Acquisition of Scientific Literatures based on Citation-reason Visualization

Dongli Han[1], Hiroshi Koide[2] and Ayato Inoue[2]
[1]Department of Information Science, College of Humanities and Sciences, Nihon University, Sakurajosui 3-25-40, Setagaya-ku, 156-8550, Tokyo, Japan
[2]The Graduate School of Integrated Basic Sciences, Nihon University, Sakurajosui 3-25-40, Setagaya-ku, 156-8550, Tokyo, Japan

Keywords: Paper Acquisition, Citation-reason, Machine Learning, Visualization.

Abstract: When carrying out scientific research, the first step is to acquire relevant papers. It is easy to grab vast numbers of papers by inputting a keyword into a digital library or an online search engine. However, reading all the retrieved papers to find the most relevant ones is agonizingly time-consuming. Previous works have tried to improve paper search by clustering papers with their mutual similarity based on reference relations, including limited use of the type of citation (e.g. providing background vs. using specific method or data). However, previously proposed methods only classify or organize the papers from one point of view, and hence not flexible enough for user or context-specific demands. Moreover, none of the previous works has built a practical system based on a paper database. In this paper, we first establish a paper database from an open-access paper source, then use machine learning to automatically predict the reason for each citation between papers, and finally visualize the resulting information in an application system to help users more efficiently find the papers relevant to their personal uses. User studies employing the system show the effectiveness of our approach.

SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

It is essential to conduct a bibliographic survey and obtain a set of relevant papers before carrying out scientific research in a concerned field. Grabbing a great number of papers from a digital library or an online search engine by inputting a keyword is easy, whereas reading all the obtained papers in order to find the most appropriate ones is agonizingly time-consuming.

Some previous works try to cope with this problem by calculating textual similarities between papers (Baeza-Yates and Ribeiro-Neto, 1999; Dobashi et al., 2003). In these works, the authors generally focus on the keywords contained in each paper and try to estimate how close two papers might be through the common keywords. This strategy can be effective when a researcher wants to find a rough set of related works dealing with a certain topic. However, would it help at all if we need to find papers employing the same theoretical model or using a different experimental data set? In these cases, it is hard to believe that similarity-based approaches would work effectively.

Reference-relation based approaches come from the observation that two papers are probably related to each other regarding method, data, evaluation or any other aspects, provided that one paper cites another. Based on this idea, a number of works have been carried out using either coupling or co-citation (Kessler, 1963; Small, 1973; Miyadera et al., 2004; Yoon et al., 2010; Jeh and Widom, 2002). Coupling-based approaches attempt to compute the similarity between two papers based on the number of papers both of them reference, whereas co-citation based approaches calculate the similarity between two papers based on the number of papers that reference both of them. Another work in this direction has enhanced the reference-relation based approach furthermore by incorporating citation types (Nanba and Okumura, 1999; Nanba et al., 2001). They divide all the citations into 3 types (Type B: basis, Type C: problem presentation, and Type O: others), and cluster the papers that are citing the same paper with the same citation type from a primitive paper

Han, D., Koide, H. and Inoue, A.
Acquisition of Scientific Literatures based on Citation-reason Visualization.
DOI: 10.5220/0005693801230130
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 2: IVAPP*, pages 125-132
ISBN: 978-989-758-175-5
Copyright © 2016 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

set. This method is more efficient than basic reference-relation approaches described above, and has tried to solve the paper-acquisition problem from a new point of view.

However, none of the previous methods could help answer the questions we have raised at the end of the second paragraph. A paper may cite another paper for many reasons (Teufel et al., 2006). For example, a citation may be used to provide general background information, to justify an approach, to describe alternate or competing methods, to define terms, or to refer to data or methods used. In this paper, we make use of the citation-reason to help focus the search for relevant papers.

In section 2, we introduce some major differences between our citation-reason analysis schema and those of previous works. We then describe the process of establishing our paper database in Section 3. Section 4 presents our machine-learning based method for predicting the citation-reason between papers. Section 5 describes the visualization system we have built based on our idea, and an evaluation using the system. Section 6 gives some discussions and the conclusion.

2 CITATION-REASON ANALYSIS

Citation-reason analysis has been a popular research field in bibliometrics and sociology since the 1960's. Many studies have been made to identify the citation-reason between two papers by handcraft (Garfield, 1979; Weinstock, 1971; Moravcsik and Poovanalingan, 1975; Chubin and Moitra, 1975; Spiegel-Rosing, 1977; Oppenheim and Susan, 1978). Since the early 2000's, researchers in computational linguistics have been trying to automate this process (Garzone and Robert, 2000; Pham and Hoffmann, 2003). However, their methods are generally rule-based, which implies a high cost of development and hence low expandability. In recent work, both Radoulov (Radoulov, 2008) and Teufel (Teufel et al., 2006; Teufel, 2010) have proved the effectiveness of machine learning in automated citation-reason analysis. Citation-reason analysis has been carried out with various purposes, but none of the previous works has been directed towards paper acquisition or reorganization as we have been considering here. We believe this has been the most important contribution we have made in this paper.

Besides, there exist some differences between our machine learning approach and those of previous works. In order to conduct fast yet efficient paper classification, we need a set of classification categories that is neither too large to conduct effective machine learning, nor too small to make the classification meaningless. Based on Teufel's schema, we remove three categories which are difficult to distinguish from other categories, and redefine the remaining nine categories in this paper. Table 2 in Section 4 shows the categories and their definitions. Other differences include the features used for machine learning, scope determination for extracting the citation contexts, and the scale of training corpus. Above all, we have established a much bigger and more extendable database than most of the previous work to realize the practical use of our approach in paper acquisition. We will present more detailed descriptions of these aspects throughout Section 3, 4, and 5.

3 OUR PAPER DATABASE

We need a scientific-paper collection to generate the training data for machine learning, and evaluate the effectiveness of our approach as well. We could, in theory, employ a dataset that has been used in any of the previous works and is reusable. Unfortunately, most of the data sets are on a small scale, merely containing hundreds of citation instances at a maximum and come from a closed collection of unidentifiable papers. These disadvantages have made the previous efforts lack expandability and reproducibility, and therefore appear ad-hoc. In this paper, we establish our own paper collection using a widely accessible paper corpus with specified description of the construction process, so that our data set could be rebuilt more easily, and our system could be reused by other. In the rest of this section, we first introduce our data source, then describe the process of generating a database from it, and finally offer some discussion of the resulting database.

3.1 Data Source

We used the annual conference proceedings of the ANLPJ (The Association for Natural Language Processing in Japan) as the data source for our paper database. This conference is an annual event of ANLPJ containing hundreds of manuscripts from researchers all over Japan, and sometimes overseas. Proceedings are published in CD-ROM since 2004 and open to the public via Internet for free since 2010 (paper proceedings were published prior to 2003) (http://www.anlp.jp/guide/nenji.html). Both the accessibility and the total number of papers satisfy our requirement in establishing a paper database from a single scientific area.

We take all the Japanese papers published in the conference proceedings since 2004 as the data source, and call them citing papers hereafter. A preliminary investigation of paper references from a randomly extracted subset of the data shows that scientific papers published in the following five paper collections are most frequently referred or cited by citing papers.

- annual conference proceedings of ANLPJ
- Journal of Natural Language Processing
- IPSJ SIG Technical Reports
- IPSJ Transactions
- IEICE Transactions

Bibliographic information of all papers that have been published in the above five paper collections satisfying the following conditions are extracted using J-GLOBAL (http://jglobal.jst.go.jp/) and CiNii (http://ci.nii.ac.jp/) manually. We call these papers cited papers in the rest of this paper.

- being cited by one or more citing papers
- published after 2000
- written in Japanese

3.2 Database Generation

Two kinds of information are extracted from the citing papers and cited papers, and stored into the database: Macro information and Micro information. The former indicates the meta information of the papers themselves, and the latter includes the textual information inside and around each citation location. Table 1 shows the specific fields.

In our database, a unique number called paper No. is assigned to each citing or cited paper. Papers in ANLPJ have their own distinctive numbering system, based on which we have easily generated their Paper No. However, papers coming from other paper collections do not share a common numbering system, and therefore are numbered using CiNii. Other fields in the Macro information mainly include some bibliographic information about the papers. The last field i.e., the total number of citing paper's component sections, is considered as a potential feature for use in machine learning though we haven't yet used it so far.

Micro information is composed of a number of attributes related to the context where the authors refer to a cited paper within a citing paper.

Table 1: Fields in the paper database.

	Specific fields (possible choices)
Macro information	citing paper No.
	cited paper No.
	citing paper's title
	cited paper's title
	citing paper's authors
	cited paper's authors
	cited paper's publication year
	total number of citing paper's component sections
Micro information	citing sentence
	preceding sentence
	succeeding sentence
	citation location (main body, footnote, or headline)
	citing text when citation location is not main body
	itemization (NULL, internal, anterior, or posterior)
	citation-reason (9 categories)
	section nubmer

The first three fields indicate the scope of the context based on which the computer predicts the citation-reason with machine learning. The citing sentence is the sentence within which a reference number appears, and the preceding sentence and succeeding sentence stand for the sentences around it. In case multiple reference numbers appear in one citing sentence, multiple records are generated in the database with the same citing paper No. and contexts but different cited paper No. On the other hand, authors might want to refer to the same cited paper more than once within the same citing paper. In this case, multiple records with the same citing paper and cited paper but different citation contexts are saved in the database as well. The citation location indicates the type of the area where the citation occurs within the paper: main body, footnote, or headline. In cases where citation locations are those other than main body, the value *null* is stored in the first three fields, and the entire footnote or the headline is extracted from the paper and stored in the field of citing text when citation location is not main body. Itemization indicates the relative position between the citing sentence and itemizations. It is assigned with *null* if the citing sentence is neither situated within nor adjacent to an itemization, otherwise one of the other three values is used according to the observed relative position. The citation-reason indicates the reason why the authors are referring to the cited paper in the citing paper. Nine possible values are used here representing nine specific categories as shown in Table 2.

Based on the above descriptions, we generated 4600 records in the database by hand. Furthermore, the second author of this paper and three collaborators extracted 900 records and annotated a citation-reason to each of them with repeated discussion and careful analysis on both the citing paper and the cited paper. This process is extremely difficult and unexpectedly time-consuming due to the ambiguous borderlines between the citation-reason categories, especially for untrained first-time annotators. The 900 records are used as training data of machine learning for predicting citation-reasons as to be described in Section 4.

3.3 Discussions

We have created a scientific-paper database from which we are able to generate the training data for machine learning. Most information in the database either contributes to the machine learning process directly, or helps annotators more efficiently locate and analyze an original paper in the data source. The rest is expected to make it easier to maintain the whole database by programming.

Our database contains 4600 citation instances extracted from all the papers digitally published in ANLPJ from 2004 to 2012. Compared with the data sets used in previous citation analysis, our database has four advantages: a larger scale, a longer time span, a wide openness, and a persistent updatability. The last advantage comes from the annual renewal of ANLPJ and will definitely benefit the extendability of the database and furthermore improve the performance of the machine learning process.

On the other hand, concerns about the database include the lack of papers published prior to 2003, and the cost involved in generating new records manually from ANLPJ hereafter. We might need some semi-automated process to solve these problems to make the database more comprehensive in the future.

4 CITATION-REASON PREDICTION

In this section, we describe the method proposed for predicting citation-reasons from a citing paper towards a cited paper. We take a machine-learning based approach using data extracted from the paper database described in Section 3. Here, we first introduce our citation-reason categories, then give a systematic descriptions on our machine-learning based citation-reason analysis, and finally specify our evaluation process.

4.1 Citation-reason Categories

As stated in Section 2, various categories have been employed in different works. Too many categories generally need more manually annotated training data and tend to cause confusion among similar classes, whereas too few categories will not help users solve their problems in organizing or classifying papers effectively. Teufel’s study seems to be the most thorough one among all the other works in citation-reason category definition (Teufel et al., 2006; Teufel, 2010). Following her idea, we remove three categories from Teufel’s schema which are difficult to be distinguished from other categories by non-professional annotators, and employ the remaining nine categories as elaborated in Table 2.

Table 2: Citation-reason categories.

Citation-reason category	Definition
Weak	describing general disadvantages of a cited paper
Coco	describing disadvantages of a cited paper in comparison with the citing paper
CocoGM	comparing with a cited paper in aim or method
CocoRo	comparing with a cited paper in result
PBas	taking a cited paper as a starting point
PUse	using tools, algorithms, or data described in a cited paper
PModi	modifying and using a tool, algorithm, or data described in a cited paper
PMot	demonstrating validity of the citing paper through a cited paper
Neut	describing a cited paper neutrally

4.2 Machine-learning based Prediction

We extract citation contexts i.e., citing sentence, preceding sentence, and succeeding sentence to compose the training dataset from 900 records which have been assigned with citation-reasons as described in Section 3. Then a Japanese morphological analyzer Mecab (http://mecab.sourceforge.net/) is applied to these citation contexts to generate unigram and bigram data respectively as features for machine learning. During this process, only nouns, verbs, and adjectives are extracted to generate each n-gram data.

Then we employ a Naïve Bayes classifier as the basic machine learning method to generate a classifier from the training data. The reason we use a Naïve Bayes classifier lies in its theoretical naiveness and its simplicity of implementation. We believe that we could obtain even better results with other machine-learning methods if we can succeed with the simplest approach first.

Eq. (1) shows the basic idea of a Naïve Bayes classifier, where *P(con)*, *P(cat)*, *P(cat|con)*, and *P(con|cat)* indicate the probability of a context, the probability of a category, the probability of a citation-reason category provided with a particular context, and the probability of a context provided with a particular citation-reason category.

$$P(cat \mid con) = \frac{P(con \mid cat) \times P(cat)}{P(con)} \qquad (1)$$

$$\begin{aligned} & P_{uni-gram}(con \mid cat) \\ & = P(word_1 \wedge \cdots \wedge word_j \mid cat) \\ & \approx \prod_{i=1}^{j} P(word_i \mid cat) \qquad (2) \\ & = \prod_{i=1}^{j} \frac{F(cat, word_i)}{\sum_{word' \in V} F(cat, word')} \end{aligned}$$

P(con|cat) could be estimated by Eq. (2) and Eq. (3) for a uni-gram model and a bi-gram model respectively.

$$\begin{aligned} & P_{bi-gram}(con \mid cat) \\ & = P(word_1 word_2 \wedge \cdots \wedge word_{j-1} word_j \mid cat) \\ & \approx \prod_{i=1}^{j-1} P(word_i word_{i+1} \mid cat) \qquad (3) \\ & = \prod_{i=1}^{j-1} \frac{F(cat, word_i word_{i+1})}{\sum_{word' word'' \in V'} F(cat, word' word'')} \end{aligned}$$

Here, the symbol *F* stands for *frequency*. For example, *F(cat,wordiwordi+1)* indicates the total number of citation contexts that has been assigned with a particular category, and contains the bi-gram $word_i word_{i+1}$ as well.

The symbol *V* and *V'* indicate the set of all the single words appearing in the training dataset and the set of all bi-grams respectively. The bi-gram model is an extension of the uni-gram, where a context is considered as the aggregation of all the consecutive two-word pairs.

The calculation process is simple. The system assigns one of the nine citation-reason categories to each input citation context based on the computation of each *P(cat|con)* and a comparison among them. More specifically, the category holding the highest *P(con|cat)×P(cat)* is assigned to the input citation context.

On the other hand, as all the probability values used in Eq. (2) or Eq. (3) have been calculated in advance, the final determination takes up very little time. In other words, there is no time-lag problem for our machine-learning based approach.

4.3 Evaluation Experiments

We conducted several experiments to evaluate the effectiveness of our machine-learning based proposal on citation-reason prediction. Here, in order to examine the utility of the preceding and succeeding sentence for machine learning, we carry out the experiments with two kinds of contexts: the citing sentence itself, and the whole context including the citing, preceding, and succeeding sentences. We randomly divide our data into two groups: a training data set of 800 records and a test data set of 100 records. Table 3 and Table 4 show the results from our experiments.

Table 3: Results of machine-learning based citation-reason prediction.

Language Model + Context	Precision
unigram + *citing sentence*	17%
unigram + whole context	26%
bigram + *citing sentence*	66%
bigram + whole context	71%

Table 3 shows the superiority of the bi-gram language model over the uni-gram model. This seems to be able to prove one intuition that the same collocations tend to appear in the context of citations with the same reason in Japanese scientific papers. On the other hand, using the whole context leads to a more accurate model than employing citing sentences only. In some situations, it is likely that we get even better results if we expand the contextual scope, for example, to more than one preceding or succeeding sentence, or even the whole paragraph. At the same time however, noise contained in the expanded context might produce harmful effects. For example, sometimes when multiple citation instances appear close to each other, their contexts will overlap with each other if we expand the contextual scope too widely.

Table 4: Experimental results for each citation-reason category.

	Weak	Coco	CocoGM	CocoRo	PBas	PUse	PModi	PMot	Neut
Precision	97%	33%	56%	50%	62%	78%	47%	45%	90%

Table 4 contains the experimental results using the bi-gram model and the whole context for each citation-reason category. *Weak* and *Neut* work very well, which conforms with the intuition. On the other hand, *PModi* and *PMot* seem to be unreliable. This is reasonable too. *PUse* and *PModi* are highly similar and *PModi* usually needs more contextual information to be distinguished from *PUse*.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Identifying *PMot* is even harder than *PModi* as demonstrating the validity of a citing paper through a cited paper sometimes seems more like a neutral description about the cited paper. The worst performance is observed in connection with the classifier for *Coco*. This might have come from the low amount of training data for *Coco* compared with other citation-reason categories, and indicates the necessity of increasing the amount of training data, especially for the categories with smaller datasets.

Table 5: Experimental results concerning training data scale and prediction accuracy.

Scale of traning	100	200	400	600	800
Precision	39%	43%	53%	65%	71%

In an experiment concerning the relationship between the scale of training data and the prediction accuracy, we used the same test data set of 100 records as above. We repeat the machine learning process five times with 100, 200, 400, 600, and 800 records as training data. Precisions are shown in Table 5. The figures here reveal the fact that a larger training dataset tends to enhance the performance of the machine-learning based classifier. Given this perspective, assigning citation-reasons to the remaining un-annotated records and further enlarging the database seem to be two significant future tasks.

5 CITATION-REASON VISUALIZATION

Using the paper database and the citation-reasons assigned to each citation instance, we have built a practical system attempting to visualize citation-reasons to help users find relevant papers that they might be specially interested in. In this section, we first introduce the system briefly, and then describe the evaluation process to examine the effectiveness of our system.

5.1 System Description

Our system is mainly composed of three functions: Basic Paper Search, Citation-reason Visualization, and Paper Information Display. Basic paper search is the first step in paper acquisition. In this module, the system helps users find a particular paper of interest.

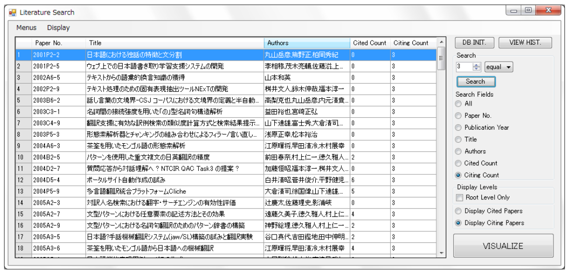

Figure 1: Basic Paper Search.

Figure 1 is a screen shot of the initial system interface. Users search the database through the grid view for a paper of interest as their starting point. During this process, users may search a paper by its title, author names, publication year, and the total number of its references or the total count of it as a cited paper. Not only full-text search but also partial-match search is accepted here. Also, a combination of multiple search functions is allowed to bring users refined search results. For example, you can try to find a paper from ANLPJ with the word *language* in its title and more than 5 reference papers in its reference list, and that has been cited 3 times since its publication year, say, 2010. When a user finally locates a paper, she may choose to open another form to view the visualized citation-reason information starting from the selected paper above.

Figure 2 shows such a form generated during the citation-reason visualizing process. Here, a hierarchical graph has been generated automatically from the root (i.e., the starting paper), spreading to the papers cited by it in the second layer, and other papers cited by the second-layer papers or deeper ones likewise recursively. Another kind of graph could be generated in a similar way for a starting paper by locating papers citing it recursively.

The numbered boxes in Figure 2 are actually document-like icons standing for papers. A paper being cited by an upper-layer paper is located in a lower position in Figure 2. The numbers are assigned in turns to each paper automatically by the system, and don't have any special meaning.

For example, Paper 16 in Figure 2 is being cited by Paper 12, and citing Paper 17, Paper 23, Paper 24, and Paper 25 at the same time. The system just visualizes the analytical results of citation-reason that have been obtained in advance using the machine-learning based approach described in Section 4.2.

Citation-reasons are represented by straight lines with different colors for different categories. Users can also choose to highlight only one or several categories by checking the radio button in front of each corresponding citation-reason.

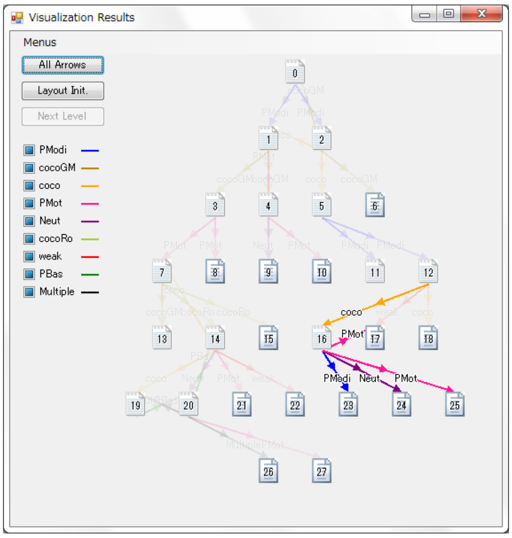

Figure 2: Visualized *citation-reason* information for the starting paper.

Other available functions include moving nodes to wherever you want for clear vision, showing the multi-layer graph layer by layer, displaying specific paper information whenever you put the mouse pointer on top of a paper node, and etc. All the functions implemented are developed to help users grasp the citation-relations between papers faster and more accurately.

Finally, when you double click any paper node in Figure 2, the third function, Paper information display, will create a new window, showing the specific information on the selected paper.

5.2 Effectiveness Evaluation

We conducted a set of experiments to examine the effectiveness of our system in helping users with their paper acquisition. Fourteen students that are not involved in this study have cooperated as the examinees. Three kinds of experiments, as shown below, were carried out with the same examinees.

- Experiment 1:not permitted to use the system
- Experiment 2: permitted to use the system except the citation-reason part
- Experiment 3: permitted to use the entire system including the citation-reason visualization function

Experiment 2 uses a simplified version of our system leaving over the straight line standing for the reference relation while removing the specific citation-reasons that were originally on top of them.

We first select one starting paper and the five most relevant papers for it from a randomly generated 50-paper subset using the database. Then in each experiment the examinees are told to select five most significantly related papers from the 50-candidate collection within 15 minutes. There is no major difference between each candidate collection, and every examinee uses a computer with exactly the same specifications. Our aim is to analyze the difference in working time and the acquisition correctness.

Table 6: Evaluation results.

	Experiemnt1	Experiment2	Experiment3
Average correct-number	0.57	1.57	2.57
Average working-time	12m35s	9m7s	9m1s

Table 6 shows the evaluation results. We can see from the table that using a supporting system improves the accuracy of obtaining relevant papers. The citation-reason based method even enhances this tendency with an average correct number of 2.57, which means that most of the examinees have on average correctly selected more than half of the correct answers with the help of citation-reason visualization. Similarly, the tendency shown by average working-time is also as expected. Examinees in the first experiment have to rely on their own efforts to obtain the relevant papers, which is the most time-consuming case in the three experiments. Besides, a subsequent questionnaire shows that *PBas*, *PUse*, and *cocoGM* are the most contributive citation-reasons during the process of trial and error in paper acquisition for Experiment 3.

There are still a few issues on the experimental method. First, the time limit is set to be 15 minutes, which might be insufficient for the examinees to carry out a good job. Another concern is about the system usage instruction. In order to reduce each examinee's burden, only 10 minutes were provided to examinees for understanding how to use the system in Experiment 2 and Experiment 3. An unstructured interview after the experiments even shows that a couple of examinees had not correctly understood the meaning of some of the nine citation-reasons. These issues are important and need to be considered in the future.

6 CONCLUSIONS

Paper acquisition is an important step in scientific research. Content-similarity-based methods and citation-reference based methods have been proposed in previous studies to search relevant papers for a given starting scientific paper. However, none of these could effectively help users find papers, for example, employing the same theoretical model or using a different experimental data set.

In this paper, we cast a spotlight on the citation-reason analysis which has been conducted previously for other purposes, and propose a method for classifying and organizing papers based on citation-reasons. Specifically, we established a paper database from a real paper corpus, predicted the citation-reasons between papers by using a machine-learning based classifier trained on an extensive hand-annotated set of citations, and finally visualized the resulting information in a practical system to help users more efficiently find the most appropriate papers close to their personal demands. By using our system, we could expect more accurate searching results for more context-specific demands, such as the ones we have raised in Section 1, as long as we could follow the appropriate citation-reasons between papers.

To the best of our knowledge, this study is the first attempt to employ citation-reason analysis in paper acquisition. Also, compared with previous studies in citation-reason analysis, our approach defines different features for machine learning, uses a more flexible contextual scope, and a much bigger training data set. We have also established a larger database covering a longer time span and an open-access data-source compared to previous work, targeted at the practical use of our method in paper acquisition, rather than a sociological study.

Evaluation results using the practical system have shown the effectiveness of our approach in paper acquisition. However, we believe that improvements could be made with more powerful machine-learning approaches, such as Support Vector Machine or Conditional Random Fields. Also, more context-specific experiments should be conducted to show exactly how effective our idea is in helping focus the search for relevant papers.

REFERENCES

Baeza-Yates R., Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. Addison Wesley.

Dobashi, K., Yamauchi, H., Tachibana, R. 2003. Keyword Mining and Visualization from Text Corpus for Knowledge Chain Discovery Support. *Technical Report of IEICE, NLC2003-24*. pp.55-60. (in Japanese).

Kessler, M., 1963. Bibliographic Coupling between Scientific Papers. *Journal of the American Documentation*, Vol.14, No.1, pp.10-25.

Small, H., 1973. Co-citation in the Scientific Paper: A New Measure of the Relationship between Two Documents. *Journal of the American Society for Information Science*, Vol.24, No.4, pp.265-269.

Miyadera, Y., Taji, A., Oyobe, K., Yokoyama, S., Konya, H., Yaku, T. 2004. Algorithms for Visualization of Paper-Relations Based on the Graph Drawing Problem. *IEICE Transactions*. J87-D-I(3), pp.398-415. (in Japanese).

Yoon, S., Kim, S., Park. S. 2010. A Link-based Similarity Measure for Scientific Paper. In *Proceedings of WWW'2010*, pp.1213-1214.

Jeh, G., Widom. J. 2002. SimRank: A Measure of Structural-Context Similarity. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pp.538-543.

Nanba, H., Okumura, M. 1999. Towards Multi-paper Summarization Using Reference Information. *Journal of Natural Language Processing,* Vol.6, No.5, pp.43-62. (in Japanese).

Nanba, H., Kando, N., Okumura, M. 2001. Classification of Research Papers Using Citation Links and Citation Types, *IPSJ Journal* Vol.42, No.11, pp.2640-2649. (in Japanese).

Garfield, E. 1979. Citation Index: Its Theory and Application in Science. *Technology and Humanities*. New York, NY:J. Wiley.

Weinstock, M. 1971. Citation Indexs. *Encyclopedia of Ligrary and Information Science*, 5:16-40. New York, NY:Dekker.

Moravcsik M., Poovanalingan, M. 1975. Some Results on the Function and Quality of Citations. *Social Studies of Science*, 5:88-91.

Chubin, D., Moitra, S. 1975. Content Analysis of References: Adjunct or Alternative to Citation Counting? *Social Studies of Science*, 5(4):423-441.

Spiegel-Rosing, I. 1977. Science Studies: Bibliometric and Content Analysis. *Social Studies of Science*, 7:97-113.

Oppenheim, C., Susan, P. 1978. Highly Cited Old Papers and the Reasons Why They Continue to Be Cited. *Journal of the American Society for Information Science*, 29:226-230.

Garzone, M., Robert, F. 2000. Towards an Automated Citation Classifier. In *Proceedings of the 13th Biennial Conference of the CSCI/SCEIO (AI-2000)*, pp.337-346.

Pham, S., Hoffmann, A. 2003. A New Approach for Scientific Citation Classification Using Cue Phrases. In *Proceedings of the Australian Joint Conference in Artificial Intelligence*, Perth, Australia.

Radoulov. R., 2008. Exploring Automatic Citation Classification. *Master thesis in University of Waterloo*.

Teufel, S., Advaith, S., Dan, T. 2006. Automatic Classification of Citation Function. In *Proceedings of EMNLP-06*.

Teufel, S. 2010. *The Structure of Scientific Articles – Applications to Citation Indexing and Summarization*. CSLI Publications.