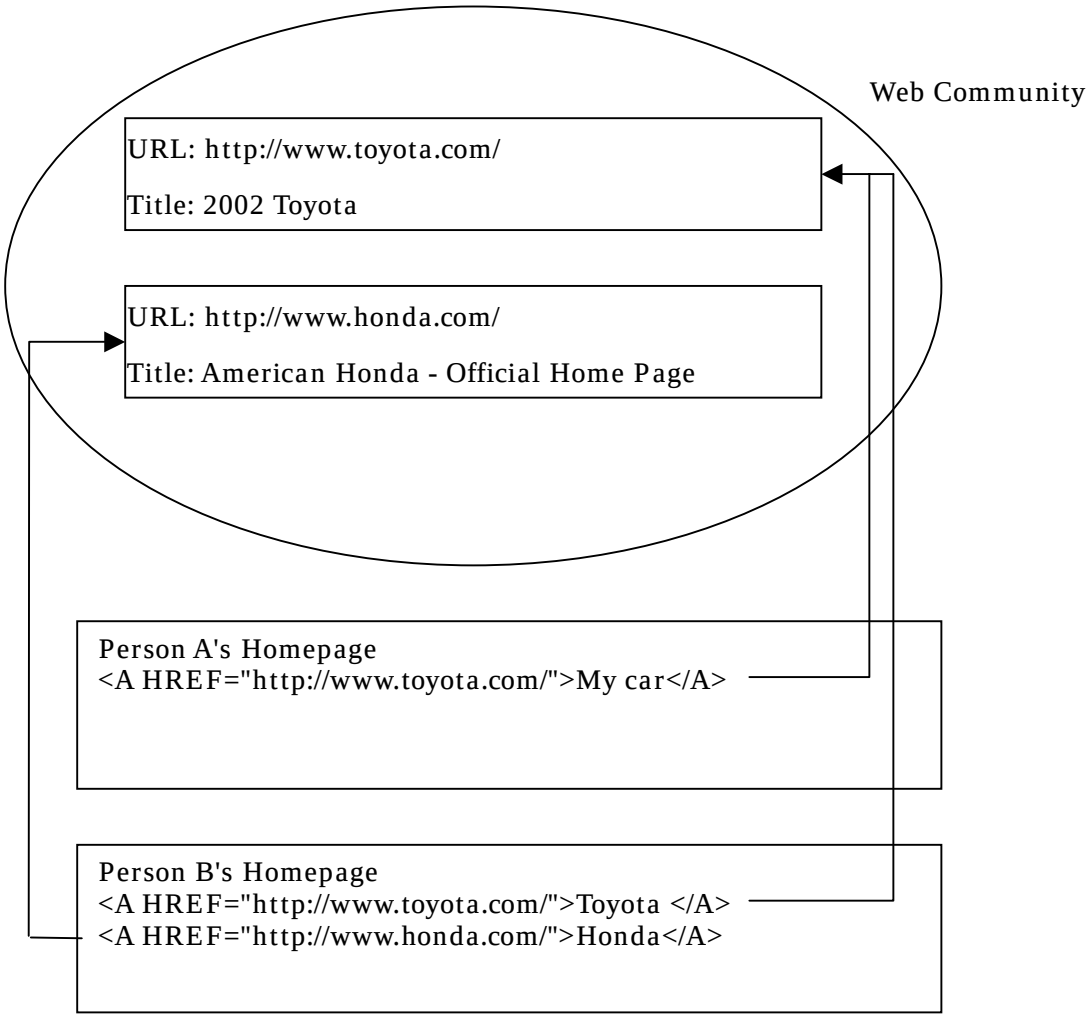


Figure 3.8 An example of "Description of Web community"

There are two personal homepages and one Web community that is linked by them. The following is "Description of Web community" of its Web community.

2002 Toyota (http://www.toyota.com/) [My Car/Toyota]
American Honda - Official Home Page (http://www.honda.com/) [Honda]

If the length of a character sequence of "Description of Web community" is larger than 60, the character sequence is cut and added "..." to. However, you can see the character sequence of "Description of Web community" in detail, if you put mouse cursor on the Web community (Figure 3.9).

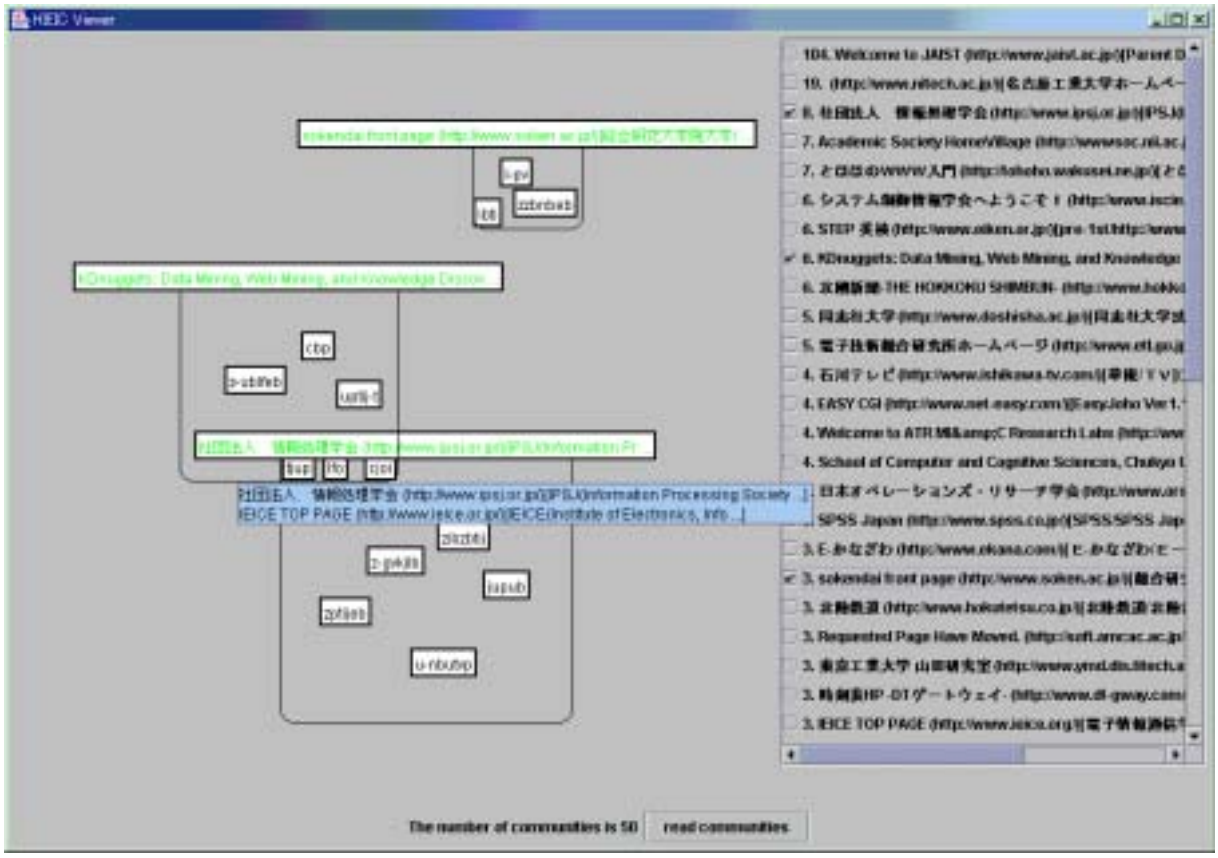


Figure 3.9 Detailed "Description of Web community"

This visualization is based on spring model [Eades 84], [Sugiyama 95] that is a well-known technique for drawing general undirected graphs. In this model, vertices are replaced with steel rings and each edge with a sprint to form a mechanical system, and repulsive and attractive forces among rings. Then the rings are placed in some initial layout and moved iteratively according to the forces so that the system reaches a minimal energy state. Two types of force are given by Eadges's model.

(1) Fs: attractive or repulsive forces exerted by the springs between neighbors

$$Fs = Cs \log(d / k)$$

(2) Fr: repulsive forces between every pair of non-neighboring vertices

$$Fr = -Cr 1/d^2$$

where d is the distance between a pair of vertices, k is an ideal distance between

neighbors, Cs, Cr >0 are parameters for tuning the model.

We used only Fs and changed into next formula, since we think computational cost is more importance than it drawing precision in our visualization.
(1)' $Fs = Cs(d - k)$ where if d > k then Fs=0

The following are three relations that are acted by Fs(1)'.
① Among all parent nodes that express Web communities
② Among all child nodes, which express Personal Homepage, in the same parent node
③ Between top parent node and its child nodes, where *Fs* places on only y-axis.

These forces are based on the following principles. Web communities should not be drawn too close to each other ①. Personal Homepage in the one Web community should not be drawn too close to each other ②. Since explanation of Web Community is drawn in the top of Web community node, Personal homepage should not be drawn too close to the top of Web community node ③.

Moreover, there is a restriction between a parent node and its child nodes that child node must be inside its parent node; there are two interactive forces from/to parent node to/from child node. The following is an explanation about it.

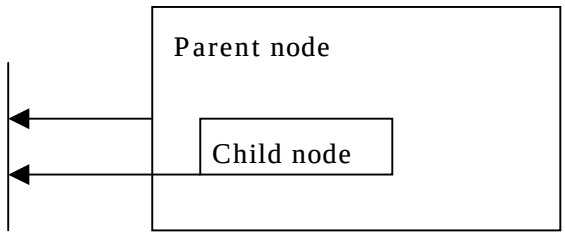


Figure 3.10 Relation between parent and child node

In case that the force placed on the child node is too big that it would end up with getting out of the parent node, the minimum force required to keep child node

inside the parent node is placed on the parent node as well, and vice versa.

The following are the other functions of our visualization system where one can:

1. move nodes using a mouse (Figure 3.11)
2. fix node positions (Figure 3.12)
3. fire up a browser displaying URLs making up a Web community (Figure 3.13)

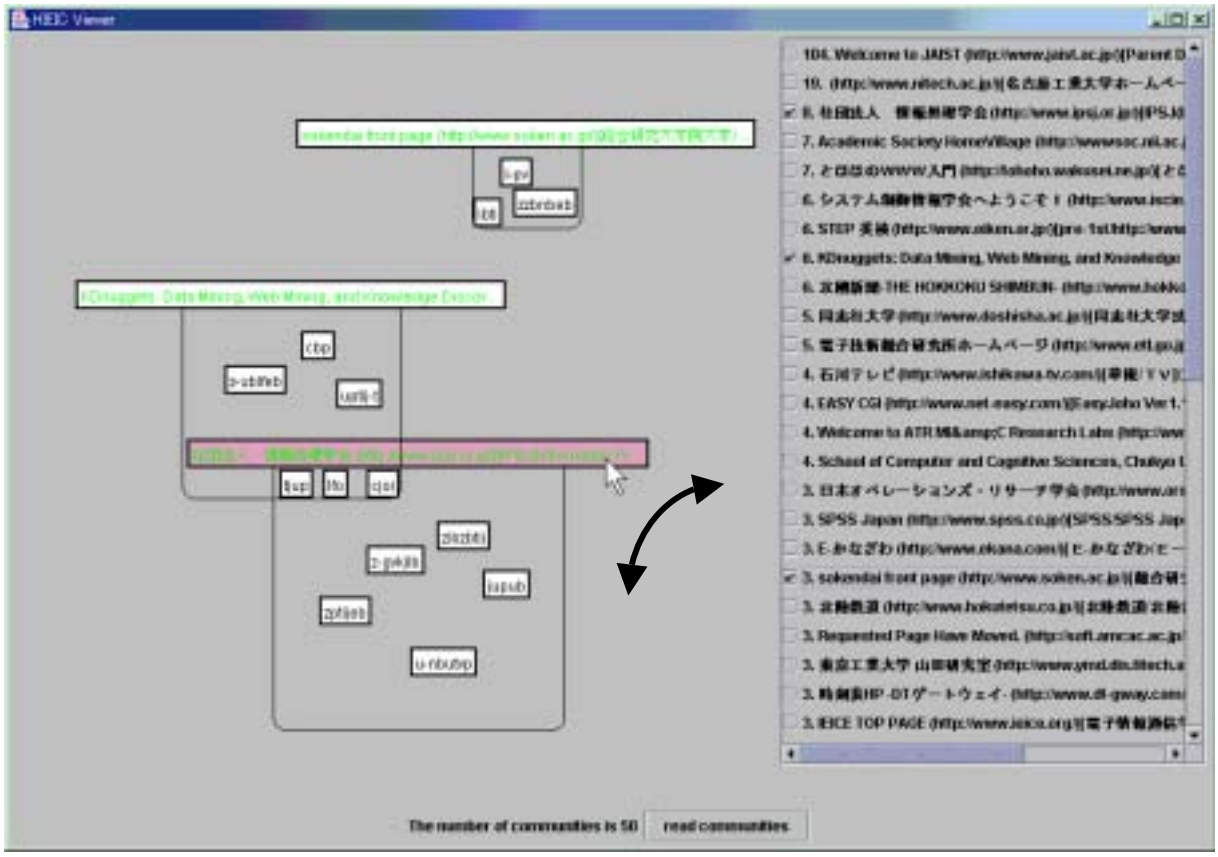


Figure 3.11 Moving nodes using a mouse

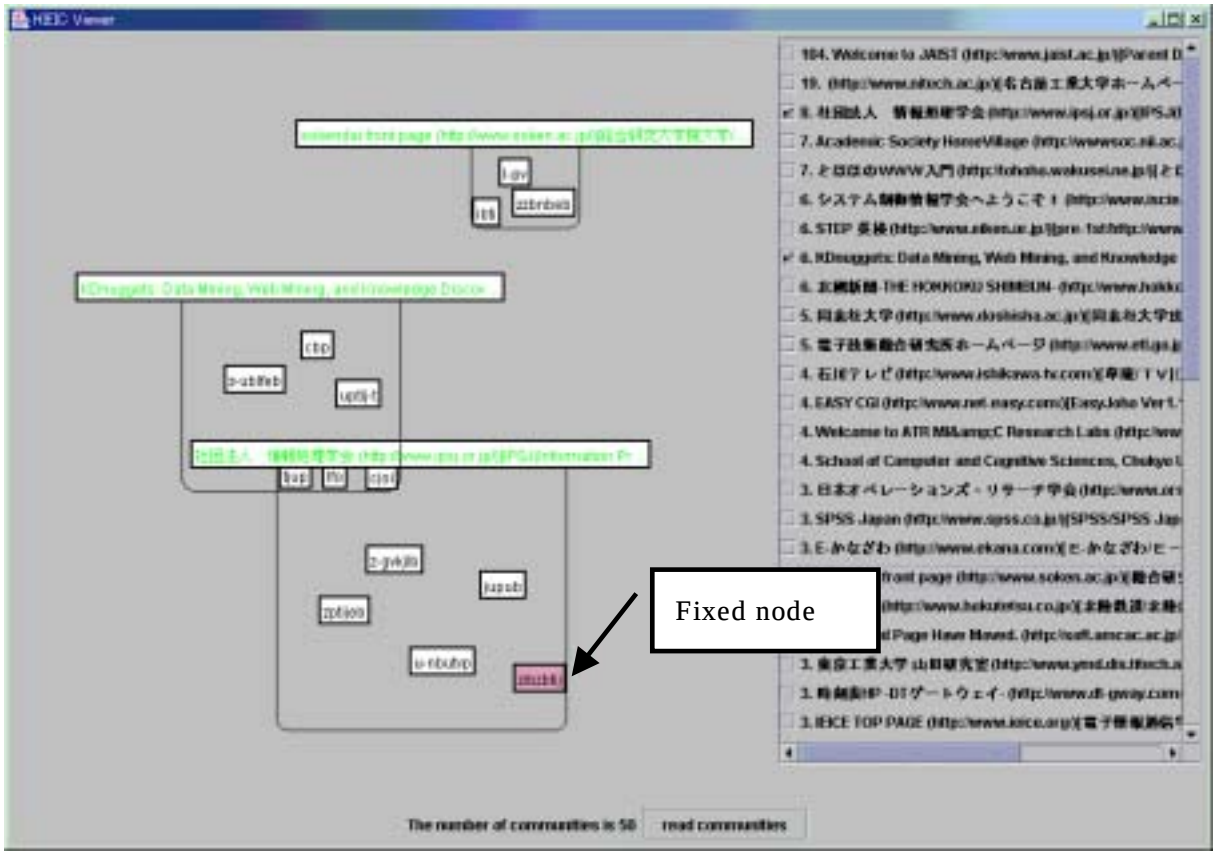


Figure 3.12 Fixing node positions

Figure 3.13 Firing up a browser displaying URLs making up a Web community

Chapter 4

Implementation

This chapter describes how to implement our method.

4.1 Basic Design

We implemented our method using Java language whose version is 1.4.0 beta3. The reason why we selected Java is that Java has rich libraries such as HTML parser and we can easily write multi-thread programs. And the reason why we used version 1.4.0 beta3 is that library about XML is appended and the bug about read function in class HTMLEditorKit was corrected from this version. The major problem in Java is its slow calculation speed. However, the bottleneck of our method is how to receive response from a search engine. When a program is waiting the response, it can do other jobs. Therefore, this problem is not so important for our method.

One important issue in data mining is scalability. In other words, a good data mining algorithm should have ability to deal with a huge amount of data. Although we recognize the importance of scalability, we did not have to consider it in this research, because our program has enough virtual memory to function.

4.2 System Architecture

The following is a figure that shows relations of classes (that is a word defined in Java language) in our implementation.

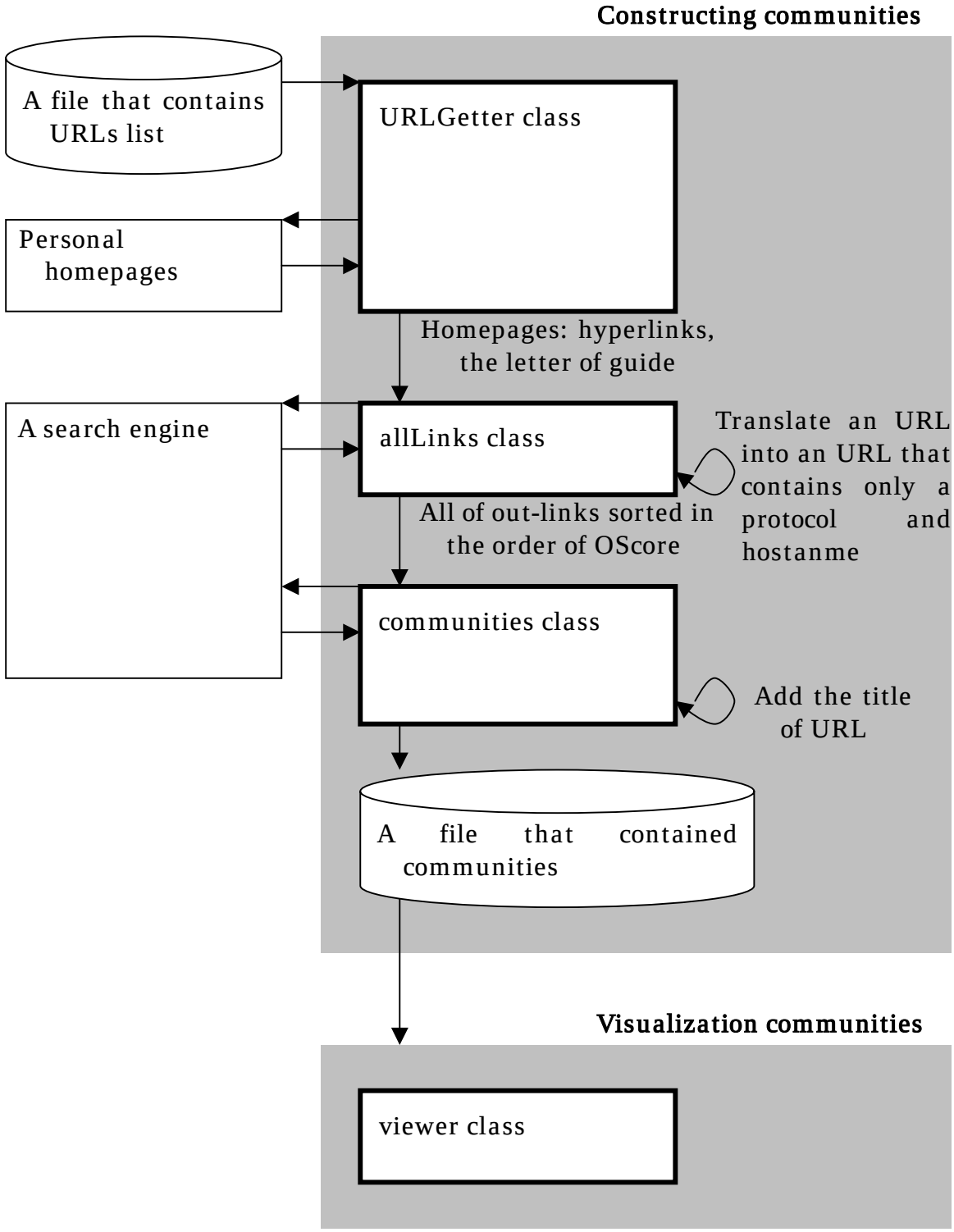


Figure 4.1 Relations of classes

The bold rectangle represents class that is a word defined in Java language, the

thin rectangle represents an outside data in system, the column represents file system, and arrows represent data flows.

The following table shows the main jobs of each class.

Table 4.1 The main jobs of each class

Class	Input	Output	Task
URLGetter	An personal homepage's list	A hyperlinks and the letter of guide	1.Get HTML data from personal homepages 2. Extract hyperlinks and letters of guide from 1.
allLinks	URLs of personal homepages, its out-links, and its letters of guide	A list of all out-links sorted in the order of OScore	1. Translate an URL into an URL that contains only a protocol and a hostanme 2. Get number of backlinks by using a search engine 3. Sort out-links in the order of OScore
communities	A list of all out-links sorted in the order of OScore	Web communities of personal interests	1. Construct Web communities 2. Get the title of URLs
viewer	Web communities of personal interests	A visualization	1. Represent Web communities by an undirected graph

When our program gets a hyperlink from personal homepages, it creates some threads in order to promote the efficiency, since personal homepages are not always located in the same HTTP server. However, when our program gets number of backlinks by using a search engine, it takes a rest every one request, since we considered a heavy load to a search engine must not be given.

Chapter 5

Experiment and Evaluation

This chapter describes experiments and evaluations of our method.

5.1 Fundamental Experiment for Similarity

Before we perform main experiments, we need to determine the value of σ: a threshold for similarity measure. Because σ is independent of the specific domain that we want to investigate, we could set σ before main experiments, if we determine which a search engine is chosen.

We selected AltaVista as a search engine and Japan Advanced Institute of Science and Technology (JAIST) as a specific domain. We carried out the experiment as follows:

1. We gathered out-links in the domain
2. We constructed all pair of out-links, where a pair of URLs that indicate the same page was not used.
3. We evaluated 191 obtained pairs using three criteria: "very similar", "similar", and "not similar".

Figure 5.1 shows the result.

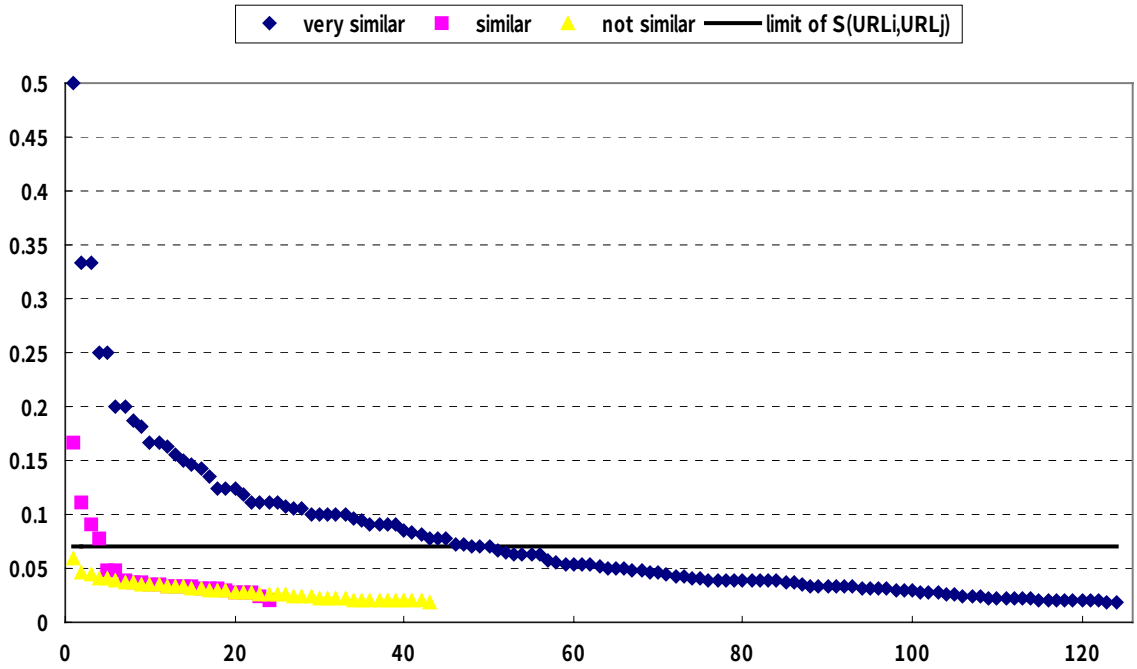


Figure 5.1 The result of the experiment for similarity

In Figure 5.1, x-axis means the number of data, and y-axis means the value of Similarity(URLi,URLj) (Definition 1). The figure shows one horizontal line and three types of data that represent the criterion respectively and sorted in the order of the value of Similarity(URLi,URLj). The problem of determining the threshold σ is one of determining the position of the horizontal line in Figure 5.1. Consequently, the horizontal line in Figure 5.1, the value is 0.07, is our decision. Therefore, we had to drop data more than a half that represents "very similar" criterion in order to drop data that represents "not similar" criterion.

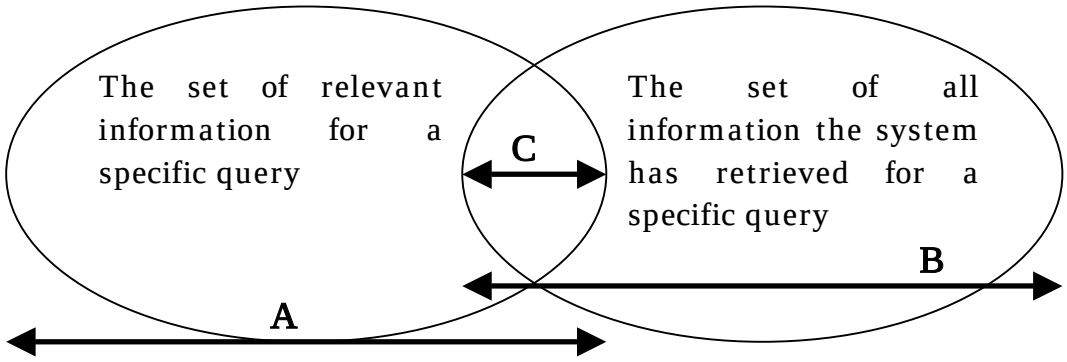
5.2 Basic Idea of Experiment and Evaluation

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?

The traditional method for evaluating information retrieval is to evaluate precision and recall. The following figure explains them.



precision: C / B

recall: C / A

Figure 5.2 An explanation of precision and recall

We are not concerned in this paper with evaluation of recall, because it is beyond the scope of this paper to know the size of "A" in Figure 5.2 that means all of Web communities in the specific domain in our case.

We used the questionnaire to evaluate the result obtained by our method, since the evaluation is subjective; Precision of a Web community and judgment on interests are due to personal background knowledge.

5.3 Experiment in JAIST domain

Following experiments were made:

1. Our method constructs Web communities from three specific domains that are JAIST (Japan Advanced Institute of Science and Technology) school of knowledge science, school of information science, and school of material science.
2. We use the questionnaire to evaluate obtained Web communities.

The period of the experiment was from January 23 to 24, 2002. The following table shows four parameters in our method.

Table 5.1 Four parameters in the experiment in JAIST

α: threshold for deleting an URL from out-links	smaller than 9 backlinks
β: maximum of a Web communities to be found	at most 50 communities
γ: threshold for seed of a Web community	larger than 1
σ: threshold for similarity measure	0.07

The following table shows the outline of the result.

Table 5.2 Result of evaluation

	Knowledge science	Information science	Material science
Number of personal homepages we got	294	388	411
Number of out-links	5426	3499	923
Number of communities	50	50	44

We set the parameter to 50 for all three domains. In school of material science, however, since we ran out of URL's to which more than two people have links, we ended up with 44 Web communities when the program finished.

Before we present the result we got, we would like to show the format of table containing the result. One line shows one Web community. The first column has

numbers of people having links to the Web community. The second column shows description of the Web community. The following is an example of its description.

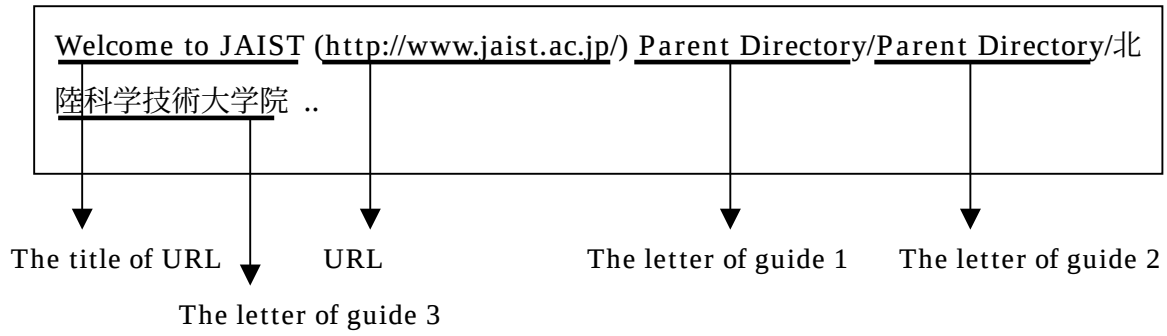


Figure 5.3 An example of description of a Web community

The letter of guide is a character sequence and if you click on it, you will move to new page indicated by its URL. Although a character sequence of description of the Web community is composed of "The title of URL", "URL", and "The letter of guide", in that order, the sequence will be reduced in order to go into the display space. There are also descriptions of a Web community that have no information of the title. Of course, there are also descriptions that have several URLs that compose the Web community.

Table 5.3 The result from school of knowledge science

Number of members	Descriptions
104	Welcome to JAIST (http://www.jaist.ac.jp/) Parent Directory/Parent Directory/北陸科学技術大学院 ..
19	(http://www.nitech.ac.jp/) 名古屋工業大学ホームページ/the Nagoya Institute of Technology/Nagoya I .. 宇都宮大学 (http://www.utsunomiya-u.ac.jp/) University of utsunomiya 豊橋技術科学大学のホームページ (http://www.tut.ac.jp/) 豊橋技術科学大学/豊橋技術科学大学 大 .. 長崎大学-Nagasaki University Official Web Sute- (http://www.cc.nagasaki-u.ac.jp/) EXCEL による統計 .. Nagoya University (http://www.nagoya-u.ac.jp/) Nagoya University Fukushima University (http://www.fukushima-u.ac.jp/) 日本の国立大学総覧 大阪府立大学 (http://www.osakafu-u.ac.jp/) 大阪府立大学 宮崎医科大学公式ＨＰ (http://www.miyazaki-med.ac.jp/) Mac で作る Internet Servers Kogakuin Univ. Home Page (http://www.kogakuin.ac.jp/) 工学院大学 Fukui Prefectural University WWW Server Home Page (http://www.fpu.ac.jp/) 福井県立大学 武蔵大学 MUSASHI University (http://www.cc.musashi.ac.jp/) 参画型情報システム Kobe-u Homepage(English) (http://www.kobe-u.ac.jp/) Faculty of Science/Kobe University Yamagata University Home Page (http://www.yamagata-u.ac.jp/) 山形大学に入学/山形大学 TOYAMA UNIVERSITY Web Site (http://www.toyama-u.ac.jp/) 清家彰敏プロジェクト /小倉研究室(経 .. Osaka University (http://www.osaka-u.ac.jp/) Osaka University/大阪大学/Osaka Uni. 大阪医科大学 ホームページ (http://www.osaka-med.ac.jp/) Vine Linux 1.1 福岡工業大学 (http://www.fit.ac.jp/) the Ninth International Conference on Industrial & Engineering A .. 京都大学 (http://www.kyoto-u.ac.jp/) 京都 富山県立大学 (http://www.pu-toyama.ac.jp/) 富山県立大学 Meiji University Home Page (http://www.meiji.ac.jp/) R.Kamei's MkLinux Station 姫路工業大学ホームページ (http://www.himeji-tech.ac.jp/) 兵庫県立姫路工業大学 (http://www.kanagawa-u.ac.jp/) シェルスクリプトを書く 筑波大学 Web ページ (http://www.tsukuba.ac.jp/) 筑波大学 金沢工業大学ホームページ (http://www.kanazawa-it.ac.jp/) 金沢工業大学ライブラリーセンター/金 ..
8	社団法人 情報処理学会 (http://www.ipsj.or.jp/) IPSJ(Information Processing Society of Japan)/IPSJ .. IEICE TOP PAGE (http://www.ieice.or.jp/) IEICE(Institute of Electronics, Information, and Communicati ..
7	Academic Society HomeVillage (http://wwwsoc.nii.ac.jp/) JSSST(Japan Socirty for Software Scienc ..

7	とほほのWWW入門 (http://tohoho.wakusei.ne.jp/) とほほの perl 入門/とほほの perl 入門/とほほの per ..
6	システム制御情報学会へようこそ！ (http://www.iscie.or.jp/) システム制御情報学会(the Institute of ..
	(http://www.sice.or.jp/) SICE(Society of Instrument and Control Engineers)/(社)計測制御自動学会(th ..
6	STEP 英検 (http://www.eiken.or.jp/) pre-1st/http://www.eiken.or.jp/index1.html/英語検定事務局
	TOEIC (http://www.toeic.or.jp/) http://www.toeic.or.jp//TOEIC/TOEIC/ＴＯＥＩＣ公式サイト
	漢検ホームページ (http://www.kanken.or.jp/) http://www.kanken.or.jp/
	国際教育交換協議会 (CIEE) (http://www.cieej.or.jp/) 国際教育交換協議会 (カウンシル) 日本代表部/ ..
6	KDnuggets: Data Mining, Web Mining, and Knowledge Discovery Guide (http://www.kdnuggets.com/) ..
6	北國新聞-THE HOKKOKU SHIMBUN- (http://www.hokkoku.co.jp/) 新聞社/北國新聞/北國新聞/北國
	osakanews.com (http://www.osakanews.com/) 大阪
	北海道新聞社ホームページ (http://www.hokkaido-np.co.jp/) 北海道
	沖縄タイムス (http://www.okinawatimes.co.jp/) こちら
	四国新聞社 (http://www.shikoku-np.co.jp/) うどん屋検索サイト (四国新聞)
	琉球新報 (http://www.ryukyushimpo.co.jp/) 琉球
	東京新聞ホームページへようこそ (http://www.tokyo-np.co.jp/) 東京
	さきがけ onTheWeb - 秋田魁新報社／秋田さきがけスポーツ新聞社 (http://www.sakigake.co.jp/) 「 ..
	河北新報 THE KAHOKU SHIMPO World Wide Web (http://www.kahoku.co.jp/) 河北
5	同志社大学 (http://www.doshisha.ac.jp/) 同志社大学法学部政治学科/同志社大学商学部入学
	Osaka Gakuin University Home Page (http://www.osaka-gu.ac.jp/) 大阪学院大･情報学･中川研究室 ..
	大阪工業大学 (http://www.oit.ac.jp/) 文字認識を用いたメッセージ入力システム
	Welcome.html (http://www.osakac.ac.jp/) ますい くんのページ/大阪の郊外にあるこじんま りした大 ..
5	電子技術総合研究所ホームページ (http://www.etl.go.jp/) 第 43 回音楽情報科学研究会(SIGMUS)/電子 ..
4	石川テレビ (http://www.ishikawa-tv.com/) 草庵/ＴＶ
	ようこそ！北陸朝日放送へ (http://www.hab.co.jp/) 「わくわく動物王国」/コイツ
	FM ISHIKAWA Official Site (http://www.fmishikawa.co.jp/) Ｒａｄｉｏ /FM ISHIKAWA Official Site
4	EASY CGI (http://www.net-easy.com/) EasyJoho Ver1.1/EASY CGI/EAGY CGI/EASY CGI
4	Welcome to ATR MI&C Research Labs (http://www.mic.atr.co.jp/) the Third International Confe ..
4	School of Computer and Cognitive Sciences, Chukyo University (Japanese) (http://www.sccs.chuk ..
3	日本オペレーションズ･リサーチ学会 (http://www.orsj.or.jp/) 日本 OR 学会/日本オペレーションズ ..
3	SPSS Japan (http://www.spss.co.jp/) SPSS/SPSS Japan/SPSS Japan 公式サイト
3	E-かなざわ (http://www.ekana.com/) Ｅ-かなざわ/Ｅ－金沢

3	sokendai front page (http://www.soken.ac.jp/) 総合研究大学院大学/総合研 究大学院大学 湘南工科大学 (http://www.shonan-it.ac.jp/) 湘南工科大学/材料学科
3	北陸鉄道 (http://www.hokutetsu.co.jp/) 北陸鉄道/北陸鉄道石川線/金沢近郊バス時刻表
3	Requested Page Have Moved. (http://soft.amcac.ac.jp/) アジア･ファジィシステムシンポジウム/日 ..
3	東京工業大学 山田研究室 (http://www.ymd.dis.titech.ac.jp/) 東京工業大･総合理工学･山田研究室/AI ..
3	時刻表 HP -DT ゲートウェイ- (http://www.dt-gway.com/) DT ゲートウェイ～石川の鉄道･空路･高速 ..
3	IEICE TOP PAGE (http://www.ieice.org/) 電子情報通信学会/IEICE(the Institute of Electronics, Informa ..
3	(http://dirr.nacsis.ac.jp/) NACSIS-DiRR/研究活動資源ディレクトリ/NACSIS-DiRR
2	Well come to Tawara Lab !! (http://www.tawara.ie.musashi-tech.ac.jp/) こちら/[URL]/宇野さん
2	(http://www2.mic.atr.co.jp/) ComicDiary:/Representing Individual Experiences through Comic/ATR, ..
2	ぅ (http://ifdef.udn.ne.jp/) ぅ/ぅ
2	独立行政法人 経済産業研究所 (RIETI) (http://www.rieti.go.jp/) 独立行政法人･経済産業研究所 (RIE ..
2	grips (http://www.grips.ac.jp/) 政策研究大学院大学/ＧＲＩＰＳ
2	オオサワグループホームページ (http://www.ohsawagroup.co.jp/) アートツーリスト/arttouridt
2	経営情報系 (ホーム) (http://kjs.nagaokaut.ac.jp/) 長岡技術科学大学経営情報系三上研究室/長岡技 ..
2	**王様の本ホームページ!** (http://www.samahon.co.jp/) 王様の本/王様の本 うつのみや (http://www.utsunomiya.co.jp/) うつのみや書店
2	(http://eco21.nikkeihome.co.jp/) 日経 ECO21/ECO21
2	Welcome to JASESS (http://jasess.u-shizuoka-ken.ac.jp/) 社会･経済システム学会/社会･経済シス ..
2	fcrc TOP PAGE (http://www.fcrc.titech.ac.jp/) 東京工業大学フロンティア創造共同研究センター/Ｆ .. Nubic Web Top page (http://www.nubic.adm.nihon-u.ac.jp/) 日本大学国際産業･ビジネス育成センタ .. (株) 筑波リエゾン研究所 (http://www.tliaison.com/) 筑波リエゾン研究所 New Industry Creation Hatchery Center (http://www.niche.tohoku.ac.jp/) 未来科学技術共同研究セン .. (http://www.aee.u-tokyo.ac.jp/) ＲＣＡＥＥ
2	(http://www.dee21.com/) 経営情報学会(the JApan Society for Management INformation)/経営情報 ..
2	TARA, Univ. of Tsukuba, Japan (http://www.tara.tsukuba.ac.jp/) -/筑波大学先端学際領域研究センター
2	(http://www2.nomura.co.jp/) ●/Virtual Stock Investment Club
2	Google (http://www.google.co.jp/) google(日本語)/google
2	HGC homepage (http://www.hgc.ims.u-tokyo.ac.jp/) Link to SIGMBI (Molecular Biology Informatics)/ ..
2	Welcome to 代々木ゼミナール (予備校) ホームページ (http://www.yozemi.ac.jp/) 代々木ゼミナー ..

2	Welcome to Yokoya-lab. HomePage (http://yindy1.aist-nara.ac.jp/) 竹村 治雄(TAKEMURA, Haruo)/3 次 ..
2	PADI JAPAN'S HOME PAGE (http://www.padi.co.jp/) PADI japan/project A.W.A.R.E.(Aquatic World Awa ..
2	Cora Research Paper Search (http://cora.whizbang.com/) Cora/Cora Search
2	Hoshi Lab. at Tokai univ. Home Page (http://www.fb.u-tokai.ac.jp/) 3.文理シナジー学会会員/文理 シナ ..
2	Osaka-Univ. ICS Home Page (http://www.ics.es.osaka-u.ac.jp/) INN FAQ in Japanese/Ph.D./Informatio ..
2	Test (http://www2.startshop.co.jp/) unix の部屋/ネットワークプログラミングの基礎知識
2	Home of Lab. for Information Synthesis (http://www.islab.brain.riken.go.jp/) 情報創成システム研究チ ..
2	CGI Pocket (http://pocket.727.net/) CGI Pocket/CGI Pocket

Table 5.4 The result from school of information science

Number of members	Descriptions
86	Welcome to JAIST (http://www.jaist.ac.jp/) 変人の部屋//Parent Directory/Japan Advanced Institute o ..
14	Ochimizu Lab. Home-Page (http://ochimizu-www.jaist.ac.jp/) 落水研究室のメンバー リスト/落水研究 .. Tojo Lab. (http://cirrus.jaist.ac.jp:8080/) 東条先生 (http://fish.jaist.ac.jp:8080/) 就職活動やってますか？ Shinoda Lab., Chair of Software Engineering, JAIST (http://shinoda-www.jaist.ac.jp/) Shinoda Labora .. Acoustic Information Science Laboratory (http://gelgoog.jaist.ac.jp:8000/) the Acoustic Information ..
10	IEICE TOP PAGE (http://www.ieice.or.jp/) The Institute of Electronics, Information and Communicatio .. 社団法人 情報処理学会 (http://www.ipsj.or.jp/) Vol. 40, No. SIG3(TOD1)/Vol. 40, No. SIG3(TOD1)/IP ..
9	JAIST ソフトウェア基礎講座 (http://kt-www.jaist.ac.jp:8000/) http://kt-www.jaist.ac.jp:8000/~toshiak .. Ruby programming: source code samples, examples, fragments, classes, methods and modules. .. (http://www.ale.cx/) Ruby Mine Nakamura Lab Web Page (http://easter.kuee.kyoto-u.ac.jp/) ruby K.Hiwada. (http://www.ruby.ch/) Ruby.CHannel
6	Tokyo Institute of Technology (http://www.titech.ac.jp/) Tokyo Institute of Technology/東京工業大学/ .. 京都大学 (http://www.kyoto-u.ac.jp/) Kyoto University the University of Tokyo (http://www.u-tokyo.ac.jp/) University of Tokyo/東京大学 Kobe-u Homepage(English) (http://www.kobe-u.ac.jp/) Kobe University 宮崎大学<Miyazaki Univ.> (http://www.miyazaki-u.ac.jp/) 宮崎大学のホームページ/工学 部

5	Horiguchi-Abe lab. (http://mitsuko.jaist.ac.jp/) Horiguchi-Abe? Lab)/堀口･阿部研究室/マルチメデ ..
4	Image Laboratory Home Page (http://awabi.jaist.ac.jp:8000/) Miyahara Lab.'s page/Kotani Laborator .. Okada Laboratory Web Page (http://www.media.cs.chubu.ac.jp/) 論文の書き方･注意点 Sato Laboratory Home Page (http://hilbert.elcom.nitech.ac.jp/) 名古屋工業大学 佐藤･佐藤研究室 (http://isg.ap.eng.osaka-u.ac.jp/) ISG
4	LDL home page (http://www.ldl.jaist.ac.jp/) Standard ML Local Guide (JAIST)/http://www.ldl.jaist.ac.jp ..
4	Home Page of Sagayama & Shimodaira Lab, JAIST (http://www-ks.jaist.ac.jp/) JAIST 知能情報処理学 ..
4	EASY CGI (http://www.net-easy.com/) EASY LOG V2.1/EasyBBS Ver1.05/EasyBBS Ver1.05/EASY L ..
4	NTT Basic Research Laboratories (http://www.brl.ntt.co.jp/) NTT 情報科学研究部/NTT 基礎研究所/ .. NTT Communication Science Laboratories (http://www.kecl.ntt.co.jp/) Toshio Irino's Homepage/Dr. ..
3	The chair of Natural Language Processing (http://galaga.jaist.ac.jp:8000/) Back to Home Page/Bick ..
3	NSK (lttp://w2292.nsk.ne.jp/) 理容室はやし//理容室 はやし の ホームページ
3	バーチャルネットアイドル･ちゆ１２歳 (http://tiyu.to/) ちゆ１２歳/へーチャルネットアイドル･ち .. Selfish! (http://selfish.ug.to/) Selfish!
3	ワーナーマイカル (http://www.warnermycal.com/) ワーナーマイカル/★シネマズ御経塚/ワーナー /.
3	Technical Program Area (http://icassp2000.sdsu.edu/) 2000 IEEE International Conference on Acou .. ICASSP-99 Home Page (httt://icassp99.asu.edu/) official homepage/ICASSP'99 (http://www.icslp2000.org/) ICLSP'2000 DTT-TUB - Department of!Telecommunications and Telematics (http://tel.tvt.bme.hu/) EuroSpeech'99
3	いいねっと金沢 (http://www.city.kanazawa.ishikawa.jp/) Kanazawa/金沢市のホームペー ジ/いいね ..
3	辞書･辞典･用語集のリンク集 (http://jisyo.com/) jisho.com/拡張子辞典/拡張子辞典
2	Home Page of Yukimitsu Izawa (http://minerva.jaist.ac.jp;8080/) ポケットステーションでの開発/イ ..
2	System Control and Management Laboratory Lab. Home Page (http://grampus.jaist.ac.jp:8080/) 宮地 ..
2	Perceptual Computing Group Home Page (http://tk01.tk.elec.waseda.ac.jp/) Special Interest Group o ..
2	(http://nandenkanden.com/) なんでんかんでん/なんでんかんでん
2	熱流体解析 (http://www.csl.shinshu-u.ac.jp/) 信州大学工学部/さぁ F.E.M を学びましょう
2	北陸鉄道 (http://www.hokutetsu.co.jp/) 22 時 15 分が終電/北鉄/北陸鉄道
2	Japan NetBSD Users' Group (http://www.jp.netbsd.org/) http://www.jp.netbsd.org/ja/Documentation/n ..
2	IEICE TOP PAGE (http://www.ieice.org/) IEICE/電子通信情報学会
2	TONIC - Project Description (http://www-nrc.nokia.com/) 1999 IEEE Workshop on Robust Methods fo ..

2	The Computational Linguistics Lab. (http://cactus.aist-nara.ac.jp/) 情報処理学会研究報告,98-NL-127/ ..
2	高等学校紹介ホームページ (http://www.gdpec.smile.pref.gifu.jp/) Tonojitugyou high school/Kaizukita ..
2	Mizuno Laboratory (http://mizuno-labo.cs.inf.shizuoka.ac.jp/) http--mizuno-labo.cs.inf.shizuoka.ac.jp ..
2	北陸イチバン ネットマガジン ZAZi(ザジ) (http://www.kanazawaclub.com/) movie_kanazawa/ZAZI 金沢倶楽部ホームページ (http://www.k-club.co.jp/) CLUB/金澤倶楽部
2	JR おでかけネット (http://www.jr-odekake.net/) おでかけネット(JR 西日本)/★JR ハイウェイバス/★ ..
2	Computer Vision & Image Media LAB. UNIV. of Tsukuba (http://www.image.esys.tsukuba.ac.jp/) ..
2	OMG ジャパンウェブサイト (http://www.omgj.org/) OMG ジャパンウェブサイト/OMG ジャパン
2	Welcome ¦ The Official Fayray.net (http://www.fayray.net/) Fayrey/Fayray Home Page
2	Nishio Laboratory Home Page (http://www-nishio.ise.eng.osaka-u.ac.jp/) S. Nishio/M. Tsukamoto/htt ..
2	スラッシュドット ジャパン: アレゲなニュースと雑談サイト (http://slashdot.jp/) スラッシュドット ..
2	Information (http://www.ec.t.kanazawa-u.ac.jp/) Department of Electrical and Computer Engineering, ..
2	(http://wwwbase.nacsis.ac.jp/) Acoustical Society of Japan (ASJ)/日本音響学会
2	(http://www.pcunix.org/) アプリケーション/ペンギン活用委員会
2	Cora Research Paper Search (http://cora.whizbang.com/) Cora/Cora Research Paper Search
2	とほほのWWW入門 (http://tohoho.wakusei.ne.jp/) とほほの WWW 入門/とほほのスタイルシート入門 ..
2	(http://www.theoricon.com/) NO.1!!!!!THE ORICON - MENU/THE ORICON
2	TOEIC (http://www.toeic.or.jp/) TOEIC/TOEIC
2	Miyagi University of Education (http://www.miyakyo-u.ac.jp/) 宮城教育大学/宮城教育大学
2	Index of isWeb19.infoseek.co.jp (http://www19.freeWeb.ne.jp/) 学部時代作った HP/すずすけ
2	Home page : Department of Computer Science (http://www.cs.titech.ac.jp/) CS Dept./Operating Sys ..
2	WWW of Dept. of Administration Engineering, Keio Univ. (http://www.comp.ae.keio.ac.jp/) ソフトウ ..
2	Pocketstudio.jp - ポケットスタジオ (http://pockets.to/) ICQ 道場/ICQ 道場 2000
2	日本 Linux 協会/Japan Linux Association (http://jla.linux.or.jp/) 日本 Linux 協会/日本 Linux 協会

Table 5.5 The result from school of material science

Number of members	Descriptions
58	Welcome to JAIST (http://www.jaist.ac.jp/) Parent Directory/Parent Directory/Parent Directory/Paren ..
7	(http://wwwsoc.nacsis.ac.jp/) 日本応用磁気学会/日本物理学会/●日本物理学会/日本物理学会/日本 ..
6	Home Page of Tohoku University (http://www.tohoku.ac.jp/) 東北大学/東北大学 京都大学 (http://www.kyoto-u.ac.jp/) 京都大学 Tokyo Institute of Technology (http://www.titech.ac.jp/) 東京工業大学/東京工業大学 Kyushu Institute of Technology:#000-e:KIT Home (http://www.kyutech.ac.jp/) ftp.kyutech.ac.jp 構造
5	Chemistry.org: Science that Matters - brought to you by the American Chemical Society (http://www ..
4	(http://www.twmc.ac.jp/) 東京女子医科大学/Unix Manual 上智大学ホームページ (http://www.sophia.ac.jp/) Sophia University/理工学研究科 応用化学専攻 三重大学－Mie University (http://www.mie-u.ac.jp/) 三重大学 Infor. Proces. Center of Y.U. (http://www.cc.yamaguchi-u.ac.jp/) 山口大
4	(http://www.ntt.jp/) NTT Home Page/Japan/●HTML マニュアル (基礎編) /Japanese Information/
4	asahi-net - Homepage (http://www.asahi-net.or.jp/) 上田篤史/第 10 回あるしす記念演奏会/Futoshi Eb .. @nifty:@homepage:メンバーズホームページ移転のお知らせ (http://member.nifty.ne.jp/) ByeByeSep .. ERROR (http://www2s.biglobe.ne.jp/) Santana HP www.ne.jp (http://www.ne.jp/) 精神現象学 (村のホームページ)
4	Yahoo! JAPAN (http://www.yahoo.co.jp/) Yahoo! JAPAN/ヤフー/YAHOO! Yahoo!ジオシティーズ (http://www.geocities.co.jp/) 不涸井荘/Fool's Paradise
3	日本化学会 (http://www.chemistry.or.jp/) 応用化学会/日本化学会 (CSJ)/日本化学会
3	JSAP 応用物理学会 (http://www.jsap.or.jp/) 応用物理学会/応用物理学会/応用物理学会
3	電気学会ホームページ (http://www.iee.or.jp/) 電気学会/電気学会/電気学会
3	(http://navi.ntt.jp/) ●日本のWWWサーバー (NTT 版) / ODIN has moved (http://kichijiro.c.u-tokyo.ac.jp/) ODIN (http://yahho.ita.tutkie.tut.ac.jp/) Company.help_wanted Department of Information and Computer Science Home page (http://www.info.waseda.ac.jp/) Senri .. (http://www1.sony.co.jp/) BIGTOP
3	Wiley-VCH (http://www.wiley-vch.de/) Macromol. Rapid. Comm./Wiley VCH/Helvetica Chimica Acta/C ..

3	サントリーホームページ (http://www.suntory.co.jp/) 第 5 回芥川作曲賞/モルツ/サントリーの HP
3	American Institute of Physics - Home Page (http://www.aip.org/) American Institute of Physics/47th .. www.iop.org from The Institute of Physics (http://www.iop.org/) ●Journal of Physics: Condensed ..
3	Welcome to InfiNet (http://www.infi.net/) The Translators Home Page/●色見本/Color
3	(http://www.elsevier.nl/) Elsevier Science - Home Page/Elsevier Science/Internet Catalogue Biomed .. Oxford University Press - OUP - UK Official Home Page of Oxford University Press - Oxford Books (h .. (http://www.wkap.nl/) Biotechnology Letters/Perspectives in Drug Discovery and Design/Journal of ..
3	goo (http://www.goo.ne.jp/) goo/グー NIKKEI NET (http://www.nikkei.co.jp/) ●日経サイエンス
2	(http://oe.bk.tsukuba.ac.jp/) 文部省科学研究費補助金・特定領域研究A「新しい材料システム構築 ..
2	Research Institute of Electrical Communication (http://www.riec.tohoku.ac.jp/) 電気通信研究所/東 ..
2	高分子学会－HomePage－ (http://www.spsj.or.jp/) 高分子学会 (SPSJ) /The Society of Polymer Sc ..
2	日本学術振興会 (http://www.jsps.go.jp/) 日本学術振興会/日本学術振興会(JSPS)
2	Institute for Chemical Research, Kyoto University (http://www.kuicr.kyoto-u.ac.jp/) 生体分子情報研究 ..
2	軽部研究室 (http://t-rex.bio.rcast.u-tokyo.ac.jp/) 東京大学先端科学技術研究センター軽部研究室
2	鳥取県の情報 (アピオン) (http://www.apionet.or.jp/) 鳥取県倉吉市/鳥取県倉吉市/●アピオネット ..
2	Institute for Solid State Physics (http://www.issp.u-tokyo.ac.jp/) ISSP-Kashiwa 2001 'Correlated Ele ..
2	Nature Japan (http://www.naturejpn.com/) Nature Japan/Nature Japan Home Page -- Main Menu
2	理化学研究所 RIKEN (http://www.riken.go.jp/) RIKEN/理化学研究所 (RIKEN)
2	National Institute of Genetics WWW Title Page (http://www.nig.ac.jp/) National Institute of Genetics ..
2	RSC - Home Page (http://chemistry.rsc.org/) The Royal Society of Chemistry/Royal Society of Chemis ..
2	NIH-NET Cover Page (http://www.nih.go.jp/) Research Tools/Quadrophenia Home Page
2	産業技術総合研究所 (http://www.aist.go.jp/) 強相関電子物性の研究 (産総研) /独立行政法人 産業技 ..
2	宇宙開発事業団ホームページ (http://www.nasda.go.jp/) 宇宙開発事業団 (NASDA)/▲宇宙開発事業団
2	Yahoo!天気情報 - トップ (http://weather.yahoo.co.jp/) Yahoo! ゴルフ場の 天気予報 (石川) /石 川県 ..
2	www.pdb.bnl.gov Redirection Page (http://www.pdb.bnl.gov/) PDB WWW Server/PDB WWW Home ..
2	電子技術総合研究所ホームページ (http://www.etl.go.jp/) 物理 化学関連サーバーへのリンクページ/W ..
2	Harcourt International - Where Learning Comes To Life (http://www.hbuk.co.uk/) Academic Press - .. Welcome to Academic Press (http://www.apnet.com/) Academic Press
2	ExPASy has moved (http://expasy.hcuge.ch/) ExPASy - Compute pI/Mw tool/Swiss-Model: Automate ..

2	(http://www.threeWeb.ad.jp/) douga.html/▼WAKUWAKU HP BIRTHDAY (http://www.bekkoame.or.jp/) Software Catalog
2	GenomeNet WWW server (http://www.genome.ad.jp/) GenomeNet WWW server/GenomeNet WWW .. The RCSB Protein Data Bank (http://www.rcsb.org/) Protein Data Bank
2	kyoto-Inet (http://Web.kyoto-inet.or.jp/) 化学同人/Kyoto-Inet FTP Server
2	Navigator of Web! (http://www.iijnet.or.jp/) CSJ INDEX/ＭＳ機器
2	NCBI HomePage (http://www3.ncbi.nlm.nih.gov/) Entrez/Entrez MEDLINE query/Entrez MEDLINE query NCBI HomePage (http://www.ncbi.nlm.nih.gov/) The National Center for Biotechnology Information/Pu ..
2	Yahoo! (http://www.yahoo.com/) Yahoo!/Yahoo

We used the following questionnaire to evaluate above results.

Please select one item from the following on each Web community

① I think this Web community explains some interest.

② I do not judge whether this Web community explains some interest.

③ I think this Web community does not explain some interest.

④ I do not understand what is the topic of the Web community.

Figure 5.4 An questionnaire to evaluate a Web community

We used the ①, ②, and ③ to evaluate Goal 2 explained previously and the ④ to evaluate Goal 1.

We asked seven people to answer the above questionnaire. It should be noted that such seven people belong to the school of knowledge science. Therefore, we recognize unbalanced answer among three domains. Moreover, we also recognize unbalanced answer among seven people, since the questionnaire was performed without additional investigation into a Web community such as looking the homepages that compose the Web community. In other words, although the Web community really explains some interest, he/her may answer "I do not understand what is the topic of the Web community" to the questionnaire, when he/her does not know any information about the topic.

First, we present the whole result that presents the average values from three domains.

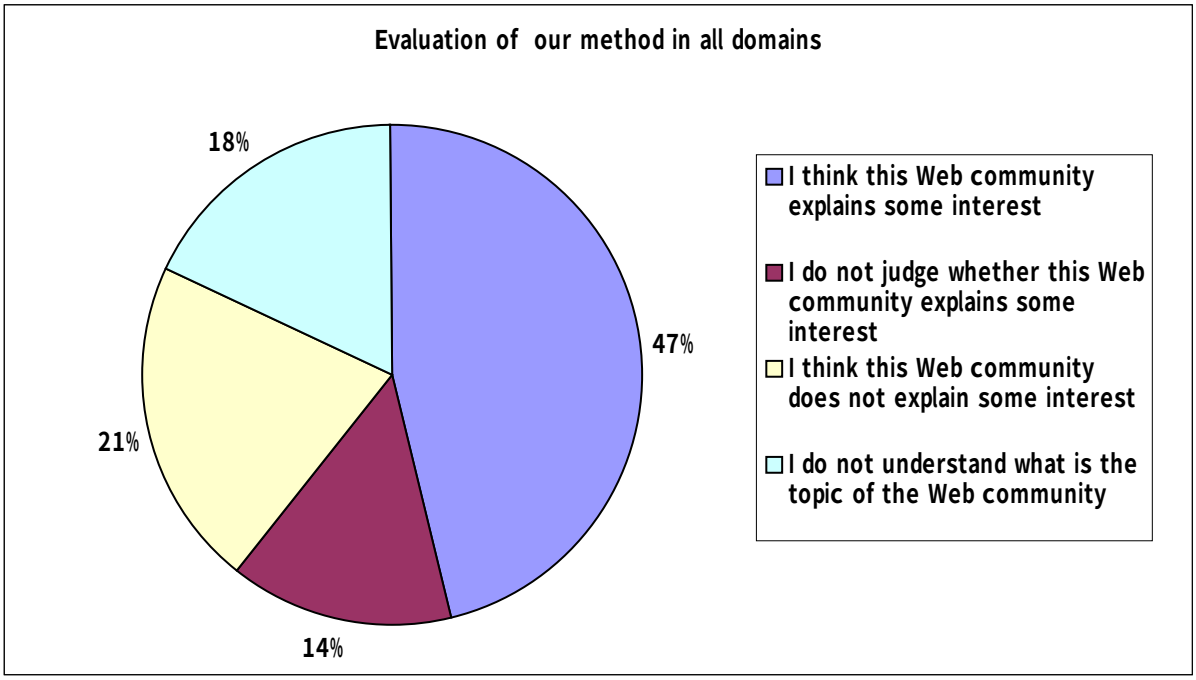


Figure 5.5 Evaluation of our method in all domains

Next we present the result of each of domains

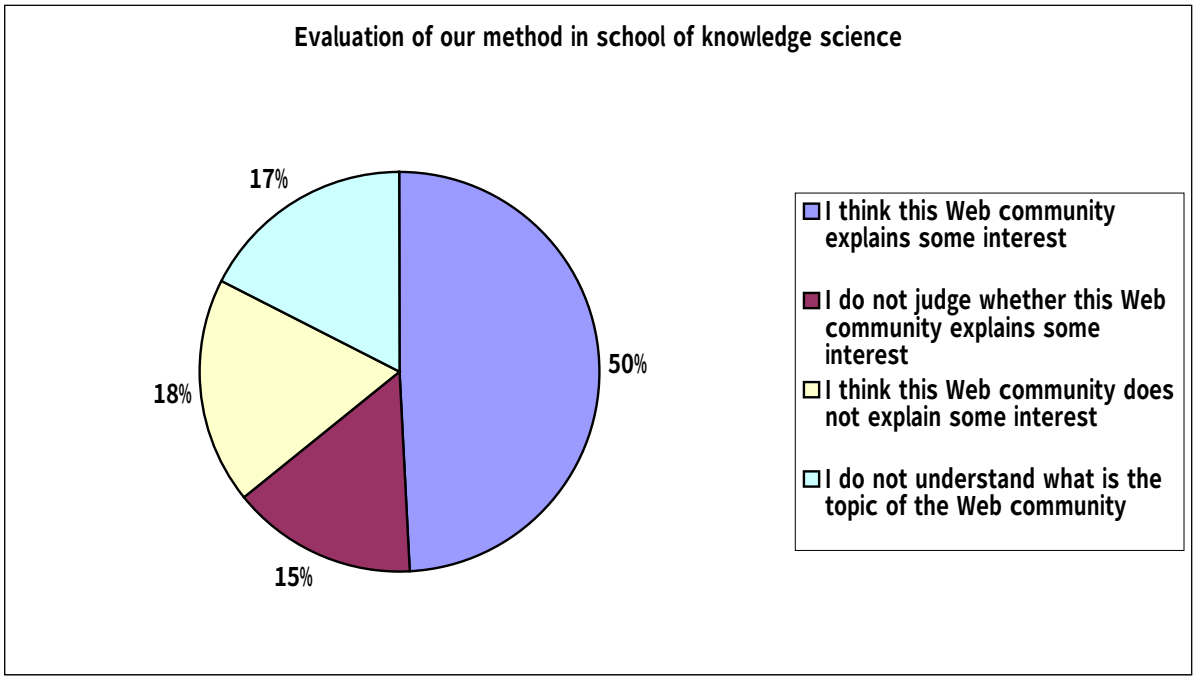


Figure 5.6 Evaluation of our method in school of knowledge science

In Figure 5.6, because all people who answered the questionnaire belong to the domain, The number of "I think this Web community explains some interest" answer is larger than the other domains.

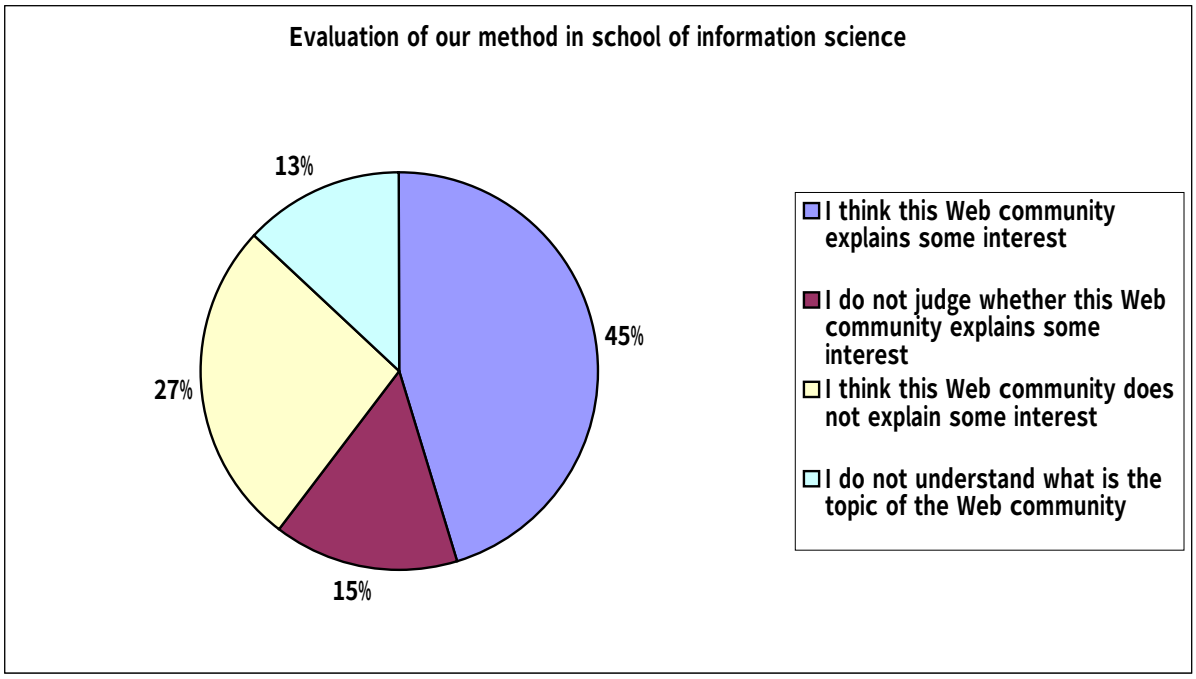


Figure 5.7 Evaluation of our method in school of information science

The feature of this result is that number of "I do not understand what is the topic of the Web community" is small compared with the other domain.

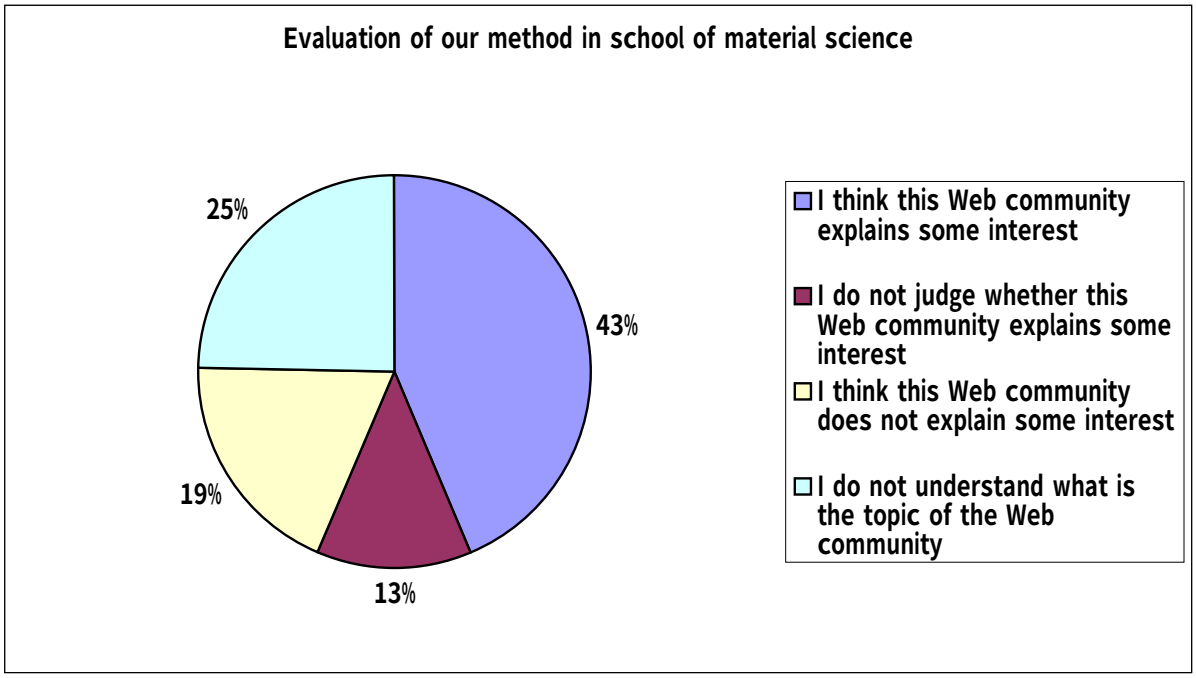


Figure 5.8 Evaluation of our method in school of material science

The feature of this result is number of "I do not understand what is the topic of the Web community" is large compared with the other domain. It may be due to background knowledge of the people that answered the questionnaire. In our investigation, there is much domain specific information in the domain.

Next we present a figure in order to confirm that the answer is imbalance among people.

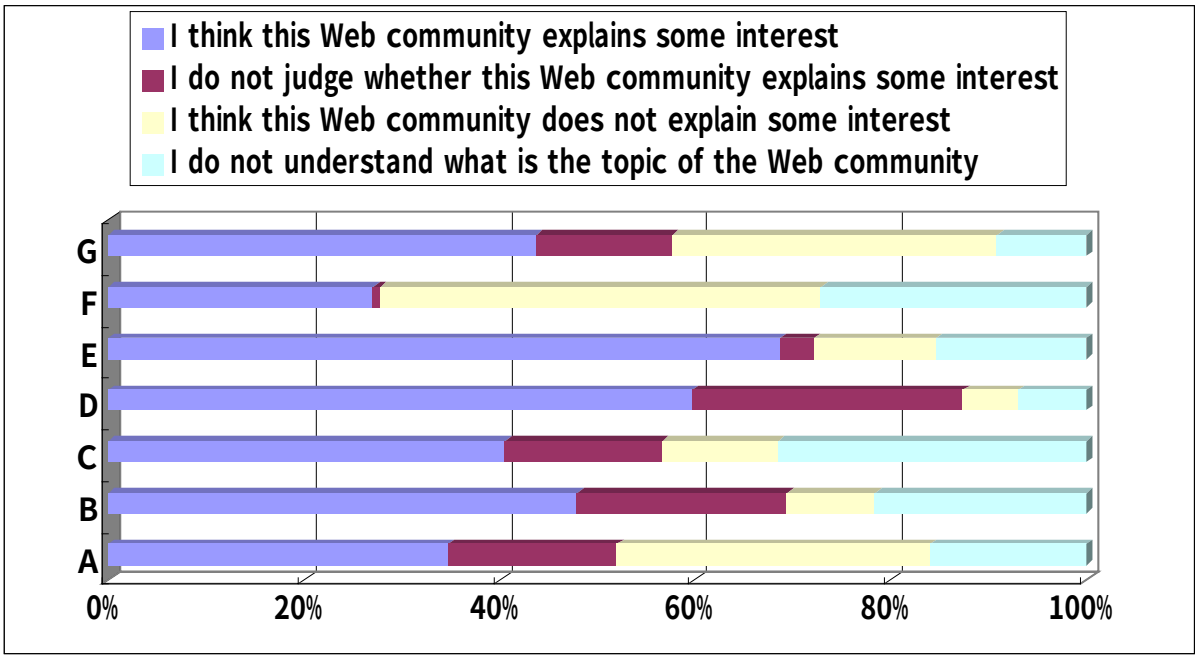


Figure 5.9 Percentages of the four items

In Figure 5.9, x-axis is the rate of four items and y-axis is seven persons that answered the questionnaire. Unbalanced answer is clear from this figure. However, other viewpoints can be seen from a following figure that shows imbalance among four items.

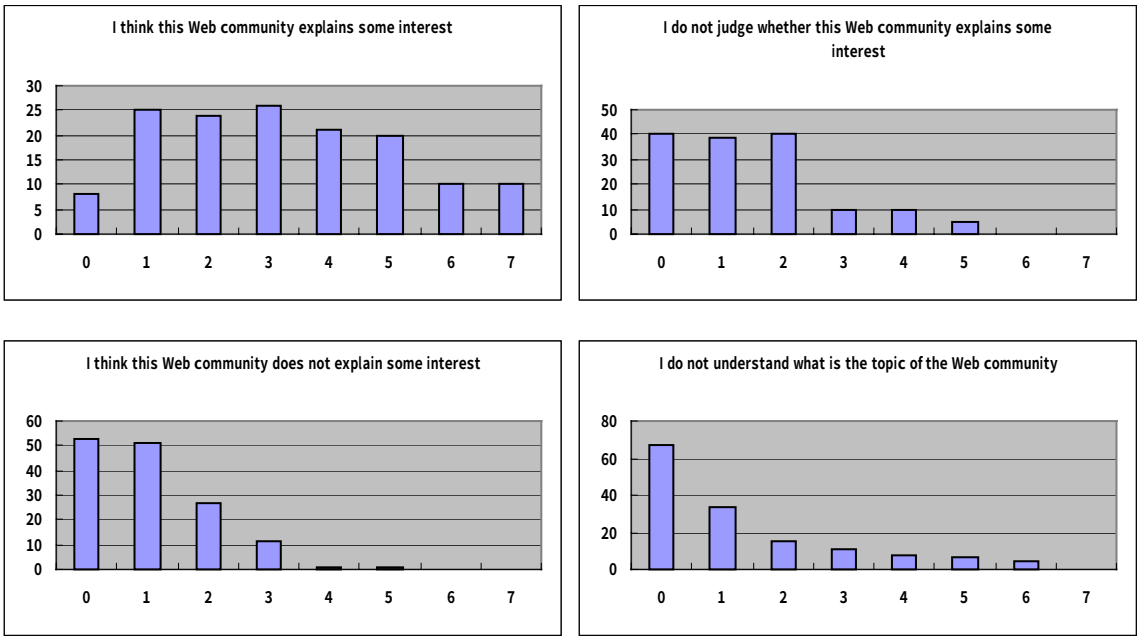


Figure 5.10 The number of web communities that people selected the item that corresponds to the graph title for

Figure 5.10 shows distributions of each of four items. In this figure, x-axis is the number of persons that judged the item that corresponds the title of the graph, y-axis the number of Web communities. For example, in left top figure, the number of Web communities judged "I think this Web community explains some interest" by seven persons is 10, and the number of it by five persons is 20.

The bottom right figure in figure 5 shows the fact that answers for the question “I do not understand what is the topic of the Web community” is not imbalance. Percentages of Web communities that no one judged that the Web community is non-understandable is about half. Moreover, percentages of Web communities that at most two people judged is about 80%. Therefore, it seems to be quite reasonable to consider 80% Web communities are understandable.

Left top figure in Figure 5.10, however, shows that there are many unclear Web communities judged "I think this Web community explains some interest".

We would like to evaluate the result in the light of questions we described previously.

Regarding the question 1 it seems to be quite reasonable to consider 80% discovered Web communities are understandable. Although Web Trawling [Kumar 00] that shares similarities with our method stated that 96% Web communities are reliable, it is not shown clearly how the experiment performed. However, according to many related other works on discovering of Web communities, the result we obtained can 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.

5.4 Experiment in Other Domains

We would like to present the result performed in the other domains that are Stanford University and MIT in U.S.A. Personal homepages in such domains could be gotten at http://www.stanford.edu/leland/dir.html and http://web.mit.edu/search.html respectively. The results are created from 200 homepages selected at random. In addition, we did not use the URL whose character sequences was contained "stanford.edu" or "mit.edu", because those two domains contained a lot of HTTP servers, for example http://www.humnet.ucla.edu/, http://www.oakland.edu/, http://www.cs.umd.edu/, http://web.syr.edu/, and so on. The period of the experiment was from February 4 to 6, 2002.

Table 5.6 The result from Stanford University

Number of members	Descriptions of the Web community
14	Yahoo! (http://www.yahoo.com/) Yahoo/Yahoo/Yahoo/Yahoo!/Yahoo/Yahoo!/Yahoo/Yahoo!/Yahoo/Yaho .. Welcome to MSN.com (http://www.msn.com/) Mis sonrisas! (http://www.excite.com/) Excite.com.
10	AltaVista - The Search Company (http://www.altavista.com/) Alta Vista/Alta Vista/Alta Vista/Alta Vis .. Google (http://www.google.com/) Google/Google/Google,/Search Engine ? Google/driver drowsiness ..
10	Yahoo! GeoCities - Your Home on the Web® (http://www.geocities.com/) www.geocities.com/cpe .. AOL Hometown (http://members.aol.com/) CELEBRATION OF ONENESS CENTER/Anthrax/Judas Priest Angelfire (http://www.angelfire.com/) The Writer's Realm Writing Ring//Submit Site/LIFE NOW/[jen kr .. Tripod (http://members.tripod.com/) INSPIRATIONS FROM THE LIGHT//ShiZhang LIN
7	washingtonpost.com - News Front (http://www.washingtonpost.com/) Taken from the Washington P .. ABCNEWS.com: Home (http://www.ABCnews.com/) ABCNews The New York Times on the Web (http://www.nytimes.com/) HEADLINE: RUSSIA WANTS BINDING AR .. MSNBC Cover (http://www.msnbc.com/) dogs BBC News ¦ Front Page (http://news.bbc.co.uk/) BBC news, CNN.com (http://www.cnn.com/) CNN/?CNN,/visual evidence/CNN
5	Massachusetts Institute of Technology (http://web.mit.edu/) Paul Krugman/MIT Japan Program/PGP i .. Center for Reliable and High Performance Computing (http://www.crhc.uiuc.edu/) 1st Workshop on .. Princeton University (http://www.princeton.edu/) Lars Svensson/Bernasek, S.L.
4	AltaVista - The Search Company (http://www.altavista.digital.com/) Alta Vista/Alta Vista/Alta Vista Bartleby.com: Great Books Online (http://www.bartleby.com/) where to find anything and everything ..
4	TheCounter.com: The Full-Featured Web Counter with Graphic Reports and Detailed Information (ht ..
4	Amazon.com--Earth's Biggest Selection (http://www.amazon.com/) The Papers of Martin Luther Ki ..
3	Journal of Biological Chemistry (http://www.jbc.org/) J Biol Chem/?JBC,

	The Journal of Cell Biology (http://www.jcb.org/) J Cell Biol The EMBO Journal Online (http://www.emboj.org/) EMBO, Proceedings of the National Academy of Sciences (http://www.pnas.org/) Proceedings of the Nation .. Molecular Biology of the Cell (http://www.molbiolcell.org/) Mol Biol Cell
3	(http://www.disney.com/) Disney/Disney Online/Disneyland Merriam-Webster OnLine (http://www.webster.com/) ?Webster,
3	Yahoo! Maps and Driving Directions (http://maps.yahoo.com/) map/Cambridge, MA/directions/Yahoo ..
3	(http://www.concentric.net/) Michael's Tennis Page Welcome to GlobalCrossing (http://www.primenet.com/) web pages
3	NCBI HomePage (http://www.ncbi.nlm.nih.gov/) here/PubMed/PubMed,/?Blast 2 sequences
3	Welcome to the Microsoft Corporate Web Site (http://www.microsoft.com/) Microsoft V-Chat/Micros ..
2	(http://pet.pagecount.com/)
2	PennState Physics Department (http://www.phys.psu.edu/) The Liu Lab at Penn State/Diehl, R.D.
2	Bulgaria.com Home Page (http://www.bulgaria.com/) Bulgaria/Bulgaria
2	Stanford Alumni Association (http://www.stanfordalumni.org/) Office of Alumni Volunteer Relations/ ..
2	JHU Biomedical Engineering Homepage (http://www.bme.jhu.edu/) Scot C. Kuo/Centers for Computa ..
2	Welcome to Cornell University! (http://www.cornell.edu/) Cornell University/Cornell University
2	Welcome to The University of Hong Kong (http://www.hku.hk/) The University of Hong Kong/Linguistics
2	Yahoo! (http://yahoo.com/) Yahoo/YAHOO
2	MSR Home (http://research.microsoft.com/) ICDE/Michael B. Jones/Microsoft Research/Systems an ..
2	Welcome to Logitech (http://www.logitech.com/) iFeel Mouse/WingMan
2	FEMINIST MAJORITY FOUNDATION ONLINE HOMEPAGE (http://www.feminist.org/) Empowering wom ..
2	ProQuest (http://proquest.umi.com/) http://proquest.umi.com/pqdweb/ERP system
2	The Onion ¦ America's Finest News Source (http://www.theonion.com/) the onion/the onion
2	Lonely Planet Online (http://www.lonelyplanet.com/) England/Turkey,/Japan,/Korea,/Taiwan,/Thailand ..
2	HowStuffWorks - Learn how Everything Works! (http://www.howstuffworks.com/) why onions make ..

2	Columbia University in the City of New York (http://www.cc.columbia.edu/) THE PROPHET by Kahlil G ..
2	IBM Research (http://www.research.ibm.com/) T.J.Watson Research Center/Avouris, P.
2	Nature science journals: nature.com (http://www.nature.com/) Nat Cell Biol/?Nature,/?NatureBiotech .. Science Magazine Home (http://www.sciencemag.org/) ?Science
2	The Johns Hopkins University (http://www.jhu.edu/) The Johns Hopkins University/Denis Wirtz/consi ..
2	Scientific American (http://www.sciam.com/) performed the surgery from a facility in New York/?Sci ..
2	(http://www.nps.gov/) National Park/National Park Service/Yosemite National Park
2	Apple (http://www.apple.com/) apple/HERE
2	Massachusetts Institute of Technology (http://www.mit.edu/) 6.857: Network and Computer Security/ ..
2	Nokia on the Web (http://www.nokia.com/) Nokia's Bluetooth page/Nokia's WAP index/Nokia Mobile .. Nokia Press Services (http://press.nokia.com/) Nokia Mobile Phones-Mobile Payment
2	American Express Personal Card, Financial, and Travel Products and Services (http://www.americ ..
2	Verio and Webcom (http://www.webcom.com/) Santa Cruz Sentinel Triathlon/Kansas
2	Library of Congress Home Page (http://lcweb.loc.gov/) US Executive websites/US Legislative websit ..
2	AMA - American Medical Association Home Page (http://www.ama-assn.org/) Elliott, Victoria Stagg ..
2	Top Stories from Wired News (http://www.wired.com/) http://www.wired.com/wired/6.05/europe.ht ..
2	(http://www.webring.org/) King Diamond//Poetry Webring

Table 5.7 The result of MIT

Number of members	Descriptions of the Web community
10	Welcome to NCSA (http://www.ncsa.uiuc.edu/) A Beginner's Guide to HTML/Learn HTML!/NCSA HTML ..
8	AltaVista - The Search Company (http://www.altavista.digital.com/) Alta Vista/alta vista HotBot (http://www.hotbot.com/) hotbot STARS Online: Film, Music, Science, Technology, Places, People, Life (http://www.stars.com/) JavaS .. Art on the Net (art.net) (http://www.art.net/) I also have a studio/Art on the Net GO.com (http://www.infoseek.com/) InfoSeek Home Page/infoseek

	Google Groups (http://www.dejanews.com/) deja news ualberta.ca - University of Alberta home page (http://www.ualberta.ca/) University of Alberta Lycos (http://www.lycos.com/) Lycos Computers Guide: Cyberculture/lycos (http://www.infoseek/) [Infoseek]
5	Welcome to Boston.com (http://www.boston.com/) Boston/Boston/Boston.com//Boston Globe
5	NBA.com (http://www.nba.com/) [NBA]/Celtics NHL.com - The National Hockey League Web Site (http://www.nhl.com/) Bruins ESPN.com (http://espn.go.com/) ここ/Edgerrin James/Dameyune Craig/Griese/NHL/Buffalo/NBA/Celti ..
4	Welcome to Harvard University (http://www.harvard.edu/) Harvard University/ハーバード大学
4	Home Page: School of Computer Science, Carnegie Mellon (http://www.cs.cmu.edu/) Tsinghua Clas ..
4	The Ohio State University Computer and Information Science Department (http://www.cis.ohio-state ..
4	EFF Homepage (http://www.eff.org/) The EFF/electronic frontier foundation
4	Princeton University (http://www.princeton.edu/) ES2001/Chemistry Department/Princeton Universit .. Welcome to Oxford University Computing Laboratory (http://www.comlab.ox.ac.uk/) Audio Page SU Personal Home Pages (http://web.syr.edu/) Austin 'Swinger' Wei Department of Computer Science (http://www.cs.umd.edu/) Encyclopedia of Virtual Environments a2i Communications (rahul.net) (http://www.rahul.net/) Architectour JAPAN 95/Architects Abroad (http://weber.u.washington.edu/) AIAS NorthWest Pre-Forum 1995
3	(http://www.lysator.liu.se:7500/) very weird/pinball/Abrahamsson, Thomas
3	Wind River: Operating Systems: BSD/OS (http://www.bsdi.com/) jad@MIT.EDU/(dougie@athena.mit. .. The NetBSD Project (http://www.netbsd.org/) netbsd
3	Texas Instruments Welcomes You (http://www.ti.com/) Texas Instruments DSP Challenge/ti broadb .. ADI - Homepage (http://www.analog.com/) Analog Devices, Inc.
3	SF Gate: News and Information for the San Francisco Bay Area (http://www.sfgate.com/) The Gate/In .. MSNBC Cover (http://www.msnbc.com/) NBC News/My Turn: I 致 e Seen the Worst That War Can Do/ ..
3	TheCounter.com: The Full-Featured Web Counter with Graphic Reports and Detailed Information (ht .. (Sonic.net, Inc.) (http://www.sonic.net/) Robert Ghostwolf: Native American Spiritual Spokesman or ..
3	The Internet Movie Database (IMDb). (http://www.imdb.com/) the internet movie database/[Movie Dat .. Movie Review Query Engine (http://www.MRQE.com/) [Movie Review]

2	Simmons College - Boston, MA (http://www.simmons.edu/) Simmons College/the Simmons College ..
2	AllFreeStats.com Your Free Tracking Software (http://www.allfreestats.com/)
2	Cambridge, MA - Official Web Site Home page (http://www.ci.cambridge.ma.us/) Cambridge/Cambrid ..
2	core77 design magazine and resource (http://www.core77.com/) Paul Lucas' Inconspicuous Consu ..
2	(http://www2.whidbey.net/) foxes/Sustainable Society
2	Urban75 ezine - direct action, rave, useless games, bulletin boards, drugs, football, photos and mo ..
2	ESPN.com (http://espn.sportszone.com/) ESPN Sportszone/College Football
2	University of Florida College of Liberal Arts and Sciences (http://www.clas.ufl.edu/) LSRL 30/Religiou ..
2	The J.S. Bach Home Page (http://www.jsbach.org/) バッハ/J.S. Bach Title (http://w3.rz-berlin.mpg.de/) 作曲家について
2	Welcome to the UIUC Student/Staff Computing Cluster! (http://www.students.uiuc.edu/) Young Min C ..
2	University of California, Berkeley (http://www.berkeley.edu/) University of California at Berkeley/ber ..
2	IBM Research (http://www.research.ibm.com/) Mark Lucente/[IBM Research]
2	Red Meat - from the secret files of Max Cannon (http://www.redmeat.com/) Red Meat/red meat
2	Charm Net Inc .- Advanced Internet (http://www.charm.net/) PC Game Center/HTML Tables Tutorial
2	The American Society of Civil Engineers World Headquarters (http://www.asce.org/) ASCE/America ..
2	tagesschau (http://www.tagesschau.de/) Tagesschau/
2	Rice Computer Science: Department of Computer Science (http://www.cs.rice.edu/) treadmarks/Dr ..
2	Active Window Productions (http://www.actwin.com/) Movies Index/Islamic Student Center
2	(http://www.wimsey.com/) shower curtain/Steel House in Vancouver, A
2	EDV-Pool Welcome (http://bau2.uibk.ac.at/) the UK/Starting Points for Architecture and Visualization
2	X.org (http://www.x.org/) The X Consortium/x windows
2	KZSU 90.1 fm (http://kzsu.stanford.edu/) The Web's Edge/Blade Runner/2019: off-world (http://www.wpi.edu:8080/) Star Wars
2	Home Page do Laboratório de Sistemas Integráveis (LSI) (http://www.lsi.usp.br/) Ar ..
2	Purdue University-West Lafayette, Indiana (http://www.purdue.edu/) Purdue University/purdue
2	(http://travel.roughguides.com/) rough guides/The Rough Guide
2	Rensselaer Polytechnic Institute - Engineering, Information Technology, Management, Science, Arc ..
2	AMG All Music Guide (http://www.allmusic.com/) allmusic

2	Electronic Engineering at Surrey (http://www.ee.surrey.ac.uk/) Isobitis/Kokoras/KYR!/UK/UK
2	AnyBrowser Pages (http://www.anybrowser.org/)
2	Colby College ¦ Four Year Private Undergraduate Liberal Arts College in Waterville, Maine (http://w ..
2	The Onion ¦ America's Finest News Source (http://www.theonion.com/) The Onion/the onion FuckedCompany.com - Official lubricant of the new economy (http://www.fuckedcompany.com/) fuc ..
2	Välkommen till Nada (http://www.nada.kth.se/) Wierd Religions/[NP Optimization Problems]
2	Free Music Download, MP3 Music, Music Chat, Music Video, Music CD, ARTIST direct Network (http:/ ..
2	The Internet Movie Database (IMDb). (http://us.imdb.com/) Phantom of the Paradise/Movie Reviews
2	Real.com - RealOne Player (http://www.real.com/) Real Player

Chapter 6

Conclusion

The objective of our study is to propose the method for extracting Web communities of personal interests in a specific domain. Such Web communities are useful for human recommender system, knowledge management, and so on. Based on our analyses of existing methods for extracting Web communities, we pointed out that such methods could not apply to our problem.

We developed a new method to solve the problem that we formulated. The method explained previously 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 a Web community. First, our method gathers hyperlinks from personal homepages in a specific domain. Second, our method sorts such hyperlinks in the order of OScore that was proposed in chapter 3. Finally, our method gets one URL as a seed of a Web community from the hyperlinks, and gathers URLs that have a similarity to the seed from hyperlinks, and then created groups of URLs are Web communities. 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.

We performed experiments in five domains, and evaluated our method by using a questionnaire. Roughly 80% discovered Web communities were judged understandable, and about 40% to 50% Web communities explained some interests.

Considering that the similarity measure of URLs is of most importance in our

study, we observed that among URLs found to be similar by the used measure, many of them were judged similar by the human, but also many of them were judged “not similar”. In order to improve the method, we think that it is necessary not only to solve this problem but also to use effectively other available information such as texts in a personal homepage, bookmarks, and so on.

Finally, we should note that a few people who answered a questionnaire used in our evaluation said, "I want to know who links this Web community". It is an instance of a human recommender system that is an example of application of our study.

Acknowledgements

I am indebted to the participants in the studies for their gracious cooperation. Thanks also go to professor Ho Tu Bao, associate professor Masato Ishizaki, associate professor Takashi Hashimoto, associate Nguyen Trong Dung, and other participants for their support, assistance, and efforts.

References

[AltaVista] http://www.altavista.com/

[AllTheWeb] http://www.alltheWeb.com/

[Eades 84] P. Eades, A Heuristics for Graph Drawing, Congressus Numerantium, Vol. 42, pp.149-160, 1984

[Google] http://www.google.com/

[Google a] http://www.google.com/terms_of_service.html

[Google b] http://www.google.com/technology/

[Kauts 97] H.Kauts, B. Selman, and M.Shah: “The Hidden Web”, AI Magazine, vol.18, no.2, pp.27-36, 1997

[Kleinberg 99] Authoritative Sources in a Hyperlinked Environment”, Jornal of the ACM Vol. 46 Num.5 pp.604-632, 1999

[Kumar 00] R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins: “Trawling the Web for emerging cyber-communities”, Proceedings of the 8th WWW conference, 1999

[Murata 00a] Tsuyoshi Murata: “Discovery of Web Communities Based on the Co-occurrence of References”, Proc. of the Third International COnference on Discoverty Science (DS’2000), 2000

[Murata 00b] Tsuyoshi Murata: “Discovery of the Structures of Web Communities”, JSAI SIG-KBS-A002-2,pp 7-12,2000

[Nonaka 01] Ikujiro Nonaka, Katsuhiro Umemoto: “Managing Existing Knowledge Is Not Enough: Recent Developments in Knowledge Management Theory and Practice in Japan”,Journal of Japanese Society for Artificial Intelligence,Vol.16 No.1, pp.4-14, 2001

[Page 98] Page, L. Brin, S.Motwani R. and Winograd T.: "The PageRank Citation Ranking: Bringin Order to the Web, Online manuscript, http://www-db.stanford.edu/~backrub/pageranksub.ps, 1998

[Salton 89] Salton, R. Automatic Text Processing, Reading, Mass.: Addison-Wesley, 1989

[Stanley 94] Stanley W.,Katherine F.: “Social Network Analysis Methods and Applications”, Cambridge, University Press, 1994

[Sugiyama 95] Kozo Sugiyama, Kazuo Misue: "A Simple and Unified Method for Drawing Graphs: Magnetic-Spring Algorithms", Proc. DIMACS Int. Work.

Graph Drawing, GD (Princeton, U.S.A.; 10-12 Oct, 1994); Springer-Verlag, Lecture Notes in Computer Science, 894:364-375, 1995

Contributions

[1] Toyohisa Nakada, A Hyperlink-induced Method for Extracting Implicit Communities, SIG-KBS JSAI, MITSUBISHI ELECTRIC CORPORATION, September 14, 2001