Community Mining Tool using Bibliography Data

Ryutaro Ichise, Hideaki Takeda
National Institute of Informatics
2-1-2 Hitotsubashi Chiyoda-ku
Tokyo, 101-8430, Japan
{ichise,takeda}@nii.ac.jp

Kosuke Ueyama
TRIAX Inc.
4-18-2-203 Takadanobaba, Shinjyuku-ku
Tokyo, 169-0075, Japan
ko@triax.jp

Abstract

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

1 Introduction

As information technologies progress, we can obtain research information faster then before. However, technologies covering a wide area can change just as rapidly. Therefore, all researchers not only must continuously follow new trends of research but also investigate new research topics. When we undertake new research topics, we need to know the research communities of researchers with the same research topic or same interest. As a result, we need an effective community mining tool for finding them.

In order to identify existing research communities, bibliography information is widely used. 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 individual researchers. As a result, realizing that each researcher has an individual research area is difficult. In contrast, CiteSeer [11] and Google Scholar [4] 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 local communities involving an author, they are not suitable for finding research communities close to the author. Moreover, it is not suitable for finding the position of the author within a global research community. In the present paper, we propose a community mining system for researchers. This mining system has both local and global viewpoints. The proposed system will facilitate understanding of researcher communities and will advance new areas of research.

The present paper is organized as follows. In Section 2, we discuss the proposed design for the community mining system. In Section 3, we describe and analyze the bibliography data used in our system. In Section 4, we describe our visualization technologies. In Section 5, we explain the proposed system using a number of examples. Finally, in Section 6, 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 a bibliography, which constitutes a paper list of the researchers and citation information of the papers in the list. In the networks of research, several relationships can be obtained from this information. The relationships in the knowledge domain include co-authorship [10], bibliographic coupling [6], citation [3] and co-citation [12]. In this paper, we use the following three relationships in the knowledge domain to find research communities: 1) co-authorship, 2) citation, 3) author citation.

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

2.2 Community Mining

In this section, we describe the community mining method for the networks of research. We use a policy for community mining borrowed from the active mining approach [7]. The basic concept of active mining is to utilize spiral interactions between the user and the computer. We separate the mining method into two steps. The first step automatically mines communities by computer, and the second step visualizes the information in order to facilitate the user's understanding of the information. After the communities are discovered by computer in the first step, the user can see the result in the second step. If the user is not satisfied with the communities discovered, the user can then specify parameters to further refine the communities. Then, the first step is conducted again with the new settings to discover more communities. As we described above, the proposed method can facilitate the finding of communities through spiral manipulation of the mining results obtained by computer.

Several types of indexes have been proposed for finding communities [2]. In the present study, we use the following three indexes for finding communities: 1) simple weight, 2) maximum flow, 3) closeness. The first index, simple weight, is a measure for weighting arcs: for example, a weight can be applied to represent 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 a 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 communication capacity 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 has a high value. Therefore, this index may be a good measure for finding communities. The third index, closeness, is a measure 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.

2.3 System Architecture

In the present study, we implemented a community mining system using the concepts presented in the previous section. The components of the system are shown in Figure 1. The system has two databases and five program units. The two databases are the CiNii(Citation Information by the National Institute of Informatics) [9] database and an experimental database. The CiNii database will be described in detail later.

The five program units are shown in Figure 1. The database generation unit selects records from the CiNii database and sends them to the database management unit. MySQL [8] is used as the database management unit. If possible, a 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 experimental 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 (Scalable Vector Graphics) Viewer. Then, when a user inputs data from the web browser, the data is sent to the I/O control unit via the Internet. The I/O control unit, which is constructed by a web server that has a CGI(Common Gateway Interface) 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.

3 Data extraction and analysis

3.1 Bibliography Database

In the present study, we use the CiNii database to obtain bibliography information. The database is composed of approximately 320 megabytes of SGML(Standard Generalized Markup Language) data. The CiNii database contains bibliography entries, such as title, author and publication year. The number of data are listed in Table 1. The entries in the CiNii column denote the number of records in the original database. 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 three when three researchers write a paper corroboratively. The Co-author entries denote

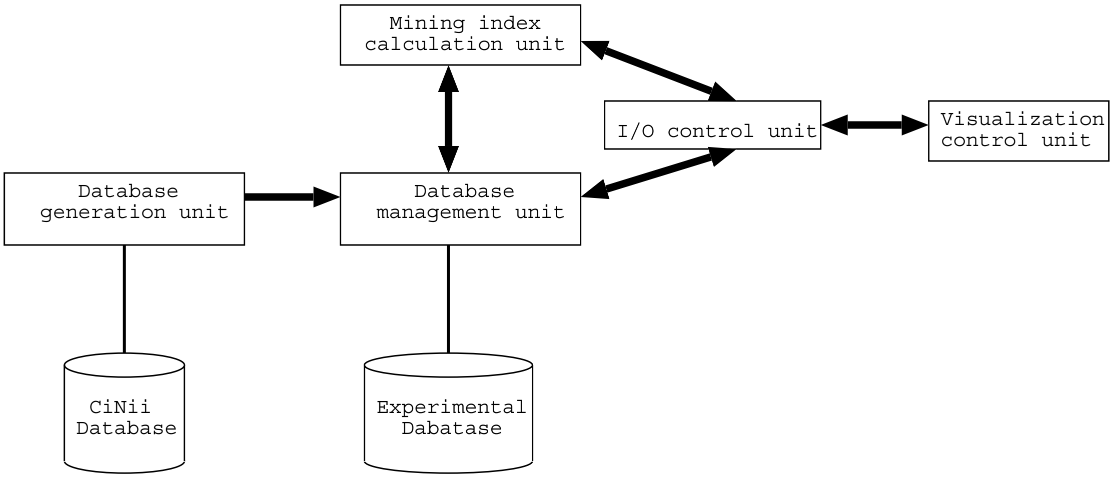


Figure 1. System architecture.

Table 1. Number of data records.

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

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 does not include citation information for all papers. The Citation (Author) entries denote the number of researchers who wrote cited papers, with duplication. In a comparison of the Paper and Researcher entries, the number of researchers is less than the number of papers. In contrast, in a comparison of the Citation (Paper) and Citation (Author) entries, the number of Citation (Author) entries is larger than the number of Citation (Paper) entries, which implies that researchers prefer citing papers by a variety of authors 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.

3.2 Preprocessing of the Database

In order to construct the experimental database, the CiNii data was preprocessed by the database generation unit. First, the database generation unit selected the useful attributes from the CiNii database; most of the attributes, such as the ISSN (International Standard Serial Number), were not relevant to this experiment. In addition, the database generation unit conducted record linkage for the author records. It has heuristics for dividing author records, because the author records in the CiNii database sometimes contain multiple author names in a single author field. For treating such records, the database generation unit divides long author names using special characters, such as $\star$. Also, a number of other small record linkage techniques were used in this stage.

After the preprocessing was completed, the experimental database was created from the CiNii database by the database management unit. The resulting number of records for the experimental database are listed in Table 1. The experimental database contains much less data than the original CiNii database because only the original bibliographies of papers were used. Although the CiNii database contains cited paper information, the information is not complete and contains several mistakes, and was therefore not used in the experimental database.

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,

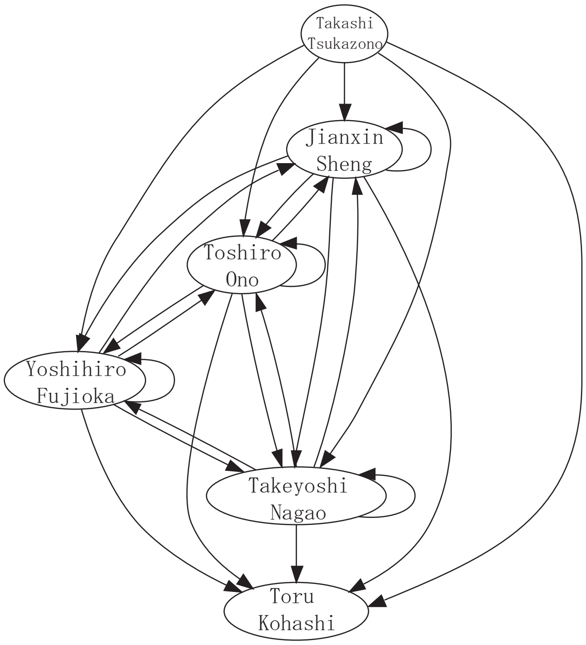


Figure 2. An example of a graph representation by SVG.

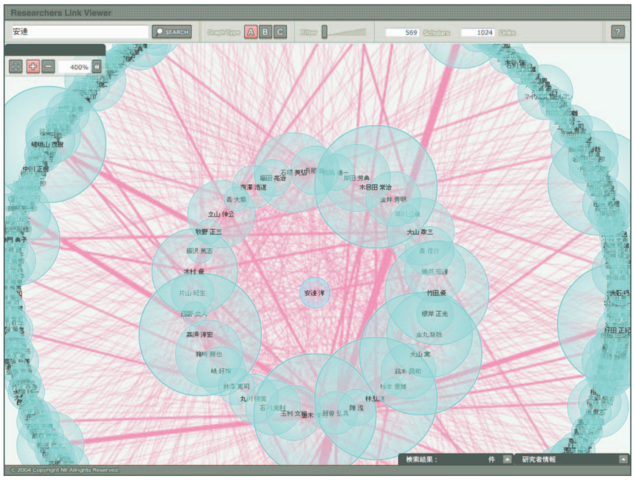


Figure 3. An example of a local view by Flash.

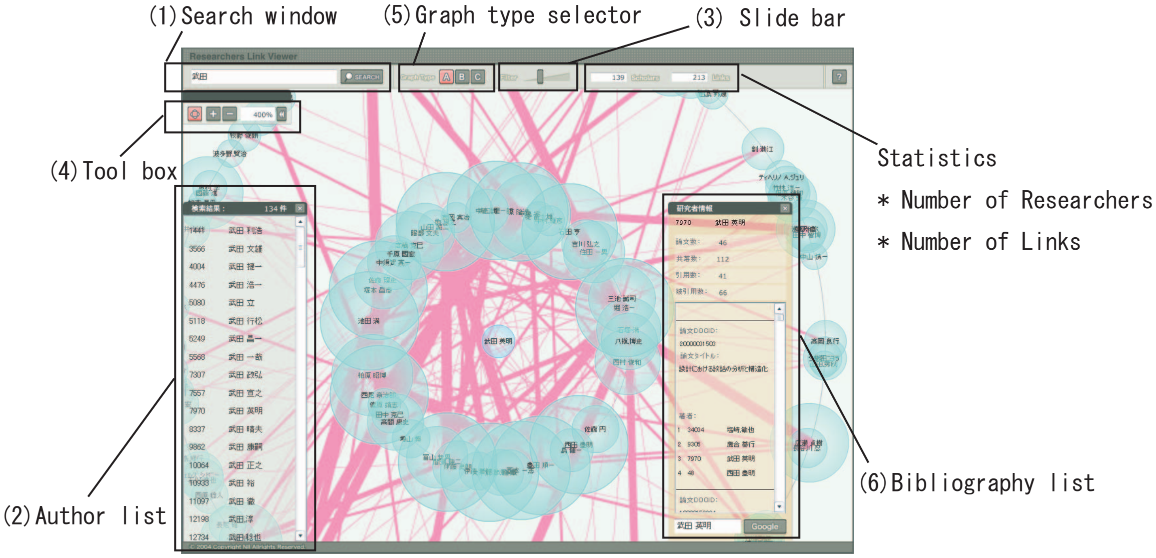
which includes Flash Player and SVG Viewer, as the visualization control unit. The data for this unit is provided by a web server on the Internet. The server is referred to as the I/O control unit. We used two types of visualization data: global visualization by SVG [13] and local visualization by Flash. Global visualization by SVG facilitates the location of global communities by showing global relationships on a graph, whereas local visualization by Flash facilitates the location of local communities by showing the local relationships near the individual. Both of these methods are discussed below.

4.1 Global Visualization

We first discuss the graph visualization by SVG. As discussed in Section 2.1, representing the relationships between communities in the form of a graph facilitates the finding of communities. SVG is an XML(eXtensible Markup Language) format for representing graphs. The SVG file can be visualized by an SVG viewer. An example of a 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, because most of the nodes would be connected [1]. As a result, the graph could be very large and the user might not be able to find the communities. In order to discern communities from a large graph, the system needs a function that shows the parts of the graph that 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 for the I/O unit to specify an index to match the intentions 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 a simple weight.

4.2 Local Visualization

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. We propose visualizing the local communities to which a specified researcher belongs in order to facilitate the refinement of communities globally. Locally, a specified researcher is located at the center of a field, and related researchers are placed around the circumference of the field. 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 the co-author and author citations, 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 initially specified researcher. The surrounding circles represent researchers who are related to the specified researcher. 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 pa-



武田

139 Scholars 213 Links

検索結果： 134 件

1441	武田 利浩
3566	武田 文雄
4004	武田 捷一
4476	武田 浩一
5080	武田 立
5118	武田 行松
5249	武田 昌一
5568	武田 一哉
7307	武田 政弘
7557	武田 宣之
7970	武田 英明
8337	武田 晴夫
9862	武田 康嗣
10064	武田 正之
10933	武田 裕
11097	武田 徹
12198	武田 淳
12734	武田 稔也

© 2004 Copyright NII Allrights Reserved.

研究者情報

7970 武田 英明

論文数： 46

共著数： 112

引用数： 41

被引用数： 66

論文DOCID：
20000031503
論文タイトル：
設計における談話の分析と構造化

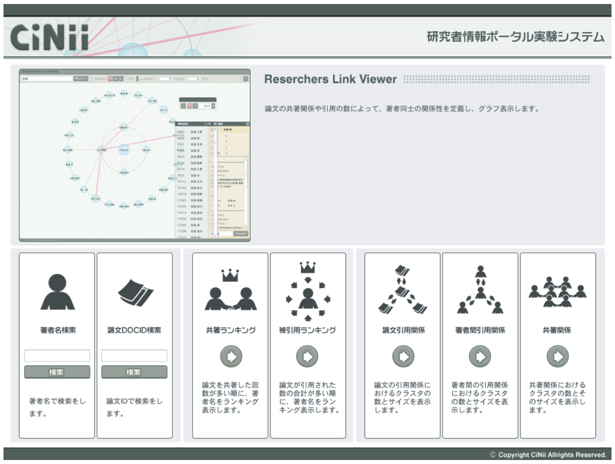
著者：

1	34034	塩崎 敏也
2	9305	鷹合 基行
3	7970	武田 英明
4	48	西田 豊明

論文DOCID：

武田 英明 Google

Figure 5. Local view for searching a particular researcher.



論文の共著関係や引用の数によって、著者同士の関係性を定義し、グラフ表示します。

著者名で検索をします。

論文IDで検索をします。

論文を共著した回数が多い順に、著者名をランキング表示します。

論文が引用された数の合計が多い順に、著者名をランキング表示します。

論文の引用関係におけるクラスタの数とサイズを表示します。

著者間の引用関係におけるクラスタの数とサイズを表示します。

共著関係におけるクラスタの数とそのサイズを表示します。

© Copyright CiNii Allrights Reserved.

Figure 4. Starting screen of the proposed system.

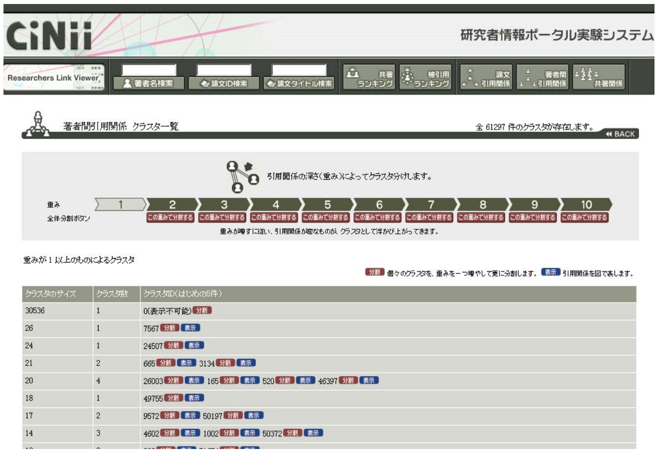
pers that were written by the researcher. The thickness of the line represents the strength of the relationships.

5 Case Study of the 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 shown in Figure 3. By clicking the center of the starting screen, a window is opened. Let us explain by an example, shown in Figure 5. First, the user inputs the name of a researcher in the search window (1). In this example, we input Takeda (in Japanese). Then, a list of authors named Takeda is shown in the author list (2). 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 researchers who have coauthored papers with Takeda. In order to find communities that are built around this specified researcher, the user can view related researchers. In addition, the system has a function for eliminating relationships that do not meet a user-specified threshold. The user selects the threshold using the slide bar (3). Then, to more clearly display the resulting information, the user can change the scale or location of circles by using the tool box (4) . In addition to the coauthor relationships shown in Figure 5, the proposed system can also show the author citation relationships in different colors by specifying the graph type (5). The bottom-right window (6) 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 the bottom-left in Figure 4 are used for searching 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 additional knowledge for community mining.



著者間引用関係 クラスター一覧

全 61297 件のクラスタが存在します。

引用関係の深さ（重み）によってクラスタ分けれします。

重みが増すに従い、引用関係が密なものが、クラスタとして浮かび上がってきます。

重みが 1 以上のものによるクラスタ

分割 個々のクラスタを、重みを一つ増やして更に分割します。 表示 引用関係を図で表します。

クラスタのサイズ	クラスタ数	クラスタID（はじめの5件）
30536	1	0(表示不可能) 分割
26	1	7567 分割 表示
24	1	24507 分割 表示
21	2	665 分割 表示 3134 分割 表示
20	4	26003 分割 表示 165 分割 表示 520 分割 表示 46397 分割 表示
18	1	49755 分割 表示
17	2	9572 分割 表示 50197 分割 表示
14	3	4602 分割 表示 1002 分割 表示 50372 分割 表示

Figure 6. List of communities.

The three items in the box at the bottom-right in Figure 4 are used for visualizing the graphs for community mining. These items show citation relationships, author citation relationships, and co-author relationships. Since each of these functions is similar, we will show only an author citation relationship. 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, the weight of the relationships is set to 1, as shown at the center-left of the screen. 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 divide the researchers into a number of communities. For example, we can obtain a community constructed with 30,536 researchers or a community constructed with 26 researchers. The list of communities is shown at the bottom in Figure 6. The user can select a community from the list and then view a graph using the SVG viewer, as shown in Figure 2. If the community appears too large to understand easily, the user can adjust the threshold for dividing the communities. If the threshold is increased, weak relationships will be eliminated, and the user would obtain a screen similar to that shown in Figure 6. In this way, the user can interactively use this system to find the desired communities.

6 Conclusions

In this paper, we propose a novel system for community mining. Our system presents both a global view of all researchers and a local view of individual researchers. We implemented and demonstrated the proposed system, which can be accessed at the web site stated in reference [5]. From our case study, we can conclude that the system provides easy understanding of the global and local relationships between researchers. Therefore, we think our system can be used not only as a community mining system but also as a new analytical tool for research. On the other hand, the relationships we used in this paper have several meanings, because the links not only represent the same interests or topics, but also the references for broad areas of research. This produces complicated graphs in some cases. Therefore, in future research, we plan to analyze our assumptions about community mining relationships.

Also, 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 conditions of actual 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 options for such a task. We intend to investigate the support of a user-driven community mining system, but also to develop a computer-driven community mining system. Finally, we must investigate the seamless integration of the global and local views of communities.

References

[1] A.-L. Barabási. *LINKED: The New Science of Networks*. Perseus Books Group, 2002.

[2] L. C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1:215–239, 1979.

[3] E. Garfield. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159):108–111, 1955.

[4] Google scholar, 2004. http://scholar.google.com/.

[5] R. Ichise and H. Takeda. Community mining system, 2005. http://irweb.ex.nii.ac.jp/.

[6] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10–25, 1963.

[7] H. Motoda. *Active Mining*. IOS Press, 2002.

[8] Mysql, 2004. http://www.mysql.com/.

[9] Cinii (citation information by nii). National Institute of Informatics, 2004. http://ci.nii.ac.jp/.

[10] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the USA*, 101(suppl. 1):5200–5205, 2004.

[11] Scientific literature digital library, 2004. http://citeseer.ist.psu.edu/.

[12] H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society of Information Science*, 24:265–269, 1973.

[13] Scalable vector graphics (svg) 1.1 specification. W3C SVG Working Group, 2003. http://www.w3.org/TR/SVG/.