Classification of Research Papers Focusing on Elemental Technologies and Their Effects

Satoshi Fukuda, Hidetsugu Nanba, Toshiyuki Takezawa

Graduate School of Information Sciences,
Hiroshima City University 3-4-1 Ozuka-higashi, Asaminami-ku,
Hiroshima 731-3194, Japan
{fukuda, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

Akiko Aizawa

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyodaku,
Tokyo 101-8430 JAPAN
aizawa@nii.ac.jp

Abstract

We propose a method for the automatic classification of research papers in the CiNii article database in terms of the KAKEN classification index. This index was originally devised to classify reports for the KAKEN research fund in Japan. It is organized as a three-level hierarchy: Area, Discipline, and Research Field. Traditionally, research papers have been classified using machine-learning algorithms, using the content words in each research paper as features. In addition to these content words, we focus on elemental technologies and their effects, as discussed in each research paper. Examining the use of elemental technology terms used in each research paper and their effects is important for characterizing the research field to which a given research paper belongs. To investigate the effectiveness of our method, we conducted an experiment using KAKEN data. From the results, we obtained average recall scores of 0.6220, 0.7205, and 0.8530 for the Research Field, Discipline, and Area levels, respectively.

Keywords: text classification, information extraction, research paper

1. Introduction

The volume of scientific information has increased exponentially in recent times, because of the increase in the number of active researchers, which has made it difficult for researchers to find relevant research papers. For example, if a researcher were to retrieve papers using the query "mobile phone", many papers from a variety of research fields, such as *Database Engineering*, *Educational Technology*, and *Social Science*, would be shown in the retrieved results, even though the researcher would only need papers from one particular field. This motivated us to investigate the automatic classification of search results from academic repositories.

Several methods for text classification have been proposed. A typical approach is to extract content words from each text and then to use them as features for machine-learning algorithms. In addition to these content words, we focus on phrases that play particular semantic roles, the elemental (underlying) technologies used in each research paper, and their effects. Elemental technologies and their effects are considered useful for characterizing research fields. For example, "Support Vector Machines" (SVMs) and "Hidden Markov Models" (HMMs) are often used as elemental technologies in *Intelligent Information* fields, such as *Natural Language Processing*, *Speech Recognition*, or *Image Processing*, whereas these technologies are seldom used in *Humanities* or in *Agriculture*. Expressions of effects are also useful for the classification of research papers. For example, expressions such as "improvement of precision" are often used in evaluations in the *Intelligent Information* field, while "improvement of educational effects" and "improved motivation" are often used in the *Educational Technology* field. Therefore, we extract elemental technology terms and expressions of their effects from the research papers and use them as additional features for machine-learning-based text classification.

The remainder of this paper is organized as follows. Section 2 describes related work. Section 3 explains our method for the classification of research papers. Section 4 reports on the experiment, and discusses the results. We present some conclusions in Section 5.

2. Related Work

There has been much relevant research in the field of cross-genre information access. An example is the subtask of research paper classification at the Patent Mining Task conducted in NTCIR-7 (Nanba et al., 2008) and NTCIR-8 (Nanba et al., 2010) Workshops. In this subtask, research papers were classified using the International Patent Classification (IPC) system, which is a hierarchical patent-classification system used worldwide.

Most participant groups employed the k-Nearest Neighbor (k-NN) method, which is a comparatively easy solution when dealing with a large classification index, because the classification is based only on extracting similar examples, with no training process being required. Furthermore, the k-NN method is itself a ranking, which enables it to be applied directly to the IPC code ranking. In our system, the number of research fields is much smaller than for NTCIR-7. Therefore, we could use two methods for the basic framework; namely, the SVM and k-NN methods.

For the Patent Mining Task, Xiao et al. (2008) used the k-NN framework, in which various similarity-calculation and ranking methods were examined. In our work, we also examined several ranking methods. In our pilot study, we employed the Listweak method as our ranking method, which we will describe in detail in Section 3.2.

The participant groups for NTCIR-8's subtask were asked to extract expressions of elemental technologies and their effects from research papers and patents. In their work, Fukuda et al. (2012) proposed a system that applied a domain-adaptation method, using both research papers and patents, and confirmed its effectiveness. We utilize their system for extracting elemental technologies and their effects from research papers, and use for the classification of research papers, as described in detail in Section 3.1.

3. Automatic Classification of Research Papers Focusing on Elemental Technologies and Their Effects

In this section, we describe a method for the automatic classification of research papers in the CiNii article database in terms of the KAKEN classification index, focusing on elemental technologies and their effects. In Section 3.1, we explain how to extract and use elemental technologies and their effects. In Section 3.2, we present an overview of our system.

3.1. Automatic Creation of Lists of Elemental Technologies and Their Effects for each Field

3.1.1. Extraction of Elemental Technologies and Their Effects from Title and Abstract

Elemental technologies and their effects are considered useful for characterizing each research field, as discussed in Section 1. Therefore, we extract elemental technologies and their effects from research papers using the information extraction method of Fukuda et al. (2012), which is based on machine learning. They formulated information extraction as a sequence-labeling problem, after which they analyzed and solved it using an SVM. The tag set was defined as follows.

- **TECHNOLOGY** includes algorithms, materials, tools, and data used in each study or invention.
- **EFFECT** includes pairs of ATTRIBUTE and VALUE tags.
- **ATTRIBUTE** and **VALUE** include effects of a technology that can be expressed by a pair comprising an attribute and a value.

A tagged example is given in Fig. 1.

Through <TECHNOLOGY>closed-loop feedback control </TECHNOLOGY>, the system could <EFFECT><VALUE> minimize</VALUE> the <ATTRIBUTE>power loss </ATTRIBUTE></EFFECT>.

Fig. 1: A tagged example (translation from Japanese)

Fukuda et al. conducted an experiment using the dataset from the NTCIR-8 Patent Mining Task. From their experimental results, which included partial-match results, they obtained recall and precision scores of 0.2756 and 0.5393, respectively, for the analysis of research papers.

3.1.2. Extraction of Key Phrases and Creation of Their Lists

We now explain the procedure for extracting elemental technologies and their effects from research papers. First, we extract items in three categories (elemental technology, attribute, and value) from the <TITLE>, <ABSTRACT>, and <KEYWORDS> sections of 672,397 Japanese KAKEN data entries, using the technical trend analysis system created by Fukuda et al. that enables comprehensive collection. Using this information, we create an "Elemental Technology list", an "Attribute list", and a "Value list".

In addition to the above categories, we extract items in two categories (author and publication) from the <AUTHORS> and <PUBLICATIONS> sections of the 283,686 KAKEN data entries annotated with a research field code relevant to our experiments, using regular expressions to create an "Author list" and a "Publication list". Here, we consider that authors and academic conferences tend to specialize in particular research fields. Using as a key phrase the name of an author involved in various research fields or an academic conference involving researchers specializing in different disciplines may not lead to the correct assignment of the research field. We therefore examine the number of research fields associated with each author and academic conference in the 283,686 KAKEN data entries annotated with a research field code relevant to our experiments. The Japanese author names associated with fewer than three research fields are placed in "Author list 1", with the remaining Japanese author names being placed in "Author list 2". The names of academic conferences associated with fewer than 10 research fields are placed in "Publication list 1", with the remainder being placed in "Publication list 2". Examples of key phrases in the seven lists and their numbers are shown in Table 1. The weight given to each list was determined via a pilot study.

We also consider that if there is no entry for the <ABSTRACT> section of a research paper, it might not be possible to annotate a research field code because of a lack of information. We therefore use a co-authorship network (Backstrom et al., 2006; Zhang et al., 2008; Sendhilkumar et al., 2012). In this paper, we create Japanese co-authorships from the CiNii article database and KAKEN data, respectively. Currently, we use the co-authorships that appear at least five times in the 5,924,669 CiNii data entries and at least once in the 283,686 KAKEN data entries annotated with a research field code relevant to our experiments. This rule was determined via a pilot study. This enables us to obtain 3,268,625 co-authorship pairs from the CiNii article database and 1,094,510 pairs from the KAKEN data.

Key phrase list	Examples of key phrases	Number	Weight
Author list 1	Omatsu Machiko, Hirai Seiji	144,108	50
Author list 2	Sarai Akinori, Suzuki Satoru	15,567	1
Publication list 1	Japan Biogeography Society	88,598	40
Publication list 2	Information Processing Society of Japan	2,838	4
Elemental Technology list	PCR method, Monoclonal, Local government	424,482	14
Attribute list	Precision, Procedure, Precipitation	589,116	6
Value list	Efficiency, Decrease, Clear	68,520	3

Table 1. Examples of key phrases belonging to each list

3.2. System Configuration

The goal of our study is to classify research papers in the CiNii article database into the KAKEN classification index. Our system comprises two modules; namely, an Indexing Module and a Document Classification Module (see Fig. 2). We now describe both of these modules.

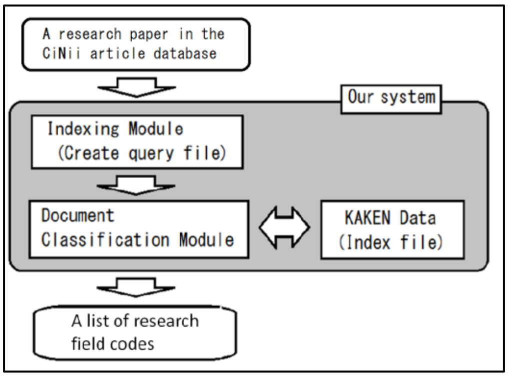
A research paper in the CiNii article database
Our system
Indexing Module (Create query file)
Document Classification Module
KAKEN Data (Index file)
A list of research field codes

Fig. 2: Schematic diagram of our system

3.2.1. Indexing Module

We create a query file for a research paper in the CiNii article database using the key phrase lists described in Section 3.1. Here, we introduce a method for setting the weights in each section. Changing the weight given to words appearing in the various sections of a document has been confirmed as effective by multiple experiments (Larkey, 1998; Fall et al., 2003).

First, we extract the noun words (plus any prefix) from the <ABSTRACT> section of the research paper entry. At this time, one-letter words and pure numerical words are deleted. Next, if the word is contained in the "Elemental Technology list", the "Attribute list", or the "Value list", the weight corresponding to that list is assigned. However, if the word is not contained in any of the three lists, a weight of 1 is assigned. We also extract noun words from the <TITLE> section. If the word is contained in the "Elemental Technology list", a weight of 17 is assigned. Otherwise, a weight of 1 is assigned.

Finally, we extract the Japanese author names from the <AUTHORS> section and the society names or the publication names from the <PUBLICATIONS> section. We then extract any author names that relate to the co-authorships in the KAKEN and CiNii article databases. If an extracted author name is contained in either "Author list 1" or "Author list 2", the weight corresponding to that list is assigned. If an extracted society (publication) name is contained in either "Publication list 1" or the "Publication list 2", the weight corresponding to that list is assigned. However, if an extracted author name or society (publication) name is not contained in any of the lists, we do not use it.

We create an index file from the KAKEN data entries using the above method. However, we extract the noun words from the <TITLE>, <ABSTRACT>, and <KEYWORDS> in the first step and do not use the co-authorship information.

3.2.2. Document Classification Module

For our basic framework, we use two classification methods, namely k-NN and SVM.

k-NN method

- **Similarity measure**

In the design of our k-NN classifier, we use the SMART measure (Salton, 1971) to calculate the similarity between the query file for research papers in the CiNii article database and the index file for the KAKEN data.

- **Ranking method**

We use the ranking method proposed by Xiao et al. (2008). First, our system extracts the top k documents $\{d_1, d_2, \dots, d_k\}$ with the highest similarities (k nearest neighbors) and calculates a score *Score(c)* for the research field of the extracted documents. Here, *Score(c)* can be regarded as a measure of the likelihood that the input document has label *c*. Next, these research fields are sorted in terms of scores. Finally, our system assigns the highest similarity of the research field to the input document and outputs it. In our system, the following ranking method (the Listweak method) is used, as chosen via a pilot study.

$$Score_{Listweak}(c) = \sum_{i=1}^{k} occur(c, d_i) Sim(q, d_i) r_1^i \qquad (1)$$

where r_1 is a parameter in the range (0, 1). The r_1^i term can be regarded as a penalty that punishes documents of lower rank. In our system, r_1 is set to 0.95 by default.

SVM method

We use the SVM method as another approach in the document classification module. We choose a linear classifier for the various kernel functions, because our method should achieve high speed automatic classification to be useful as a retrieval system. We now describe the method that annotates research fields for a research paper using SVM.

First, we create the classification categories for the research fields from the index file. Next, we apply each classifier to the query file. If a classifier outputs a positive value, the research field that it represents is assigned to the query file (research paper). Here, our task should be to annotate one research field for each given research paper. However, in the above method, if all classifiers output negative values for a query file, a candidate research field does not exist. A simple solution to this problem is to rank the classifier results. We use the distance from the hyperplane to rank the query files, and we assign the research field that represents the highest ranking.

4. Experiments

4.1. Experimental Methods

4.1.1. Datasets

KAKEN data

Research papers in the CiNii article database are not annotated with any research field codes. Therefore, we made use of research project data (hereinafter referred to as "KAKEN data") from the database of Grants-in-Aid for Scientific Research (KAKEN). In the KAKEN data, each research project is annotated with a research field code. Each project report contains a publication list, and some research papers in the list have been linked manually to the CiNii article database. We considered that such reports were also annotated with the research field codes for these papers, and we therefore used them as training and test data in our experiment.

The KAKEN classification index comprises three hierarchical levels; namely Area, Discipline, and Research Field, and each research project in the KAKEN data is annotated with a Research-Field-level code. Examples from the KAKEN classification index are shown in Table 2. The fields used have been modified over the course of several years. We used the research fields in Area, Discipline and Research Field that were in use in 2011, for which the Area (first level) contains 10 fields, the Discipline (second level) contains 69 fields, and the Research Field (third level) contains 297 fields.

Area (first level)	Discipline (second level)	Research Field (third level)
Interdisciplinary Fields	*Informatics*	*Intelligent Informatics, Software*
	Science Education and Educational Technology	*Science Education, Educational Technology*
Humanities	*Philosophy*	*Religious Studies, History of Thought*
	History	*Japanese History, Asian History*

Table 2. Examples from the KAKEN classification index

The KAKEN data contain 672,397 entries written in Japanese and published during the period 1965-2011. Each data item comprises several sections. We use 28,400 KAKEN data items for a training dataset that contains seven sections (<ID>, <TITLE>, <AUTHORS>, <ABSTRACT>, <KEYWORDS>, <PUBLICATIONS>, and <FIELD>). Currently, we have created 200 data entries per research field, thereby avoiding bias at the third level caused by differing amounts of data. We have 10 research fields at the first level, 44 research fields at the second level, and 142 research fields at the third level. Note that the number of research fields at the first and second levels has created bias.

CiNii article database

From the CiNii article database, we use 1,000 data items with an <ABSTRACT> section (the Abst dataset) and 1,000 data items without such a section (the Title dataset). These data are classified manually at the third level in the KAKEN classification index for use as test data. Currently, we have created 20 data items for each research field, giving a total of 100 research fields. There is therefore no bias at the third level. For the <FIELD> section, we use the research fields covered by the training data.

4.1.2. Evaluations

We used recall and Mean Reciprocal Rank (MRR) as evaluation measures. We covered the research field list for the top-3 output using our system. We also evaluated the performance for the research field that was annotated in the training and test data at the first, second, and third levels.

4.1.3. Comparison Methods

We conducted tests using our two methods and two baseline methods.

Our methods

- **k-NN**: Determined by the score calculated by summing the similarities, which penalizes documents with lower rank.
- **SVM**: Uses a method that annotates a query file with the research field code representing the highest classifier result.

Baseline methods

- **BASE_k-NN**: Does not use the elemental technology and effect features in k-NN.
- **BASE_SVM**: Does not use the elemental technology and effect features in SVM.

4.2. Experimental Results and Discussion

The experimental results are shown in Tables 3-5. For the k-NN method, we show the best performance using a threshold k (=1~50) for the Title dataset and the Abst dataset. In these tables, the k-NN method performed best for each hierarchical level of both datasets. Therefore, the k-NN method is more useful than the SVM method for our task. Moreover, the k-NN method obtained higher recall and MRR values than the baseline methods. This result indicates that including elemental technologies and their effects can be useful.

We investigated the overall effectiveness of extracting elemental technologies and their effects and using them as features for each research field, such as *Engineering* or *Social Science*. Here, we investigated the research fields at the first level, which deals with the most general classification of research fields. The top-1 recall scores for the k-NN and BASE_k-NN methods are shown in Table 6. This table also shows the number of correct answers and research papers. From Table 6, there were improvements not only for the recall scores in the *Engineering* and *Chemistry* fields but also for the *Social Science* and *Humanities* fields. These results show that elemental technologies and their effects are useful for a variety of fields. We also found that the recall score for *Interdisciplinary Fields*, which refers to complex research fields across two or more research areas with much interdisciplinary/cross-sectional research, was improved. *Interdisciplinary Fields* tends to focus on the areas of engineering and biological systems. For example, the *Information Science* and *Biomedical Engineering* fields belong to *Interdisciplinary Fields*. From Table 6, our method would appear to be effective for the *Engineering* field. As a result, we consider that our method will also be effective for *Interdisciplinary Fields* that deal with two or more research areas.

5. Conclusion

In this paper, we report on the construction of a method for the automatic annotation of research papers in the CiNii article database using KAKEN research field codes. We have focused on the terms used for elemental technologies and their effects in academic resources, which are used as keywords to improve the classification performance. To investigate the effectiveness of our method, we conducted an experiment using KAKEN data. From the results, we obtained average recall scores of 0.6220, 0.7205, and 0.8530 at the third, second, and first levels, respectively, when using the k-NN version of our method.

References

Backstrom, L., Huttenlocher, D., Kleinberg, J. and Lan, X. (2006). Group Formation in Large Social Networks: Membership, Growth, and Evolution. *Proceedings of the 12th ACM SIG KDD International Conference on Knowledge Discovery and Data Mining*, pp. 44-54.

		RECALL						MRR	
		@1		@2		@3			
		Title	Abst	Title	Abst	Title	Abst	Title	Abst
Our method	k-NN	**0.8270**	**0.8790**	**0.9260**	**0.9660**	0.9570	**0.9860**	**0.8858**	**0.9262**
	SVM	0.7390	0.8240	0.8200	0.9070	0.8660	0.9340	0.7948	0.8745
Baseline	BASE_k-NN	0.8160	0.8510	0.9250	0.9450	**0.9590**	0.9770	0.8810	0.9072
	BASE_SVM	0.7550	0.7840	0.8630	0.8730	0.8890	0.9170	0.8177	0.8432

Table 3. Recall and MRR values at the first level

		RECALL						MRR	
		@1		@2		@3			
		Title	Abst	Title	Abst	Title	Abst	Title	Abst
Our method	k-NN	**0.6830**	**0.7580**	**0.8240**	**0.8870**	**0.8700**	**0.9300**	**0.7662**	**0.8285**
	SVM	0.5650	0.6830	0.6740	0.7940	0.7040	0.8370	0.6295	0.7528
Baseline	BASE_k-NN	0.6790	0.7130	0.8190	0.8460	0.8680	0.9000	0.7630	0.7965
	BASE_SVM	0.5920	0.6410	0.7050	0.7550	0.7570	0.8030	0.6658	0.7140

Table 4. Recall and MRR values at the second level

		RECALL						MRR	
		@1		@2		@3			
		Title	Abst	Title	Abst	Title	Abst	Title	Abst
Our method	k-NN	**0.5970**	**0.6470**	**0.7440**	**0.8060**	**0.7920**	**0.8580**	**0.6823**	**0.7428**
	SVM	0.5160	0.6040	0.6360	0.7090	0.6640	0.7550	0.5853	0.6718
Baseline	BASE_k-NN	0.5810	0.6090	0.7350	0.7370	0.7810	0.8020	0.6725	0.6890
	BASE_SVM	0.5100	0.5570	0.6440	0.6520	0.6860	0.7140	0.5910	0.6251

Table 5. Recall and MRR values at the third level

	Engineering		*Social Science*		*Interdisciplinary Fields*		*Humanities*		*Agriculture*	
	Title	Abst	Title	Abst	Title	Abst	Title	Abst	Title	Abst
k-NN	**0.8346** (217/260)	**0.9500** (247/260)	**0.5667** (34/60)	**0.8833** (53/60)	**0.5714** (40/70)	**0.6000** (42/70)	**0.8000** (16/20)	**0.8500** (17/20)	0.8375 (67/80)	**0.9500** (76/80)
BASE_k-NN	0.8192 (213/260)	0.8962 (233/260)	0.5500 (33/60)	0.8167 (49/60)	0.5571 (39/70)	0.5429 (38/70)	0.7000 (14/20)	0.7500 (15/20)	0.8625 (69/80)	0.8750 (70/80)
	Medicine, Dentistry, and Pharmacy		*Chemistry*		*New Multidisciplinary Fields*		*Mathematical and Physical Sciences*		*Biology*	
	Title	Abst	Title	Abst	Title	Abst	Title	Abst	Title	Abst
k-NN	**0.9556** (344/360)	0.9333 (336/360)	**0.8667** (26/30)	**0.6333** (19/30)	**0.4000** (8/20)	0.5000 (10/20)	0.8250 (66/80)	0.8750 (70/80)	0.4500 (9/20)	0.4500 (9/20)
BASE_k-NN	0.9444 (340/360)	0.9333 (336/360)	0.8000 (24/30)	0.6000 (18/30)	0.3500 (7/20)	0.5500 (11/20)	0.8375 (67/80)	0.8750 (70/80)	0.5000 (10/20)	0.5500 (11/20)

Table 6. Top-1 recall scores for each research field at the first level

Fall, C.J., Torcsvari, A., Benzineb, K. and Karetka, G. (2003). Automated Categorization in the International Patent Classification. *Proceedings of the ACM SIGIR Forum*, pp. 10-25.

Fukuda, S., Nanba, H. and Takezawa, T. (2012). Extraction and Visualization of Technical Trend Information from Research Papers and Patents. *Proceedings of the 1st International Workshop on Mining Scientific Publications, collocated with JCDL 2012*.

Larkey, L.S. (1998). Some Issues in the Automatic Classification of U.S. Patents. *Working Notes for the AAAI-98 Workshop on Learning for Text Categorization*, pp. 87-90.

Nanba, H., Fujii, A., Iwayama, M. and Hashimoto, T. (2008). Overview of the Patent Mining Task at the NTCIR-7 Workshop. *Proceedings of the 7th NTCIR Workshop Meeting*, pp. 325-332.

Nanba, H., Fujii, A., Iwayama, M. and Hashimoto, T. (2010). Overview of the Patent Mining Task at the NTCIR-8 Workshop. *Proceedings of the 8th NTCIR Workshop Meeting*, pp. 293-302.

Salton, G. (1971). The SMART Retrieval System - Experiments in Automatic Document Processing. *Prentice-Hall, Inc., Upper Saddle River, NJ*.

Sendhilkumar, S., Mahalakshmi, G.S. and Dilip, S.S. (2012). Enhancement of Co-authorship Networks with Content-Similarity Information. *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, pp. 1225-1228.

Xiao, T., Cao, F., Li, T., Song, G., Zhou K., Zhu, J. and Wang, H. (2008). KNN and Re-ranking Models for English Patent Mining at NTCIR-7. *Proceedings of the 7th NTCIR Workshop Meeting*, pp. 333-340.

Zhang, X., Hu, X. and Zhou, X. (2008). A Comparative Evaluation of Different Link Types on Enhancing Document Clustering. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 555-562.