MULTI-LABELED PATENT DOCUMENT CLASSIFICATION USING TECHNICAL TERM THESAURUS

Yoshimi Suzuki and Fumiyo Fukumoto

Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, 4-3-11 Takeda, Kofu, Japan

Keywords: Thesaurus, Patent, Document classification.

Abstract: This paper presents a method for patent document classification by using an expanded technical term thesaurus. For classifying structural documents such as patent documents, structural information is very useful. However, if we use documents divided into several applicant tags, the number of words are limited. For example, 'Title of invention' tag is very important for patent document classification. However, the number of words in the tag is very few. Therefore, in order to deal with this problem, we employ two methods. One is to classify applicant tags into semantic tags, the other is word expansion using an expanded technical term thesaurus. For thesaurus expansion, our system integrates technical terms into a thesaurus using patent documents. The classification results showed the method using the expanded thesaurus was better than that without thesaurus. Although our method is very simple, it is comparable to other methods. These results suggest that thesaurus and our method to expand thesaurus can be useful for patent document classification.

1 INTRODUCTION

Patent document classification is an important issue of NLP. Some workshops for patent documents classification have been held, and many researchers proposed various methods. However, there are few methods using thesaurus. Because technical term thesaurus are required for patent document classification.

Currently there are a lot of machine readable thesauri, e.g. WordNet(Fellbaum, 1998), BunruiGoi-Hyo (National Language Research Institute, 1964) and EDR concept dictionary. However, they are thesauri of common words and they have few technical terms or their hierarchical semantic features are not for technical terms. JST Thesaurus (Japan Science and Technology Agency, 1999) is a technical term thesaurus. It consists of 43,314 index words, while many technical terms are not listed in it. Therefore, it is necessary to construct thesaurus of technical terms.

There are a lot of studies for thesaurus construction and thesaurus expansion. Tokunaga, et al. (Tokunaga, 1997) and Uramoto (Uramoto, 1996) proposed methods for extending an existing thesaurus by classifying new words in terms of that thesaurus. However, their studies are for words commonly used and not for technical terms.

For thesaurus construction or thesaurus expansion, we have to extract similar word pairs. In order to extract similar word pairs, some methods ((Hindle, 1990), (Lin, 1998) and (Hagiwara et al., 2006)) based on dependency relationships are proposed. However, their methods are for commonly-used words and they did not mentioned whether their methods were effective for technical terms.

For extracting similar words of a technical term, we have to deal with the following two difficulties.

- Some technical terms do not appear frequently.
- Some technical terms are used in the same contexts.

Therefore, it is difficult to extract various dependency relationships of technical terms.

In this paper, we use an expanded thesaurus for text categorization. We propose a method to integrate new technical terms into a core thesaurus. We also propose a method to use the thesaurus for patent document classification. We perform some experiments using the thesaurus in order to confirm the expanded thesaurus is effective for multi-labeled patent document classification. We compare the results of our system with the results using other methods. Our method is very simple, while our system is competitive to other systems.

Suzuki Y. and Fukumoto F..
MULTI-LABELED PATENT DOCUMENT CLASSIFICATION USING TECHNICAL TERM THESAURUS.
DOI: 10.5220/0003658504250428
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2011), pages 425-428
ISBN: 978-989-8425-80-5

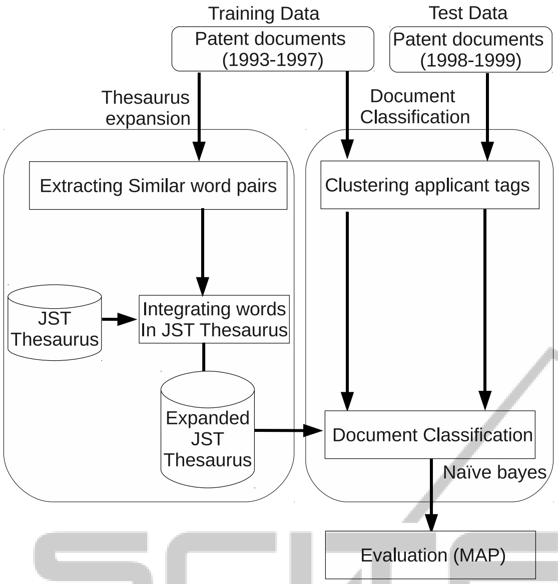

Figure 1: System overview.

2 SYSTEM DESIGN

Our system consists of two phases: "Thesaurus expansion" and "Document classification". Figure 1 shows our system overview.

2.1 Thesaurus Expansion

Some technical terms do not frequently appear in even if large corpora, then we have to use hierarchical semantic features for smoothing technique. Firstly, we extract similar word pairs using dependency relationships. Dependency relationship between two words is used for extracting semantic similar word pairs. For example, Lin proposed "dependency triple" (Lin, 1998). A dependency triple consists of two words: w, w' and the grammatical relationship between them:r in the input sentence. $||w,r,w'||$ denotes the frequency count of the dependency triple (w,r,w'). $||w,r,*||$ denotes the total occurrences of (w,r) relationships in the corpus, where "$*$" indicates wild card.

We used 10 kinds of Japanese case particles as r. Table 1 illustrates the case particles which we used as r.

In order to extract the corresponding semantic feature of the new word, we extract dependency triples of the new word and the extracted words. Using some extracted words, many types of dependency triples are extracted. For extracting the similar words from the core thesaurus, first, $I(w,r,w')$ is calculated using Formula (1).

Table 1: 10 case particles.

case particle (r)	description
ga	nominative case
no	genitive case
wo	accusative case
ni	dative case
he	goal
to	comitative case
kara	elative case
yori	from, at
de	inessive case
ya	coordination

$$\begin{aligned} & I(w,r,w') \\ &= -\log(P_{MLE}(B)P_{MLE}(A|B)P_{MLE}(C|B)) \\ &\quad -(-\log P_{MLE}(A,B,C)) \\ &= \log \frac{||w,r,w'|| \times ||*,r,*||}{||w,r,*|| \times ||*,r,w'||} \qquad (1) \end{aligned}$$

where P_{MLE} is the maximum likelihood estimation of a probability distribution.

Let $T(w)$ be the set of pairs (r,w') such that $\log \frac{||w,r,w'|| \times ||*,r,*||}{||w,r,*|| \times ||*,r,w'||}$ is positive. The similarity $Sim(w_1,w_2)$ between two words: w_1 and w_2 is defined by Formula (2).

$$\begin{aligned} & Sim(w_1,w_2) \\ &= \frac{\sum_{(r,w)\in T(w_1)\cap T(w_2)} (I(w_1,r,w)+I(w_2,r,w))}{\sum_{(r,w)\in T(w_1)} I(w_1,r,w) + \sum_{(r,w)\in T(w_2)} I(w_2,r,w)} \qquad (2) \end{aligned}$$

Finally, candidates of corresponding semantic features of the new word are detected using the hierarchical semantic features of the core thesaurus.

2.2 Document Classification

In document classification phase, we classify each document into some relevant themes.

In patent documents, there are many applicant tags: "Title of invention", "Abstract", "Purpose", "Claims", and so on. Each document has about 56 applicant tags. Most of applicant tags are used in most of documents. However, there are some notational variant in applicant tags. We classify these applicant tags into 6 semantic tags (Kim et al., 2005). Each label of semantic tag and classified applicant tags are shown in Table 2. In Table 2, "# of nouns" means average number of nouns in each documents.

Table 3 illustrates some examples of themes.

Many of themes correlate with "Purpose" of the semantic tags. Therefore, we decided "Purpose" is

Table 2: Examples of classified applicant tags into semantic tags.

Semantic tag	Examples of Applicant tag	# of nouns
Technological field	industrial application field	80.5
Purpose	title of the invention, purpose of the invention	134.1
Method	means of solving the problem	71.2
Claim	claim	151.2
Explanation	composition	166.4
Example	embodiment example	72.5

Table 3: Examples of themes.

Theme code	Description
2B011	Mushroom Cultivation
3C036	Drilling and boring
4D057	Centrifugal separators
5E022	Connector installation
5K068	Stereo-broadcasting methods

the most important semantic tags, and we used word expansion using expanded thesaurus for documents of "Purpose" tag, in document classification.

For document classification, we used "Bag Of Words" and distribution of words. Although we use many training data, many words in test data do not appear in training data. Therefore, we have to use word expansion using the expanded thesaurus. Although word expansion is useful, if we expand all words using thesaurus, the results must be worse by noise. Therefore, we use word expansion using the expanded thesaurus for documents of "Purpose" tag. We used Naive Bayes classifier for document classification. The themes $\hat{theme}$ which are selected as the relevant themes using the following equation.

$$\hat{theme} = \arg\max_{themes} P(theme) \prod_i P(w_i|theme) \qquad (3)$$

where w_i and w'_i mean $word_i$ and a related word of $word_i$. If w_i is a word in the expanded thesaurus, the words which are next neighbours in the thesaurus also used as w'_i.

3 EXPERIMENTS

We have an experiment to evaluate the effectiveness of the expanded thesaurus for document classification. We used the thesaurus for patent document classification. In the experiments, we decided corresponding themes for each patent documents out of about 2,900 themes.

Table 4: The number of related words.

related words	# of words
NT(narrower terms)	102,645
BT(brother terms)	122,606
RT(related terms)	26,958

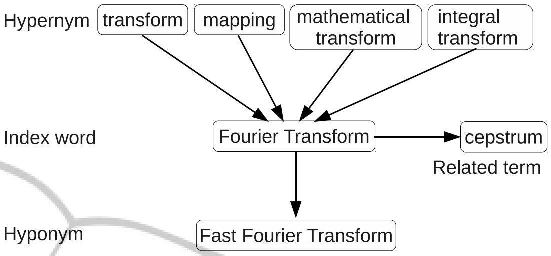

Figure 2: A part of JST Thesaurus (Fourier transform).

3.1 Experimental Setup

For the experiments we used Japanese patent documents and technical term thesaurus which was expanded by our method.

We used patent publication bulletins written in Japanese (1993-1999) which were provided by patent retrieval task of NTCIR Workshop 5 (Iwayama et al., 2005). The training data we used is the documents from 1993 to 1997. For test data we used the patent documents from 1998 to 1999. The number of training data was 1,707,194 documents. The number of test data was 2,008 documents. Each training data and test data had multiple themes. The number of themes was 2,903. Average number of themes of each document was about 2.26.

Firstly, we obtained similar word pairs and integrated them into a core thesaurus. We used JST Thesaurus (Japan Science and Technology Agency, 1999) which had 43,314 index words. Each index word had about 6 related words on average. The related words were classified into 3 categories: NT (narrower terms), BT (brother term) and RT (related terms). Table 4 shows the number of related words.

Figure 2 illustrates an index word (Fourier transform) and its related words of JST Thesaurus. The index words appeared about 2 million times in the 2,008 patent documents (1998-1999).

3.2 Document Classification Results

We retrieved relevant documents from Japanese patent documents using thesaurus information, We compared our results with the results in NTCIR5.

Table 5: Classification results (MAP).

Method	MAP
cosine	0.45
cosine + JST Thesaurus	0.46
cosine + expanded thesaurus	0.46
Naive Bayes	0.63
Naive Bayes + JST Thesaurus	0.64
Naive Bayes + expanded thesaurus	0.64
k-NN (BOLA1 (Kim et al., 2005))	0.69
Naive Bayes (JSPAT2)	0.66
k-NN (WGLAB9)	0.62
VSM (FXDM3)	0.49

We used Mean Average Precision (*MAP*) to compare these results. *MAP* is defined by the following equation

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}Precision(R_{jk}) \quad (4)$$

where Q is set of test documents, m_j is the number of relevant documents of $document_j$, and R_{jk} means kth ranked retrieval results of $document_j$.

Table 5 shows results of document classification. In Table 5, BOLA1, JSPAT2, WGLAB9 and FXDM3 are RunID of NTCIR5 Patent Retrieval Task. BOLA1 used k-NN and structure of patent documents. JSPAT2 used Naive Bayes. WGLAB9 used k-NN, where retrieval model is BM11 or the vector space model. FXDM3 used vector space model.

4 DISCUSSION

We expanded a technical term thesaurus using Japanese patent documents. To confirm our thesaurus is useful for text classification, we compared the results using our thesaurus with results without the thesaurus. As the results, we found that our thesaurus is effective for document classification. We also compared our method with the methods in NTCIR5 patent classification task. Although our method is very simple, we found our system is competitive to other systems. We classified 6 semantic tags in the experiments, and applied word expansion in "purpose".

Future work includes (i) applying the method to other data for quantitative evaluation, and (ii) comparing the method with other classification techniques to evaluate the effectiveness of the method.

ACKNOWLEDGEMENTS

The authors would like to thank the referees for their comments on the earlier version of this paper. This work was partially supported by The Telecommunications Advancement Fundation.

REFERENCES

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Hagiwara, M., Ogawa, Y., and Toyama, K. (2006). Selection of effective contextual information for automatic synonym acquisition. In *In Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 353–360.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of 28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275.

Iwayama, M., Fujii, A., and Kando, N. (2005). Overview of classification subtask at ntcir-5 patent retrieval task. In *Proceedings of NTCIR-5 Workshop Meeting*.

Japan Science and Technology Agency (1999). JST (JICST) Thesaurus 1999. http://jois.jst.go.jp/JOIS/html/thesaurus_index.htm.

Kim, J.-H., Huang, J.-X., Jung, H.-Y., and Choi, K.-S. (2005). Patent document retrieval and classification at kaist. In *Proceedings of NTCIR-5 Workshop Meeting*.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Proceedings of the Conference*, pages 768–774.

National Language Research Institute (1964). *Bunruigoihyo*. Shuei publisher (In Japanese).

Tokunaga, T. (1997). Extending a thesaurus by classifying words. In *In Proceedings of the ACL-EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources*, pages 16–21.

Uramoto, N. (1996). Positioning unknown words in a thesaurus by using information extracted from a corpus. In *In proceedings of COLING'96*, pages 956–961.