Integrating Query Translation and Text Classification in a Cross-Language Patent Access System

Guo-Wei Bian Shun-Yuan Teng
Department of Information Management
Huafan University, Taiwan, R.O.C.
gwbian@cc.hfu.edu.tw wells0609@gmail.com

Abstract

In this paper, a cross-language patent retrieval and classification system is presented to integrate the query translation using various free web translators on the internet and the document classification. The language-independent indexing method was used to process the multilingual patent documents, and the query translation method was used to translate the query from the source language to the target language. The mono-lingual and cross-lingual retrieved patent documents would be processed to classify the research papers (the queries) in terms of the International Patent Classification (IPC). The results indicate that the performance of the cross-lingual text classification reached almost the same level of the mono-lingual text classification.

Keywords: *Query Translation, Text Classification, Machine Translation, Cross-Language Information Retrieval, International Patent Classification (IPC).*

1 Introduction

Patent documents are the extremely important sources of information for technology. The researchers and developers have more attention to survey the related patents. The patent documents are the only way for opening the technology to protect the legal rights completely. Because the technical methods of patents quickly reflect the latest scientific and technological developments and research results, the quality and quantity of patents are the important indicators of a nation's innovation. Due to the increasing international competitions in the industry, the global business companies are aggressive to maintain the leading edge technologies and the market interests by the protections of patents.

Patent processing is extremely important in the fields such as business, industry, and law. Many applications are introduced to match the different requirements like Technology Survey, Invalidity Search, Patent Map, Patent Classification, Patent Mining, and Patent Translation [7, 11].

Different methods had been proposed for the patent classification. Support Vector Machine (SVM) [12], the K-nearest Neighbor (KNN) [8, 10], Hybrid Binary Classifier based on maximum entropy [4], and the Chi-square statistics method [6] are evaluated in NTCIR-6. The methods based on the maximum entropy, SVM, and KNN obtain higher classification accuracy, but the Chi-square method is significantly faster than the other methods [6].

To deal with the multiple language document collections, the word-based and bigram-based indexing methods had been introduced for multilingual information retrieval. These methods are evaluated for English and Japanese document retrieval [2, 13]. Query translation was usually employed to unify the language in queries and documents for Cross Language (Cross-Lingual) Information Retrieval (CLIR). Web translation system had been used in an online real-time multi-lingual and cross-language information access system for query translation and document translation [1].

We present a system to provide the functions of the monolingual and cross-language patent retrieval in English and Japanese. The users can input the queries or select the topic file, and use the different web translation systems to process the query translation. In additional, the translated query can be modified manually. The different fields of the query topics and the various patent document sets are selected to perform the cross-language patent retrieval from Japanese to English, and vice versa. For the requirement of online real-time patent access system, a statistic classification method is adopted in our system.

This paper is organized as follows. Section 2 describes the process of our system. Section 3 presents the experiments and the evaluation results. Finally, Section 4 concludes the remarks.

2 System Description

The architecture of the cross-language patent access system is illustrated in Figure 1. The multi-lingual patent document sets are processed and indexed. The system uses the word-based indexing for the English text collections and the bigram-based indexing for the Asian-language text collections. The users can input the query or select the topic from the topic file. The query translation module translates the query from the source language to the target language. In our system, the query is sent to the online web translator(s) which selected by the user, and the translated query terms are obtained after term extraction and term filtering. The user can review and modify the translated terms for patent retrieval. Finally, the retrieved patent documents are processed in the classification module to obtain the related IPC codes.

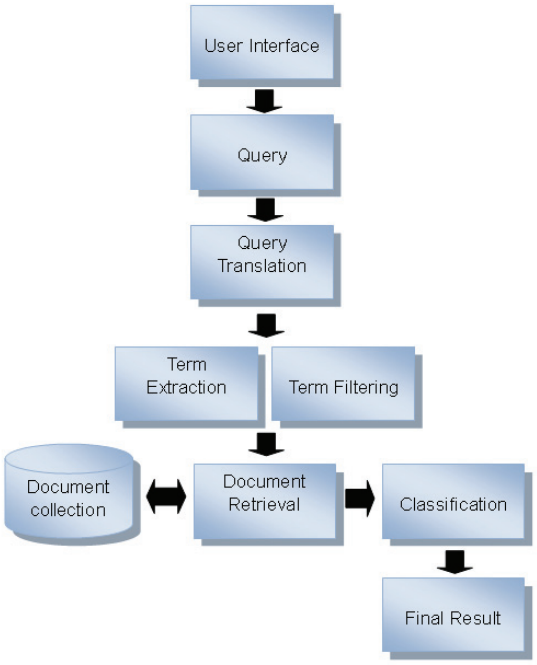


Figure 1. The architecture of the cross-language patent retrieval system

For example, Figure 2 shows the processing of Japanese-English cross-language patent retrieval. The Japanese topic is translated to the English topic. Several machine translation systems (Google [5], Excite [3], and Yahoo [15] translation systems) are used to translate the query from the source language to the target language. After the topic is translated, the tokenization processing is performed. The retrieval subsystem is used for scoring the retrieval documents to obtain the retrieval results.

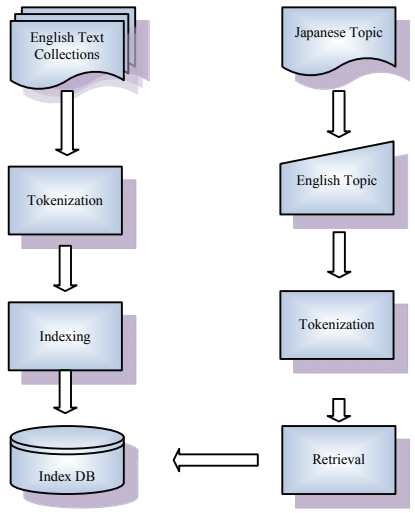


Figure 2. The processing of Japanese-English CLIR

As a newly established research group, we adapted one of the available open source information retrieval systems for our researches. Lemur [14] and Lucene [9] become the candidates for multi-lingual IR search engines. We used the Lucene toolkit developed by the Apache Software Foundation. The Apache Lucene project develops open-source search software in many programming languages, including Java, C#, C++, Delphi, Perl, Python, Ruby, and PHP. We adapted the C# version of the Lucene to build the cross-language patent retrieval subsystem. This subsystem introduced the Vector Space Model as a retrieval model and TF-IDF for term weighting.

2.1 Tokenization

The first task for Asian-language (Chinese, Japanese, and Korean) information retrieval is the text segmentation, since there are no word boundaries in Chinese, Japanese, and Korean texts. The bi-gram text segmentation and word segmentation have been widely used to parse the tokens and words of text collections.

Because the Asian languages have the different morpheme schemes, different word segmentation systems are needed for Chinese, Japanese, and Korean language processing. We adopt the language-independent technique of character bigram. The indexing unit is a pair of adjacent characters. For example, the string “航空交通管制” (Air traffic control) is indexed as the five tokens “航空”, “空交”, “交通”, “通管”, and “管制”.

In information retrieval, the punctuation marks and special characters are generally meaningless. Because Chinese, Japanese, and Korean used double-byte language coding, these symbols could be represented in ASCII or in different double-byte codes of these languages. Therefore, the system filters out these symbols and the stop words before indexing and retrieval tasks. After tokenization, the Lucene toolkit is used to index the patent collections.

2.2 Query Processing and Translation

In the monolingual English information retrieval, the query is generated from the selected field(s) of the original topic and then parsed as the stream of terms. In cross-lingual information retrieval, the query in source language is first translated to target language using different online web translation systems. The Google, Excite, and Yahoo! Babel Fish translation systems are used to translate the source languages to target languages. The user can select the translation methods (English-Japanese and Japanese-English translations) and the web translation systems to perform the query translation. To reduce the problems of translation for CLIR, the user can select one or more translation systems to translate the query and modify the translated results. Figure 3 illustrates the interface of our system and shows a example to retrieve the English patents using the Japanese topic.

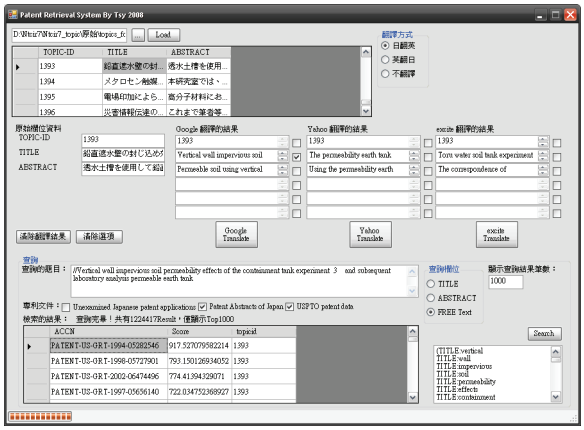


Figure 3. The interface of our cross-language information retrieval subsystem

For the translations of the proper nouns, the personal names, and the terminology terms, the Google translation generates the better results than the others. For example, the translated results of the personal name“ヘルベルト.フォン.カラヤン”(Herbert von Karajan) and the proper noun “航空管制” (Air traffic control) are listed in Table 1.

Table 1. The translation results of the web translation systems

	Personal Name	Proper Noun
Excite Translation	Herbert Von Karajan	Air-traffic control
Google Translation	Herbert Von Karajan	Air traffic control
Yahoo! Babel Fish	Hell belt. Phone. Karajan	Flight control

2.3 Text Classification

The mono-lingual and cross-lingual retrieved patent documents would be processed to classify the research papers (the queries) in terms of the International Patent Classification (IPC). In our system, the IPC codes of the query (the research paper) are obtained by the following processing:

(1) Perform the patent retrieval using the topic of the input patent (query)
(2) Retrieve the top-3000 patent documents and their IPC codes from patent data base
(3) Compute the scores of IPC Codes
Score(IPC Code) = Σ(the similarity between the query and the retrieved patent)
(4) Sorting the IPC Codes by their scores in step (3)

For example, the IPC code “A61B_5_02” are matched to the three patent documents (PATENT-US-GRT-2000-06152884, PATENT-US-GRT-1993-05181521, and PATENT-US-GRT-1998-05772600) in the retrieved results. And the similarity scores between the query and these patent documents are 21.264, 17.724, and 20.125. The score of IPC code “A61B_5_02” is summation (59.113). Figure 4 illustrates the interface of our cross-lingual text classification subsystem and shows the results (IPC codes) of classification for the Japanese topic. Here, the topic-id specifies the topic identification number, and IPC and Score specify the code and the score of IPC classification.

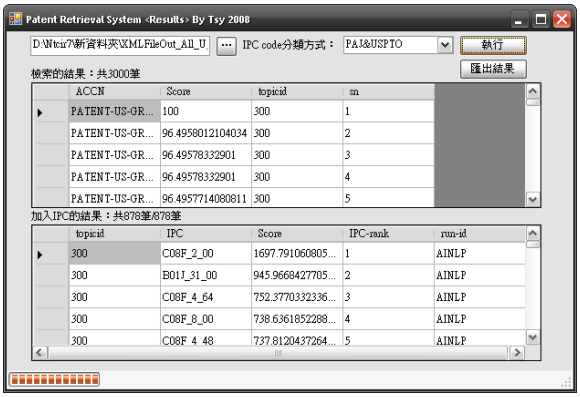


Figure 4. The interface of our cross-lingual text classification subsystem

3 Experiments

We participated in the NTCIR-7 Patent Mining task for retrieving and classifying research papers and patents. The purpose is the categorization of research papers (abstracts) written in Japanese or in English into the International Patent Classification (IPC). Our experiments consist of the English subtask and the Cross-lingual subtask (J2E).

3.1 Document Collection

The document sets for the NTCIR-7 Patent Mining task consisted of the research papers (NTCIR-1 and NTCIR-2 CLIR task Test Collection) and the patent documents from 1993 to 2002 in English and Japanese. The patent documents included the Unexamined Japanese patent applications, the USPTO patent data, and the Patent Abstracts of Japan. Table 2 shows the languages, the numbers of documents, and the storage sizes for the document collections.

Table 2. Document sets of the NTCIR-7 Patent Mining Task

Sources	Language	No. of Documents	Size (in MB)
NTCIR-1	Japanese	332,918	312
	English	187,080	218
NTCIR-2	Japanese	403,240	600
	English	134,978	200
Unexamined Japanese patent applications	Japanese	3,496,252	96,768
Patent Abstracts of Japan	English	2,543,488	4,102
USPTO patent data	English	1,315,470	53,351

Because the Japanese document collection is presented in the EUC coding, our system first transformed the documents from the EUC coding to UTF-8 coding. The Lucene supported the UTF-8 document format, which used for the multilingual text collections. In the English document collection, only the content of the <DOCNO>, <TITLE>, <ABST>, <SPEC>, and <CLAIM> fields could be used for the NTCIR-7 patent mining task.

Table 3 shows the sources, he number of bigram tokens, the size of the bigram index, and the indexing time for the Japanese document collection. The indexing of document collection is built on a personal computer (PC). The specification of the PC is: Intel 2.66 GHz Pentium-4 processor and 1 GB RAM; and the operating system is Microsoft Windows XP Professional. Table 4 shows the sources, the number of terms, the size of the index, and the indexing time for the English document set.

Table 3. The statistics of document collection in Japanese

Sources	No. of Bigram Tokens	Index Size	Indexing Time
NTCIR-1	1,596,747	312 MB	6.98 Hours
NTCIR-2	2,021,914	710 MB	9.56 Hours
Unexamined Japanese patent applications	7,596,840	46,445 MB	308.08 Hours

Table 4. The statistics of document collection in English

Sources	No. of Terms	Index Size	Indexing Time
NTCIR-1	1,033,575	211 MB	4.11 Hours
NTCIR-2	785,607	227 MB	2.78 Hours
Patent Abstracts of Japan	3,191,893	676 MB	22.07 Hours
USPTO patent data	2,653,186	1,942 MB	27.69 Hours

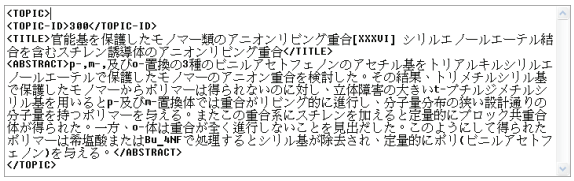
3.2 Queries

We participated in the English subtask and the Cross-lingual subtask (J2E) within the Patent Mining Task. The Japanese and English versions of the topics are used in our experiments. There are 879 topics in each language. Figure 5 lists the Japanese and English versions of the topic 300.

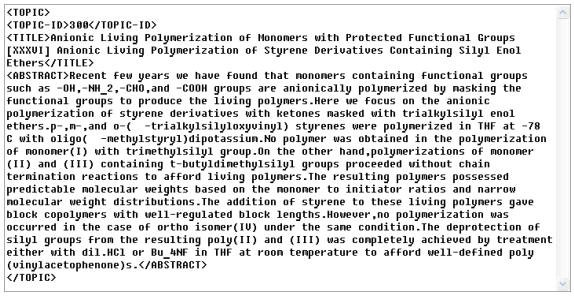
Three different queries are derived from the same topic to compare the classification performance. The three kinds of query are mentioned below.

(1) **T-run**: the short query from the title of the topic, i.e., the content of the <TITLE> field
(2) **A-run**: the long query from the topic's abstract, i.e., the content of the <ABSTRACT> field

(3) **TA-run**: the long query from the title and the abstract of the topic, i.e., the content of the <TITLE> field and <ABSTRACT> field



(a) Japanese version



(b) English version

Figure 5. Japanese and English versions of the topic 300.

3.3 Results and Discussion

In the English subtask, the queries are parsed to generate the query terms for retrieving the relevant documents. In the Cross-lingual subtask (J2E), three web translation systems can be selected by the user to perform the query translation. Table 5 shows the different translated results for the Japanese topic 309.

Table 5. The translated results of the Japanese topic 309

Topic	音声認識のための隠れマルコフ網の動的話者適応法
Google	Hidden Markov speech recognition system for dynamic speaker adaptation method
Yahoo	Dynamic speaker adaptation method of hiding Markov net for speech recognition
Excite	Dynamic speaker adjustment method of hiding Markoff net for voice recognition

Experimental results are retrieved using the Vector Space model with no relevance feedback. Because of the first participation and the integrating issues of the online web translation systems in our experiment for cross-lingual (Japanese-English and English-Japanese) information access, we spent lots of time to solve the problem of language coding and translate the queries for Cross-lingual subtasks (J2E and E2J).

For the formal run, 2 runs are submitted for English subtask, and 4 runs are submitted for the Cross-lingual subtask (J2E). After the formal run, more experiments are obtained to compare the performances using the different web translation systems for the cross-lingual text classification. The evaluation results are shown in Table 6 and Table 7. The symbol "*" in the Run field specifies the official run. Comparing the results of English subtask, the performance of the long query combined the title and abstract (TA-run) is better than the others (T-run and A-run). However, the differences of the performances of the short queries (T-run) and the long queries (A-run) are not significant.

Table 6. The evaluation results of English subtask (* specifies the official run)

Run	Retrieved	Relevant	Rel_Ret	MAP
ENG-T*	770280	2051	1380	0.0978
ENG-A	755944	2051	1409	0.0971
ENG-TA*	737595	2051	1455	0.1045

Table 7. The evaluation results of the Cross-lingual subtask (J2E) (* specifies the official run)

Web Translator(s)	Run	Retrieved	Relevant	Rel_ret	MAP
Excite (E)	T	789259	2051	1306	0.0882
	A	749716	2051	1369	0.0915
	TA	741590	2051	1405	0.0940
Google (G)	T*	776978	2051	1335	0.0941
	A	323007	2051	1221	0.1094
	TA*	727168	2051	1463	0.1070
Yahoo (Y)	T	797944	2051	1264	0.0850
	A	767818	2051	1352	0.0902
	TA	759049	2051	1384	0.0963
All (E+G+Y)	T*	755792	2051	1363	0.0934
	A	727824	2051	1433	0.0982
	TA*	724348	2051	1460	0.1041

Comparing the results of cross-lingual J2E subtask, Google translation performs better Japanese-English translation than Excite and Yahoo translations. Especially, the performance of the A-run using Google translation obtained the best score. In most cases, the performance of the long query combined the title and abstract (TA-run) is better than the others (T-run and A-run). The performance of the query translated using the three translation systems is better than the performance of the query translation using Excite translation (or Yahoo translation). Comparing the results of the English subtask with the Cross-lingual subtask (J2E), the performance of the cross-lingual text classification reached almost the same level of the mono-lingual text classification.

4 Conclusion

In this paper, we discuss the effectiveness of query translations with different machine translation systems for cross-language information retrieval and text classification. The language-independent indexing technology is used to process the text collections in various Asian languages. In the experimental results, we can find that the performance of the long query combined the title and abstract is better than the query using the title or the abstract of the research paper. For cross-lingual Japanese-English text classification, Google translation system performs better Japanese-English translation than Excite and Yahoo did. And the performance of the cross-lingual text classification reached almost the same level of the mono-lingual text classification. In the future, we will involve using the other retrieval models and combining the relevance feedback to improve the retrieval performance. For cross-language information retrieval, the dictionary-based query translation using the Wikipedia and the translation disambiguation using co-occurrence relationships will be used. And the other text classification technology like K-NN method will be adopted to improve the performance of patent classification.

Reference

[1] Bian, G.W. and Chen H.H. “Cross Language Information Access to Multilingual Collections on the Internet”, *Journal of American Society for Information Science & Technology (JASIST), Special Issue on Digital Libraries*, 51(3), pp.281-296, 2000.

[2] Cheng, C.C.; Shue, R.J.; Lee, H.L.; Hsieh, S.Y.; Yeh, G.C. and Bian, G.W. “AINLP at NTCIR-6: Evaluations for Multilingual and Cross-Lingual Information Retrieval”, Proceedings of NTCIR-6 Workshop, Japan, 2007.

[3] Excite, http://www.excite.co.jp/world/english/

[4] Fujino, A. and Isozaki, H. "Multi-label Patent Classification at NTT Communication Science Laboratories", Proceedings of NTCIR-6 Workshop, Japan, 2007.

[5] Google Translation, http://www.google.com.tw/translate_t

[6] Hashimoto, K and Yukawa, T. "Term Weighting Classification System Using the Chi-square Statistic for the Classification Subtask at NTCIR-6 Patent Retrieval Task", Proceedings of NTCIR-6 Workshop, Japan, 2007.

[7] Iwayama, M.; Fujii, A.; and Kando, N. "Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task", Proceedings of NTCIR-6 Workshop, Japan, 2007.

[8] Konishi, K. and Takaki, T. "F-term Classification System Using K-Nearest Neighbor Method", Proceedings of NTCIR-6 Workshop, Japan, 2007.

[9] Lucene, http://lucene.apache.org/java/docs/index.html

[10] Murata, M.; Kanamaru, T.; Shirado, T.; and Isahara, H. "Using the K-Nearest Neighbor Method and SMART Weighting in the Patent Document Categorization Subtask at NTCIR-6", Proceedings of NTCIR-6 Workshop, Japan, 2007.

[11] Nanba, H.; Fujii, A.; Iwayama, M.; and Hashimoto, T. “Overview of the Patent Mining Task at the NTCIR-7 Workshop”, Proceedings of NTCIR-7 Workshop, Japan, 2008.

[12] Rikitoku, M. "F-term classification Experiments at NTCIR-6 for Justsytems", Proceedings of NTCIR-6 Workshop, Japan, 2007.

[13] Shi, L. and Nie, J.Y. “Using Unigram and Bigram Language Models for Monolingual and Cross-Language IR", Proceedings of NTCIR-6 Workshop, 2007.

[14] The Lemur Toolkit, http://www.lemurproject.org

[15] Yahoo Babel Fish, http://babelfish.yahoo.com/