A mining method of communities keeping tacit knowledge

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

Satoshi Kouno, Taichi Muraki
TriAx Corporation
4-29-3 Yoyogi, Shibuya-ku,Tokyo 151-0053, Japan
{kouno,muraki}@triax.jp

Abstract

The research community plays a very important role in holding valuable scientific knowledge. The authors propose a community mining system which helps to find communities of researchers by using bibliography data. The key features of our method are a network model of papers and a word assignment technique for the communities obtained. We implemented the proposed method in a graphical computer system. In this paper, we show how research communities are found using our system. Also, we evaluate the performance of the proposed method using experiments with real world data. The results demonstrate that our system can find appropriately sized research communities for a particular scientific field.

1. Introduction

Modern life has become fundamentally supported by various technical systems, ranging from traditional systems like roads and bridges to highly technological systems like nuclear power plants, all of which are built and maintained using scientific and technical knowledge. Sustaining modern life is dependent on the maintenance of these systems. Accordingly, to avoid the catastrophic failure of such systems, preserving this scientific and technical knowledge is vital.

Scientific and technical knowledge is kept explicitly by means such as published papers, but that is not the only way. The crucial components of knowledge are kept implicitly among communities of scientists and engineers. Not only papers themselves but also communities behind the papers play an important role in keeping knowledge from generation to generation. As a result, we need an effective community mining method for finding them. In order to find research communities, we usually use bibliography information. The methods include co-citation analysis [10, 3] and bibliographic coupling [7]. Although these methods are very useful for analyzing research topics from the global viewpoint of all bibliography data, we cannot always understand what the discovered communities represent. CiteSeer [9] 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. Börner et al. [2] proposes the use of co-author networks to find research communities by using weighted graphs. Their system uses heuristics to separate communities without interaction. Ichise et al. [6] proposed a community mining method based on the interaction of users. Although the proposed system of Ichise et al. supports community mining for both a global view and a local view with several mining indexes, it does not identify the research topics of the communities obtained. In this paper, we propose a visual system to discover research communities with identified topics.

This paper is organized as follows. In Section 2, we discuss our proposed method for research community mining. Section 3 describes the proposed system using examples. In Section 4, we describe the experimental evaluation of our method and discuss the results. Finally, in Section 5, we present our conclusions.

Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)
0-7695-2702-7/06 $20.00

2. Research Community Mining

2.1. Representation of Research Community

Although several network models using bibliographies to represent research communities have been proposed [6], in this paper we focus on the co-author relationships of a research paper to find the research communities. First, we assume a simple paper model. This model consists of keywords and author names. In this case, we can consider an author's work on a research topic by noting the keywords. As a result, authors who write a paper collaboratively share the same interest, represented by the keywords. If we consider the authors as nodes and the keywords as edges, we can represent the bibliography information as researcher networks.

Let us explain our model using an example. Assume that we have two papers, as shown on each side of Figure 1. One was written by two authors, A and B, and has two keywords, W_1 and W_2. Another paper was written by three authors, A, C and D, and has the keyword W_3. We can compose graphs of the authors and edges from the two papers, as shown in Figure 1. Then, the joint representation generated from the two bibliographies of the two papers is shown in the center of Figure 1.

We can thus obtain a labeled graph from the bibliography data with our modeling. The next challenge is then how to identify research communities from this graph. We define a research community as a cluster that is densely connected by the same research interest or topic. Therefore, the research communities we want to obtain are clusters that have their edges labeled by the same keywords. Since our network model provides the research topics (edges), we can obtain the research communities by eliminating the edges of no interest to the system user. In other words, after the user specifies the research topics, most of the edges that are not related to the specified topics can be deleted. This process reveals the research communities of interest. For example, when the user specifies W_3 for the networks in Figure 1, the edges of W_1 and W_2 are eliminated. As a result, researcher B is isolated from the graph and we can find the research community consisting of researchers A, C and D.

2.2. Keyword Assignments for Communities

Since the clusters obtained by our method are only connected by user-specified relationships, we can consider each cluster to represent a research community. However, each cluster does not have its own properties or distinguishing characteristics. In other words, if the user does not have sufficient knowledge about the researchers, the user may not understand the significance of the communities because there is no information related to them. In order to resolve this problem, we propose a method of assigning keywords for each obtained community.

In our paper model, the papers written by the authors in each community have keywords. If some words appear often in such papers, we can consider these words to be a property of the community. However, if we simply counted the occurrences of the keywords in these papers, the relationships between keywords would be lost. In order to avoid this problem, we consider frequently used keywords to be units of the papers. The algorithm is as follows:

1. Select papers, which are written by the authors in the community.
2. Analyze the selected papers using the Apriori algorithm [1]. In this process, the keywords in a paper are treated as items, and the papers are treated as transactions.

As a result, we can obtain word pairs for each community. We can then assign these word pairs as the properties of the community.

3. Community Mining System

3.1. System Design

We implemented the proposed method in an actual system using Java. In order to demonstrate how the proposed system works, we will discuss it using actual examples. Figure 2 depicts the proposed system. First, the user inputs a research field of interest into the search window (1). At this time, the user can specify a parameter for use in selecting communities for the specified search term. In our current system, we use the Apriori algorithm for choosing authors of communities as well. The parameter is used in the Apriori algorithm as a support value for filtering author data and helping to identify appropriately sized communities. This parameter can be adjusted using a slide bar located under the search box. The research communities related to the search term are then shown graphically (2). When a particular community is selected from the graphs, a keyword list for the community is shown in (3). In the event that a user is not satisfied with the size of the research communities shown, they can adjust the graph size using the slide bar at the bottom left of the screen in Figure 2. The system then changes the threshold for the edges to be displayed.

3.2. Case Study of the Community Mining System

In this section, we will demonstrate how to locate communities for "robotics" by using the community mining sys-

Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)
0-7695-2702-7/06 $20.00 © 2006

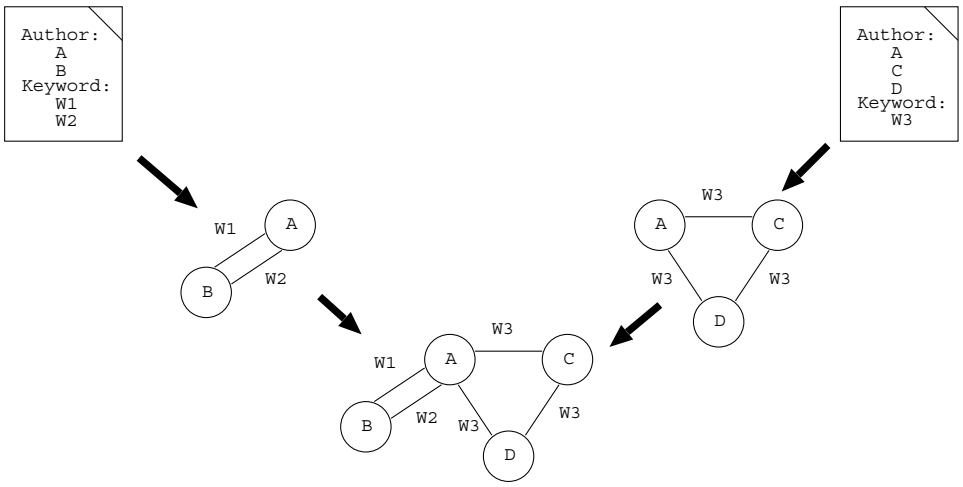

Figure 1. Network model of researchers.

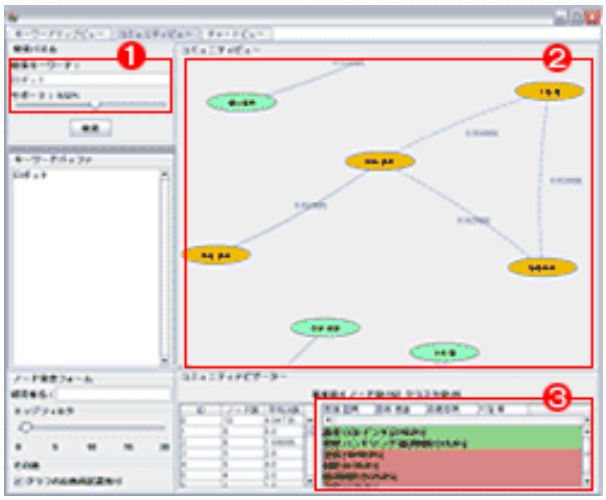

Figure 2. Screenshot of community mining system.

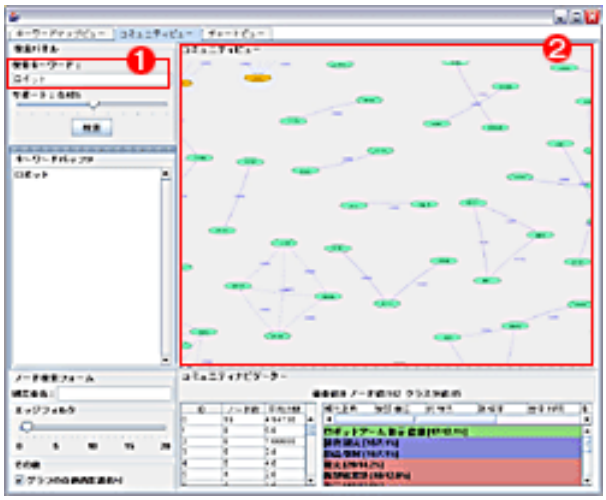

Figure 3. A search example for "robotics".

tem. First, the user inputs "robotics" in the search window (1) in Figure 3. The research communities are then displayed using graphs (2) in Figure 3. As can be seen from this example, many "robotics" communities can be obtained in this manner, including "reinforcement learning" or "sensing". In order to determine the research field for each community, the user can browse community keywords by specifying communities (3), (4) as shown in Figure 4. When the user locates a community that they are interested in, they can browse the publication list for a specified researcher. They can also browse personal information for the specified researcher.

4. Experiments

4.1. Bibliography Data

In order to evaluate our method, we conducted experiments using actual bibliography data. In this study, we used part of the CiNii database [8] to obtain bibliographic information. The database consists of SGML (Standard Generalized Markup Language) data. The CiNii database contains bibliography entries that include title, author, and publication year. The system data statistics are listed in Table 1. The paper and researcher entries denote the total number of paper records and the number of records for researcher in the system, respectively. The author record denotes the

Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)
0-7695-2702-7/06 $20.00 © 2006

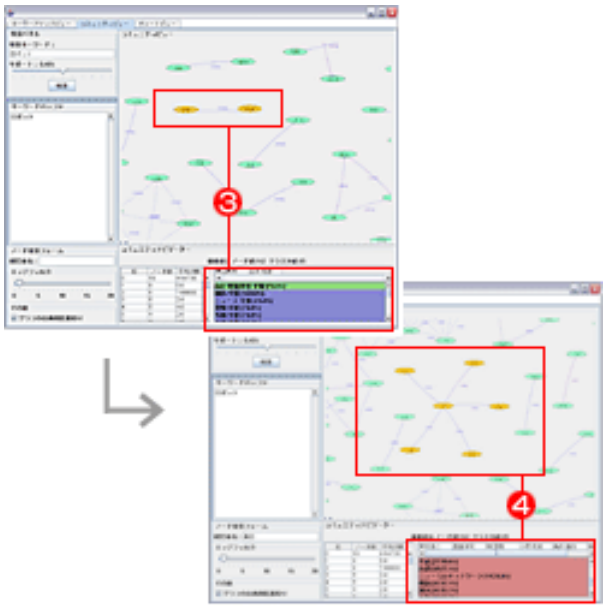

Figure 4. Keyword listing for specified communities.

Table 1. Number of data records.

	Experimental Database ($\times 1,000$)
Paper	544
Researcher	93
Author	358
Co-author	1,038
Keyword	40
Label	1,087

number of authors for each paper. For example, the record's value is three when three researchers write a paper collaboratively. The co-author entries denote the number of combinations of authors for a paper. For example, when a paper has four co-authors, it is counted as ${}_4C_2 = 6$ for the paper. The keyword entries denote the number of kinds of keywords. In this study, we used the words contained in the title as keywords. The label entries denote the number of keyword labels for each paper. For example, the record is three when the paper title has three keywords.

4.2. Experimental Results

The co-author network tends to be characterized as having large clusters. In fact, the network constructed by all the bibliography data consists of a few large clusters and numerous small clusters [5].

Table 2. Number of papers and authors for each term.

Words	Papers	Authors
entropy	38	79
vector quantization	85	138
pattern recognition	89	227
perpendicular magnetic recording	119	141
robot	372	778
genetic algorithm	433	738
chaos	444	529
algorithm	1584	2491

However, our method was capable of successfully splitting a large cluster into readable research communities. We used eight community search terms to evaluate our system, as follows:

- entropy
- vector quantization
- pattern recognition
- perpendicular magnetic recording
- robot
- genetic algorithm
- chaos
- algorithm

The number of papers and authors for each term are shown in Table 2.

We counted the number of nodes and clusters retrieved by our system for each term (Figure 5). The horizontal axis denotes the number of nodes and the vertical axis shows the number of communities. As can be seen from Figure 5, our method successfully identifies communities of a readable size. Also, communities related to particular topics of interest to the user can be mined using our method.

Next, in order to evaluate our community search method, we investigated a number of communities using different support values. Remember that the support value is a parameter used to refine the mined communities, and can be changed using the slide bar at the top left of Figure 2. The results are shown in Table 3. As you can see from the results, the support values are helpful in finding core communities. For example, for words with a small number of related authors such as "entropy" and "vector quantization", all authors are included by our community mining system. Conversely, for words with a large number of associated authors such as "robot" and "algorithm", the system selects

Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)
0-7695-2702-7/06 $20.00 © 2006

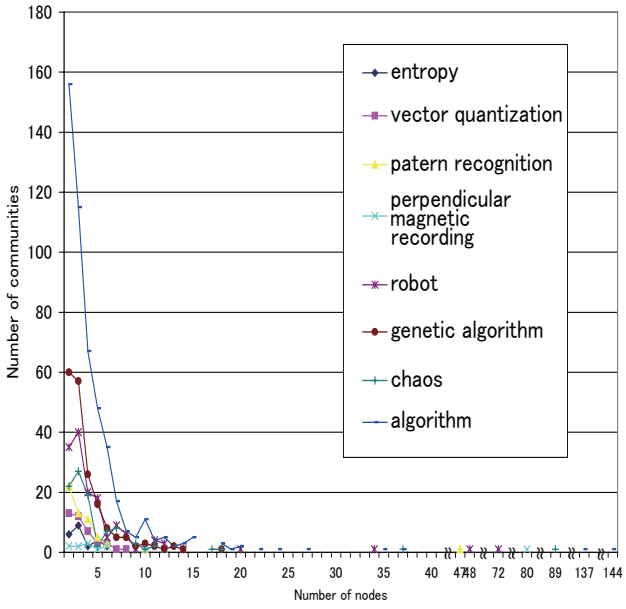

Figure 5. Number of discovered communities.

authors according to the support value. This enables the user to find appropriately sized communities for specified search terms.

4.3. Discussion

One of the main differences between the conventional community mining systems Ver.1 [6] and the system proposed here is the use of keywords. Version 1 does not use keywords for mining communities. This function enables us to find a community without an extensive knowledge of the research domains the user wants to mine. Unfortunately, the current system is limited by the level of confidence in the keywords, since we used words in the title as keywords. Since most of the records in the CiNii database do not include keywords or abstracts, we decided to employ this approach. However, this method has the disadvantage of not always being able to assign the appropriate keywords for papers. Future work will focus on the development of a method for assigning keywords for such records.

5. Conclusion

In this paper, we proposed a visual system for research community mining. The key feature of our mining method

Table 3. Relationship between support value and number of authors.

Words	Support value				
	0.2	0.4	0.6	0.8	1.0
entropy	79	79	79	79	79
vector quantization	138	138	138	138	138
pattern recognition	227	227	227	227	227
perpendicular magnetic recording	141	141	141	141	78
robot	778	210	75	75	30
genetic algorithm	738	237	88	49	26
chaos	529	208	106	64	48
algorithm	174	42	23	10	6

is the modeling of papers and researchers. This modeling enables us to eliminate the edges of large clusters. In addition, the modeling can also help to retrieve communities for particular topics. We also employ a method for assign words to separate clusters. We implemented our method and showed how to investigate bibliographic data with our system. The experimental results show that the performance of our method has considerable potential for application.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 Sept. 1994.

[2] K. Börner, L. Dall'Asta, W. Ke, and A. Vespignani. Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity*, 10(4):58–67, 2005.

[3] C. Chen and R. J. Paul. Visualizing a knowledge domain's intellectual structure. *Computer*, 34(3):65–71, 2001.

[4] Google scholar, 2006. http://scholar.google.com/.

[5] R. Ichise, H. Takeda, and T. Muraki. A discovery method of research communities. In *Proceedings of Adaptation in Artificial Biological Systems*, volume 3, pages 128–131, 2006.

[6] R. Ichise, H. Takeda, and K. Ueyama. Community mining tool using bibliography data. In *Proceedings of the 9th International Conference on Information Visualization*, 2005.

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

[8] CiNii (Citation Information by NII). National Institute of Informatics, 2006. http://ci.nii.ac.jp/.

[9] Scientific literature digital library, 2006. http://citeseer.ist.psu.edu/.

[10] 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.

Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)
0-7695-2702-7/06 $20.00 © 2006