NTCIR-7 Patent Mining Experiments at Hitachi

Hisao Mase, Makoto Iwayama
Hitachi, Ltd.
292 Yoshida-cho, Totsuka-ku, Yokohama, Kanagawa, 244-0817, Japan
hisao.mase.qw@hitachi.com
makoto.iwayama.nw@hitachi.com

Abstract

This paper reports results of our experiments on the automatic assignment of patent classification to research paper abstracts. We applied K-Nearest Neighbors Methods and three kinds of query term expansion methods using a research paper abstract dataset and a patent document dataset to improve the classification accuracy. The results show that these query expansion methods slightly improve classification accuracy when the parameter is tuned appropriately. We also compared the classification accuracy when research paper abstracts are used as input with that when abstracts or full texts of patent documents are used as input.

Keywords: *Classification of research papers, Patent Retrieval, Query Expansion, Parameter Tuning.*

1 Hitachi's Approach in NTCIR-7

The NTCIR-7 Patent Mining Task is to fully automatically assign appropriate International Patent Classifications (IPCs) to research paper abstracts. We used the K-nearest neighbors (KNN) method as a basis of automatic classification. Our system identifies IPCs through the following steps:

(1) Terms (nouns, verbs and adjectives) are extracted from an input abstract text using the Chasen[1], morphological analysis tool.
(2) A weight for each term is calculated using a general term frequency - inverted document frequency (TF-IDF) method.
(3) The top K similar patent documents are retrieved from a patent document database by a similar document retrieval engine, GETA[1][2].
(4) The IPCs assigned to each of the K patent documents are identified.
(5) For each of the identified IPCs, the retrieval scores of the patent documents with the IPC are summed up.
(6) The IPC scores are sorted in descending order. The top X IPCs are assigned to the input abstract text.

In this task, we thought that the barrier to improving the classification accuracy would be that research paper abstracts are too short to allow extraction of enough terms to identify IPCs. Thus, we used query expansion methods to add related terms to the input abstract. In these methods, documents similar to an input paper abstract are retrieved from a document database. The terms are then extracted from the top M retrieved similar documents and a weight for each term is calculated. The weighted terms are added to the term set of the input abstract, and the similar patents are retrieved using the expanded term set.

2 Query Expansion Methods

We used three kinds of query expansion method: two types that use paper abstracts similar to an input abstract and one that uses patent documents similar to an input abstract.

2.1 Query Expansion Method (QEM1)

In this method, related terms are extracted from paper abstracts similar to an input abstract and then weighted. The frequency of each term is calculated by adding the frequency in each of the similar abstracts. This method is a well-known general query expansion method.

We used the following algorithm:

(1) Terms are extracted from an input abstract text using Chasen.
(2) The weight of each term is calculated using a TF-IDF method (IDF is from paper abstracts).
(3) The top M similar paper abstracts are retrieved from a paper abstract database using GETA.
(4) The total frequency of each term in these abstracts is calculated.
(5) The top N terms with higher term frequency (not term weight) which do not appear in the input abstract are identified and added to the term set

[1] GETA is a research effort in the "Innovative Information Technology Incubation Project" promoted by the Information-technology Promotion Agency, Japan (IPA).

extracted in Step (1). The term frequency of all added terms is set to 1.
(6) With the expanded term set from Step (5), similar patent documents are retrieved from a patent document database using GETA (IDF is from patent documents).
(7) The IPCs assigned to each of the top K patent documents are identified.
(8) For each of the identified IPCs, the retrieval scores of the patent documents with the IPC are summed up.
(9) The IPC scores are sorted in descending order. The top X IPCs are assigned to the input abstract text.

In this method, parameters M, N, K and X should be tuned for better classification accuracy.

2.2 Query Expansion Method (QEM2)

In QEM1, term frequency might be disproportionately affected by one particular abstract. Thus, in this method, the average of the term weights (not term frequency) in the retrieved similar paper abstracts is calculated.

We used the following algorithm:
(1) Terms are extracted from an input abstract text using Chasen.
(2) The weight of each term is calculated using a TF-IDF method (IDF is from paper abstracts).
(3) The top M similar paper abstracts are retrieved from a paper abstract database using GETA.
(4) For each term included in these abstracts, the average of the term weights is calculated. The terms are sorted by the average weight in descending order.
(5) The top N terms are identified. The value obtained by multiplying the term weight and a constant Y is added to the original term weight calculated in Step (2).
(6) With the expanded term set from Step (5), similar patent documents are retrieved from a patent document database using GETA (IDF is from patent documents).
(7) The IPCs assigned to each of the top K patent documents are identified.
(8) For each identified IPC, the retrieval scores of the patent documents with the IPC are summed up.
(9) The IPC scores are sorted in descending order. The top X IPCs are assigned to the input abstract text.

In this method, parameters M, N, K, X and Y should be tuned for better classification accuracy.

2.3 Query Expansion Method (QEM3)

It is often said that the terms used in research papers and patents differ from each other, which impairs classification accuracy. In query expansion, it might be useful to add terms used in similar patents rather than those in similar paper abstracts to improve classification accuracy. Thus, in this method, a patent document database was used for query expansion. Patents similar to the input abstract text are retrieved and terms are extracted from the retrieved similar patents.

We used the following algorithm:
(1) Terms are extracted from an input abstract text using Chasen.
(2) The weight of each term is calculated using a TF-IDF method (IDF is from patent documents).
(3) The most similar patent document is retrieved from a patent document database using GETA.
(4) Terms are extracted from the most similar patent using Chasen. The top N terms with higher term weight are identified. The term set from Step (1) is replaced by this term set.
(5) Using the replaced term set from Step (4), similar patent documents are again retrieved from the same patent document database using GETA.
(6) The IPCs assigned to each of the top K patent documents are identified.
(7) For each of the identified IPCs, the retrieval scores of the patent documents with the IPCs are summed up.
(8) The IPC scores are sorted in descending order. The top X IPCs are assigned to the input abstract text.

In this method, parameters N, K and X should be tuned for better classification accuracy.

3 Experiments

3.1. Text Data

We used a patent document database prepared in NTCIR-5 and -6 which includes 3.5 million patent documents from 1993 to 2002. We also used a paper abstract database prepared in NTCIR-1 and -2 which includes 736,166 paper abstracts.

3.2. Parameter Tuning Patterns

For each of the three query expansion methods proposed in this paper, several parameters should be tuned. However, we had little training data to tune these parameters. In this experiment, we used only 97 query abstracts used in the dry run. Since the target classification system consists of 31,520 categories, this amount of training data was too small, which might have caused the mistuning of the parameter values.

Thus, we compared the classification accuracy for the following four kinds of parameter tuning pattern:
(1) Dry run data (97 query abstracts) was used for evaluation. This data was also used for the parameter tuning.

Table 1. Experiment Patterns.

ID	DF calculation		QEM1 (Section 2.1)	QEM2 (Section 2.2)	QEM3 (Section 2.3)	Score Merging
	whole	claim				
HTC01	Used					
HTC02		Used				
HTC03	Used		Used			
HTC04		Used	Used			
HTC05	Used	Used	Used			Used
HTC08	Used			Used		
HTC09		Used		Used		
HTC10	Used	Used		Used		Used
HTC13	Used				Used (first)	
HTC14	Used				Used (second)	

Table 2. Experiment Results (MAP Comparison).

Experiment ID	ID for comparison	Mean Average Precision (MAP)			
		Evaluation data = dry run		Evaluation data = formal run	
		(1)Tuning data =dry run	(2)Tuning data =formal run	(3)Tuning data =dry run	(4)Tuning data =formal run
HTC01	Baseline	0.4060	0.4060	0.4334	0.4334
HTC02	HTC01	0.4083 (+0.0023)	0.4083 (+0.0023)	0.4236 (-0.0098)	0.4236 (-0.0098)
HTC03	HTC01	0.4156 (+0.0096)	0.4061 (+0.0001)	0.4268 (-0.0066)	0.4371 (+0.0037)
HTC04	HTC02	0.4088 (+0.0005)	0.4041 (-0.0042)	0.4165 (-0.0071)	0.4340 (+0.0104)
HTC05	HTC03	0.4217 (+0.0061)	0.4057 (-0.0004)	0.4326 (+0.0058)	0.4388 (+0.0017)
HTC08	HTC01	0.4269 (+0.0209)	0.4036 (-0.0024)	0.4323 (-0.0011)	0.4355 (+0.0021)
HTC09	HTC02	0.4205 (+0.0122)	0.4068 (-0.0015)	0.4227 (-0.0009)	0.4339 (+0.0103)
HTC10	HTC08	0.4315 (+0.0046)	0.4075 (+0.0039)	0.4318 (-0.0005)	0.4397 (+0.0042)
HTC13	HTC01	0.4343 (+0.0283)	0.4343 (+0.0283)	0.4402 (+0.0068)	0.4402 (+0.0068)
HTC14	HTC01	0.3394 (-0.0666)	0.3394 (-0.0666)	0.3862 (-0.0472)	0.3862 (-0.0472)

Note: The values in parenthesis show the MAP difference between ID and ID for comparison.

(2) Dry run data (97 query abstracts) was used for evaluation. Formal run data (879 query abstracts)was used for the parameter tuning.
(3) Formal run data (879 query abstracts) was used for evaluation. Dry run data (97 query abstracts) was used for the parameter tuning.
(4) Formal run data (879 query abstracts) was used for evaluation. This data was also used for the parameter tuning.

Our submitted result set in a formal run was based on pattern (3).

3.3. Experiment Patterns

We did 14 patterns of experiments. In this paper, we report on the 10 patterns shown in Table 1.

HTC01 was a baseline method. To calculate a query term weight, we used the document frequency (DF) obtained from the whole body of patent texts. In HTC02, we used DF obtained only from claim texts in patent documents. Parameter K was set to 40 and X was set to 1000.

HTC03 applied QEM1 (described in Section 2.1) to HTC01, and HTC04 applied it to HTC02. Parameter M was set to 30 in HTC03 and 5 in HTC04. Parameter N was set to 20.

HTC05 merged the score of HTC03 with that of HTC04. We multiplied the score of HTC03 by a constant of 2.0.

HTC08 applied QEM2 (described in Section 2.2) to HTC01, and HTC09 applied it to HTC02. Parameter M was set to 15. Parameter N was set to 15

in HTC08 and 20 in HTC09. Parameter Y was set to 2.0 in HTC08 and 4.0 in HTC09.

HTC10 merged the score of HTC08 and that of HTC09. We multiplied the score of HTC08 by a constant of 2.0.

HTC13 applied QEM3 (described in Section 2.3) to HTC01. In HTC13, the top patent document was used to extract terms. HTC14 applied the same method, but only the second retrieved patent document was used to extract terms.

In this task, each query abstract has a patent used to define correct IPCs. For most query abstracts, GETA could retrieve the patent in the first rank. Though we did not use this patent directly to assign IPCs, this patent is useful for retrieving similar patents from a patent database. In practical cases, however, this patent might not exist or will be unknown. Thus, in HTC14, we did not use the patent in the first rank, but used the patent in the second rank for query expansion. (Since we did not know the patent ID used to define correct IPCs for each query abstract, we regarded the patent in the first rank as this). If QEM3 is an effective method, the accuracy of HTC14 should be better than that of HTC01.

3.4. Results and Discussion

The results for each experiment pattern are shown in Table 2. The behavior of the mean average precision values (MAPs) for parameter tuning patterns (2) and (3) differed from that for (1) and (4). The MAPs of HTC03, HTC04, HTC08 and HTC09 in (2) and (3), which used query expansion methods, were lower than those of the baseline HTC01 and HTC02, but were higher in (1) and (4). This shows that formal run data results in different trends than dry run data with regard to classification accuracy.

The MAPs of HTC05 and HTC10, which used score merging, were better than their comparison targets. This was mainly because the number of IPCs per query was increased by merging two results.

The MAP of HTC13, which used QEM3, was better than that of HTC01. This was because most patent documents used for the definition of correct IPCs for the query paper abstract could be retrieved in the first rank of the similar document retrieval result.

Table 3 compares the MAP by query group. The query data was divided into two groups (see the organizer's report for the details of the definition of these two groups). The MAP of group A (query #300 to #772) was much higher than that of group B (query #1000 to #1405).

Table 4 shows the MAP when patent abstracts and/or whole patent texts were the input. We used 3,464 patent documents (with 9,269 correct IPCs) published in 2002 as input. The same training data was used to identify the IPCs. When patent texts were used as input, the MAP was much higher than when paper abstract was used as input. We think that the MAP difference was due not only to the difference terms used in research papers as opposed to patents but also to the difference in how the correct IPCs were defined.

Table 3. MAP comparison for two query groups.

ID	MAP in formal run	
	Query 300-772	Query 1000-1405
HTC01	0.4922	0.3649
HTC02	0.4806	0.3572
HTC03	0.4729	0.3731
HTC04	0.4677	0.3570
HTC05	0.4776	0.3802
HTC08	0.4817	0.3747
HTC09	0.4727	0.3644
HTC10	0.4823	0.3730
HTC13	0.5062	0.3633
HTC14	0.4370	0.3270

Table 4. MAP comparison between paper and patent.

#	Input text	MAP (Method)
1	Paper abstract (Group A plus B)	0.4402 (NTCIR-7 HTC13)
2	Paper abstract (Group A only)	0.5062 (NTCIR-7 HTC13)
3	Paper abstract (Group B only)	0.3633 (NTCIR-7 HTC13)
4	Patent abstract	0.5722 (NTCIR-6 HTC01)
5	Whole patent text	0.6050 (NTCIR-6 HTC08)

4 Conclusions

We used three kinds of query expansion methods to evaluate the classification accuracy in the NTCIR-7 Patent Mining Task. Though the results showed a slight effectiveness, they were not as good as we expected.

In future work, it would be interesting to use the whole text of a research paper to assign IPCs. Our experiments on patent classification suggest that classification accuracy when using a whole patent text as input is better than when using only a first claim text and/or an abstract text. It would also be interesting to use bibliographic data such as author names, author's affiliations and publication dates to improve classification accuracy. Furthermore, it remains important to consider the difference between various research fields.

References

[1] Chasen: http://chasen.naist.jp/hiki/ChaSen/
[2] GETA: http://geta.ex.nii.ac.jp/e/index.html