Structure Mining for Intellectual Networks

Ryutaro Ichise[1], Hideaki Takeda[1], and Kosuke Ueyama[2]

[1] National Institute of Informatics,
2-1-2 Chiyoda-ku Tokyo 101-8430, Japan,
{ichise,takeda}@nii.ac.jp
[2] TRIAX Inc.
4-18-2-203 Takadanobaba, Shinjyuku-ku, Tokyo, 169-0075 Japan,
ko@triax.jp

Abstract. The research community is very important for researchers in order to undertake new research topics. In the present paper, we propose a community mining system that helps to find communities of researchers using bibliography data. The basic concept of the present study is to provide interactive visualization of communities both local and global. We implemented this concept using actual bibliography data and present a case study using the proposed system.

1 Introduction

Recently, communities have begun to include a large number of members because this helps to make information more reliable. In such communities, only informative information can be propagated among several users. As a result, we can obtain useful and reliable information from communities. This situation also exists in the field of research. Most researchers do not include all of the technology used in their experiments when writing a paper because some techniques cannot be represented in words, such as computer coding techniques. Such techniques propagating within the local research community may be useful for developing new technologies.

In order to clarify existing research communities, bibliography information is used widely. In co-citation analysis [12], all papers that are cited in a paper make up the community of a certain research area. However, this paper-based analysis overlooks the characteristics of individual researchers. As a result, realizing that each researcher has an individual research area is difficult. On the other hand, CiteSeer [3] and Google Scholar [6] are able to handle research communities from a micro viewpoint because they handle co-author and citation information from bibliographies and use the information for individual researchers. Although these systems are good for finding micro communities, they are not suitable for finding the position of a researcher within a global research community. In the present paper, we propose a community mining system for researchers that has both local and global viewpoints. The proposed system will facilitate understanding of the researcher community and will advance new areas of research.

The present paper is organized as follows. In Section 2, we discuss the proposed design for a community mining system. In Section 3, we describe a community mining

system implementation based on the design policies discussed in Section 2. In the Section 4, we explain the proposed system using a number of examples. Finally, in Section 5, we discuss our conclusions and areas for future study.

2 System Design

2.1 Relationships for Communities

The most important information for finding communities of researchers is contained in the bibliography. In scientific network research, several relationships can be obtained from this information. The relationships in the knowledge domain include co-authorship [11], citation [5] and co-citation [12]. In the present paper, we use three relationships in the knowledge domain to find research communities:

- co-authorship
- citation
- author citation

The first two relationships are well known among knowledge domain researchers. With respect to co-authorship, if we can consider the researcher as a node and co-authorship as an arc, then we can obtain networks of researchers. We can consider that researchers, whose nodes are linked to have the same research interests. With respect to citation, if we can consider the research paper as a node and citations in papers as arcs, then we can obtain networks of papers. The final relationship is author citation. This relationship also represents relationships among authors. Research papers include citations of papers, as described above. This citation indicates that both papers (which have a citation relationship) are related to the same subject. In the same manner, the authors of both papers have the same research interests. In the present study, these relationships are referred to as author citations and are used for community mining.

2.2 Community Mining

In this section, we describe the community mining method for the networks described in the previous section. We use a policy for community mining borrowed from the active mining approach [9]. The basic concept of active mining is to utilize interactions between the user and the computer. We separate mining method into two steps. The first step is to automatically mine communities by computer, and the second step is to visualize the information in order to facilitate the understanding of the information by the user. After viewing the communities that are discovered by computer, the user can specify a command to find more refined communities. As we described above, the proposed method can facilitate the finding of communities through interactive manipulation of the mining result obtained by computer.

Several types of indexes have been proposed for finding communities [4]. In the present study, we use indexes for finding communities as follows:

1. Simple weight

2. Maximum flow
3. Closeness

The first index, simple weight, is a measure for using weight on arcs such as representing the number of co-authors. When the weight is small, the arc is considered to represent an unimportant relationship. Therefore, highly weighted arcs represent communities, and thus can be used as an index for the community. The second index, max flow, is a measure that focuses on the connection between two different nodes. This index considers the weight on an arc as the thickness of a pipe and measures the connection between the two nodes. The index is able to calculate the distance between two nodes that are not directly connected. When the two nodes have a thick connection, even though they have relay nodes, the index may have a high value. Therefore, this index may be a good measure for finding communities. The third index, closeness, is a measure that is used to calculate the distance between two nodes. This index represents the shortest distance between the two nodes. Therefore, when the distance is small, the two nodes are considered to be in the same community.

3 Research Community Mining System

3.1 System Architecture

In the present study, we implemented a community mining system using the policies presented in the previous section. The components of the system are shown in Figure 1. The system has two databases, the CiNii [2] database and an experiment database, and five program units. The CiNii database will be described in detail later herein.

The database generation unit in Figure 1 selects records from the CiNii database and sends them to the database management unit. MySQL [10] is used as the database management unit. If possible, the mining index discussed in Section 2.2 is then calculated by the mining index calculation unit. The components are written in Perl. Note that the entire mining index cannot be calculated at this stage because of the need for the user query. The result of the calculated index is also stored in the experiment database, which is handled by the database management unit.

Users access the system through the visualization control unit, which is constructed using a web browser that includes Flash Player and SVG Viewer. Then, when a user inputs data from the web browser, the data will be sent to the I/O control unit via the Internet. The I/O control unit, which is constructed by a web server that includes the CGI program, generates data for visualization from the user input. The components are also written in Perl. The I/O control unit calls the database management unit to retrieve data and calls the mining index calculation unit to calculate the mining index.

In the remainder of this section, the CiNii database, the preprocessing method for data, and the visualization method are discussed in detail.

3.2 Bibliography Database

In the present study, we use the CiNii database to obtain bibliography information. The database is described by approximately 320 megabytes of SGML data. The CiNii

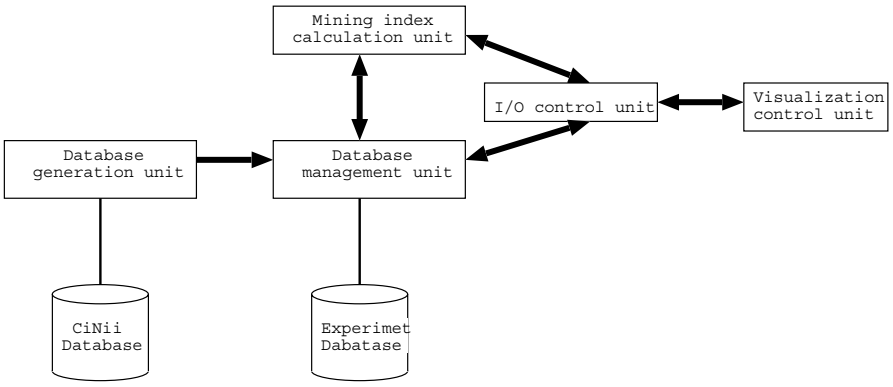

Fig. 1. System architecture.

database contains bibliography entries, such as title, author and publication year. The number of data are listed in Table 1. The number of database records in CiNii denotes the number of records in the original database, as described in SGML format. The Paper and Researcher entries denote the number of records for papers and the number of records for researchers, respectively. The number of researchers is less than the number of papers, because one researcher could write more than one paper. The Author entries in Table 1 denote the number of authors for each paper. For example, the record is counted as three when three researchers write a paper corroboratively. The Co-author entries denote the number of combinations of authors for a paper. For example, when a paper is written by four authors, it is counted as ${}_4C_2 = 6$ for the paper. The Citation(Paper) entries denote the number of citations of a paper. Although one paper usually cites a number of other papers, the Citation(Paper) entries are lower than the Paper entries because the CiNii database contains part of the citation information. The Citation(Author) entries denote the number of researchers who wrote cited papers, with duplication. Comparing the Paper and Researcher entries, the number of researchers is less than the number of papers. In contrast, comparing the Citation(Paper) and Citation(Author) entries, the number of Citation(Author) entries is larger than the number of Citation(Paper), which implies that we prefer citing various papers to citing several papers written by the same author. Note that the database used in the present study was created in October of 2003, and is therefore different from the current CiNii database.

3.3 Preprocessing of Database

In order to construct the experiment database, the data discussed in the previous section was preprocessed by a database generation unit. In this process, a few attributes are selected from the CiNii database, because most of the attributes, such as ISSN number, included in the CiNii database are not useful in the present experiment. In addition, the database generation unit conducts record linkage for author records. It has heuristics for dividing author records, because the author records in the CiNii database some-

Table 1. Number of data.

	Number of records in CiNii ($\times 1,000$)	Number of records in Experiment Database ($\times 1,000$)
Paper	544	128
Researcher	224	90
Author	787	358
Co-author	1103	231
Citation(Paper)	445	36
Citation(Author)	1562	349

times have multiple author names in single author field. For treating such records, the database generation unit divides long author names using special characters, such as $\star$. In addition, a number of other small record linkage techniques are also used in this stage.

After the preprocessing has been completed, the experiment database is created by the database management unit. The numbers of records for the experiment database are listed in Table 1. The experiment database contains much less data than the original CiNii database because the bibliography of the paper included in the CiNii database as original information was only used. Although the CiNii database contains cited paper information, the information is not complete and contains several mistakes, and is therefore not used in the present experiment.

3.4 Visualization

The visualization control unit creates a visualization screen for finding communities easily. In order for a user to find a community, the user must be able to browse the data interactively. For this purpose, we used a web browser, which includes Flash Player and SVG Viewer, as a visualization control unit. Data for this unit is provided by a web server on the Internet. The server is referred to as the I/O control unit. We used two types of visualization data provided by the I/O control unit: global visualization by SVG [13] and local visualization by Flash. Global visualization by SVG facilitates the location of macro communities by showing global relationships on the graph, whereas local visualization by Flash facilitates the location of micro communities by showing the local relationships near the individual. Both of these methods will be discussed in this section.

We first discuss graph visualization by SVG. As discussed in Section 2.1, representing the relationships between communities in the form of a graph will facilitate the finding of communities. Therefore, we use a graphic representation scheme visualization by SVG. SVG is an XML format for representing graphs. The SVG file can be visualized by an SVG viewer. An example of SVG graph visualization of the relationships among researchers is shown in Figure 2. However, using such a graphic representation is not suitable for finding communities in the graph, because most of the nodes will be connected [1]. As a result, the graph could be very large and the user might not be able to find communities in the graph. In order to discern communities from a large graph, the system should have a function that shows parts of the graph that

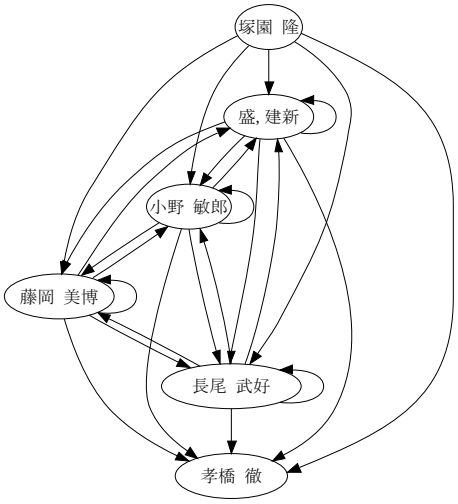

Fig. 2. An example of graph representation by SVG.

may contain the desired communities. Therefore, we herein propose interactive mining in conjunction with the indexes presented in Section 2.2. When the user specifies an index and its threshold, the system automatically divides the graph. This is very important for finding communities because it is hard to specify an index to match the purpose of the user. Although the current implementation of the mining index calculation unit calculates the three indexes presented in Section 2.2, the I/O control unit can only create an SVG graph for simple weight.

Next, we will discuss local representation by Flash. After finding a community via visualization by SVG, we must learn more about each researcher in order to refine the communities. For this purpose, users require not only a global point of view with respect to the communities, but also a local point of view that focuses on each researcher. We propose to visualize local communities, to which a specified researcher belongs, in order to facilitate the refinement of communities. Locally, a specified researcher is located at the center of a field, and related researchers are placed around the circumference of the initially specified researcher. This visualization is suitable for finding communities that are built around a researcher. Since we focused on communities of researchers, the current implementation includes only the relationships of co-author and author citation, as discussed in Section 2.1. An example of the local point of view is shown in Figure 3. The circle in the center represents the currently specified researcher of focus. The surrounding circles represent researchers who are related to the initially specified researcher. Finally, the outermost ring of circles represents researchers who are related to the researchers of the inner ring. The size of each circle represents the number of papers that were written by the researcher. The thickness of the line represents the strength of the relationships.

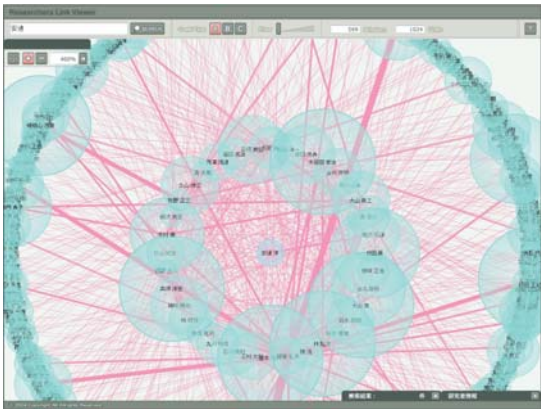

Fig. 3. An example of a local view.

The user can also perform the following operations to find communities interactively.

- Change the researcher of focus by specifying another researcher.
- Move circles to find relationships among particular researchers.
- Change the type of relationships.
- Show the bibliography of the researcher.

3.5 Other Functions

In addition, we implement basic functions in order to facilitate the finding of communities by the user as follows:

- Paper search: The user can search a paper or researcher by specifying author name or paper title. This function helps to understand the meaning within discovered communities and helps to find similar communities.
- Ranking: The user can easily find dense communities by using ranking for co-author or citation. In addition, this function also helps to show the starting point for community mining.

4 Case Study of Community Mining System

In order to demonstrate how the proposed system works, we will discuss the system using actual examples. Figure 4 illustrates the starting screen of the proposed system. From this screen, we can easily choose any of the functions of the proposed system.

From the center of the starting screen, we can access the local view by Flash, as discussed in Figure 3. By clicking the center of the starting screen, a search window is opened. After the user inputs the name of a researcher, whom the user wants to know

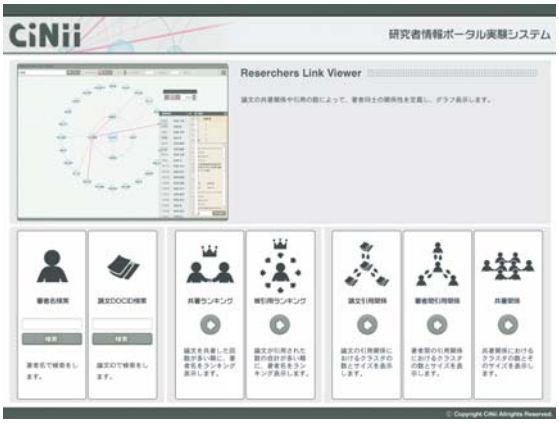

Fig. 4. Starting screen of system.

more about, a visualized screen representing the local communities for the researcher is shown. For example, if we input Takeda (in Japanese) and choose Hideaki Takeda from the Takeda list, the screen shown in Figure 5 will be displayed. The list at the bottom-left corner of the screen is a list of authors. When Hideaki Takeda is selected from this list, a circle denoting Hideaki Takeda is shown at the center of the screen. The circles around Takeda denote researches who have co-authored papers with Takeda. In order to find communities that are built around a specified researcher, the user is able visualize the researchers related to a specified researcher. In addition, the system has a function for eliminating relationships that do not exceed a particular threshold that is specified by user, as well as a function for changing the scale or location of circles in order to more clearly display the desired information. Although Figure 5 illustrates the co-author relationships, the proposed system also has the ability to show the author citation relationships in different colors. The bottom-right window shows the bibliography of the researcher at the center of the screen and is used to determine the field of the researcher.

The two items in the box at bottom-left, shown in Figure 4, are used for searching for authors and papers. The two items in the center are used to show rankings. These rankings include researchers with whom papers have been written and papers that are cited frequently. This function is used for obtaining knowledge for community mining.

The three items in the box at bottom-right, shown in Figure 4, are used for visualizing the graphs for community mining. These items are used to show citation relationships, author citation relationships, and co-author relationships. Since each of these functions is similar, we will show only the case of the author citation relationships. When the user clicks the area for the author citation in Figure 4, the screen shown in Figure 6 is displayed. From this screen, the threshold of the weight of the relationships can be set. In this example, as shown at the center-left of the screen, the weight of the relationships is set to 1. In other words, the graph is created using relationships in which the two researchers have at least one author citation relationship. As a result, we can

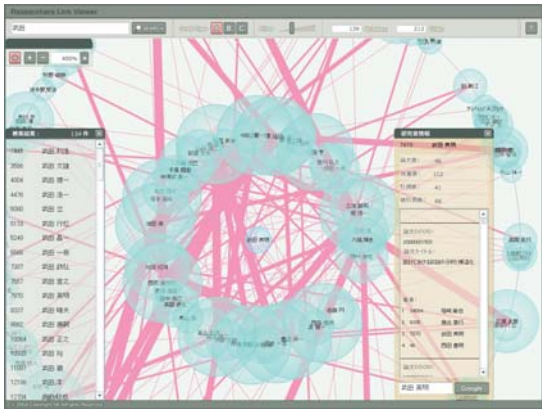

Fig. 5. Local view for searching a particular researcher.

divide the researchers into a number of communities. In this case, we can obtain, for example, a community constructed with 30,536 researchers or a community constructed with 26 researchers. The list of the communities is shown at the bottom in Figure 6. When the user selects a community from the list, the user can view a graph using SVG viewer, as shown in Figure 2. If the user judges the community to be too large to understand easily, the threshold for dividing the communities can be adjusted by the user. If the threshold is increased, weak relationships will be eliminated, and as a result the user will obtain a screen similar to that shown in Figure 6 for the threshold. In this way, the user can interactively use this system to find the desired communities.

5 Conclusions

In the present paper, we propose a novel system for community mining, which requires both a macro view of all researchers and a micro view of individual researchers. We implemented and demonstrated the proposed system, which can be accessed at the website of the following reference [7].

Although the proposed system is functional in its current state, a number of areas for improvement remain. First, the proposed system must be tested by actual users under actual conditions of use. We hope to receive feedback from researchers regarding our system and will make improvements based thereupon. Another area for improvement is how to assist the user with the indexes. In the present study, we do not use an automatic approach for finding communities because of the variety of purpose for such a task. We intend to investigate the support of a user driven community mining system, and also to develop a computer driven community mining system. Finally, we must investigate the seamless integration of global and local views of communities.

Fig. 6. List of communities.

References

1. Barabási, A-L.: LINKED: The New Science of Networks, 2002.
2. National Institute of Informatics, CiNii (Citation Information by NII), http://ci.nii.ac.jp/, 2004.
3. Scientific Literature Digital Library, http://citeseer.ist.psu.edu/, 2004.
4. Freeman, L. C.: Centrality in social networks: Conceptual clarification, Social Networks, Vol. 1, pp. 215-239, 1979.
5. E. Garfield: Citation Indexes for Science, *Science*, Vol. 122, No. 3159, pp.108-111, 1955.
6. Google Scholar, http://scholar.google.com/, 2004.
7. R. Ichise and H. Takeda: Community Mining System, http://irweb.ex.nii.ac.jp/, 2005.
8. M. Kessler: Bibliographic Coupling between Scientific Papers, *American Documentation*, Vol. 14, No. 1, pp. 10-25, 1963.
9. H. Motoda: Active Mining, IOS Press, 2002.
10. MySQL, http://www.mysql.com/, 2004.
11. M. E. J. Newman: Coauthorship Networks and Patterns of Scientific Collaboration, *Proceedings of the National Academy of Sciences of the USA*, Vol. 101, suppl. 1, pp. 5200-5205, 2004.
12. H. Small: Co-citation in the Scientific Literature, *Journal of the American Society of Information Science*, Vol. 24, pp. 265-269, 1973.
13. Scalable Vector Graphics (SVG) 1.1 Specification, http://www.w3.org/TR/SVG/, 2003.