Integrating CRF and Rule Method for Knowledge Extraction in Patent Mining Task at NTCIR-8

Jie Gui, Peng Li, Chengzhi Zhang, Ying Li and Zhaofeng Zhang
Institute of Scientific and Technical Information of China, Beijing, 100038
{guij, lipeng_cn, zhangchz, liying, zhangzf}@istic.ac.cn

ABSTRACT

We participate in the subtask "technical trend map creation" of patent mining task at NTCIR-8. In this paper, we define this task as a knowledge extraction task for patent abstracts and the CRF method and Rule method are introduced in our approach. Compare with the evaluation results, we find out the effect of method of integrating CRF model and Rule model is better than that only using CRF model. However, extraction task of <value> tag is more difficult than <technology> tags.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: Abstracting methods

General Terms

Experimentation

Keywords

Patent Mining, Knowledge Extraction, CRF model, Rule model

1. INTRODUCTION

For a researcher in a field with high industrial relevance, retrieving research papers and patents has become an important aspect of assessing the scope of the field. However, the terms used in patents are often more abstract or creative than those used in research papers, to try to widen the scope of the claims. Therefore, the Patent Mining Task aims to develop fundamental techniques for retrieving, classifying, and analyzing both research papers and patents[1].

In addition, many researches focus on content-oriented patent analysis for supporting decision-making of S&T. So, we define the subtask "technical trend map creation" as a knowledge extraction task which will provide the knowledge base for patent content analysis[2][3].

In this paper, our research aims to the English patent mining subtask--technical trend map creation. For this task, we integrate Conditional Random Fields (CRF) method and Rule method into our system for knowledge extraction of patent abstracts. This paper includes four parts: feature selection for CRF model, rule making for tag extraction, the experiments with data sets from dry run, and discussion for experiments.

2. CRF Method

2.1 Basic idea

A conditional random field (CRF) is a type of discriminative probabilistic model most often used for the labeling or parsing of sequential data, such as natural language text or biological sequences. Definition of conditional random is as followed:

Definition: Let $G = (V, E)$ be a graph such that $\mathrm{Y} = (\mathrm{Y}_v)_{v \in V}$, so that **Y** is indexed by the vertices of *G*. Then (**X**, **Y**) is a *conditional random field* in case, when conditioned on **X**, the random variables $\mathbf{Y}_v$ obey the Markov property with respect to the graph: $p(\mathrm{Y}_v \mid \mathrm{X}, \mathrm{Y}_w, w \neq v) = p(\mathrm{Y}_v \mid \mathrm{X}, \mathrm{Y}_w, w \sim v)$, where $w \sim v$ means that *w* and *v* are neighbors in *G*[4].

For English subtask of Technical Trend Map Creation, each topic must be assigned "TECHNOLOGY", "EFFECT", "ATTRIBUTE", and "VALUE" tags. So, the basic idea that we use CRF model is that these tags will be seen as sequential data that CRF model can deal with.

2.2 Feature Selection

In our CRF model, three features are defined: part of speech the word belongs to

Feature1: English article feature. When we analyze <technology> tags, we find out that commonly the first word in a phrase is an English article, such as "a", "an" or "the". For Feature 1, two parameters are defined by judging whether a word is an English article. When it is the article, Parameter of Feature 1 is a value of 1, otherwise the value 0.

Feature2: Word frequency. In this model, we try to calculate word frequency in training data set. For acquiring an objective result, a stemming processing is introduced. In this task, we use open source software--Porter Stemming Algorithm.

Feature3: Classified Information. Classified information indicates the position information that a word appears in the tag. We set seven tags for this classification shown as table 1:

Table 1. Seven tags for Classified Information

TAG	Meaning
S	No special position information
B-TECHNOLOGY	The beginning word of <technology> tag
B-ATTRIBUTE	The beginning word of <attribute> tag
B-VALUE	The beginning word of <value> tag
I-TECHNOLOGY	Word at other position in <technology> tag
I-ATTRIBUTE	Word at other position in < attribute > tag
I-VALUE	Word at other position in <value > tag

2.3 CRF model for the task

According three CRF Features, we construct the CRF model for the task (seen as Figure 1). For running the CRF model, CRF++[1], an open source software tool, is introduced for this task.

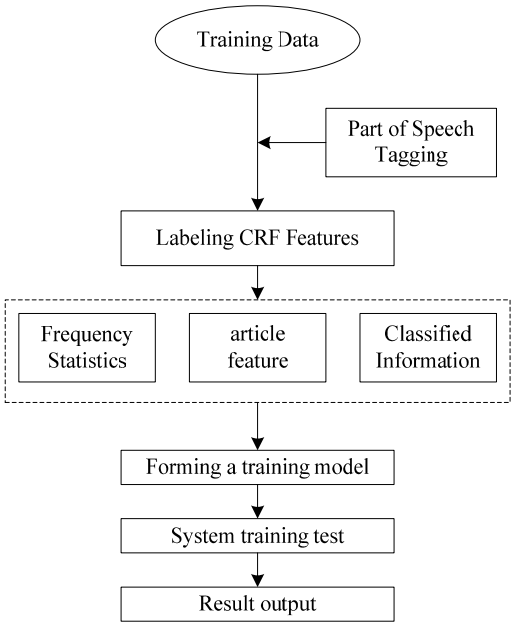

Figure 1. CRF model for the task

3. Rule Method[5]

In the dry run, we only use the CRF model to extract tags from paper and paper topics, but result of dry run is unsatisfactory. In formal run, rule method is introduced. And CRF method and Rule method are integrated to label tags.

By analyzing patent training documents of dry run, an important feature is disclosed that is some words or phrases appear repeatedly in <Technology> tags, for example, include, comprise, use, have, as well as "by means of". A large number of technology related contents would be appeared after the position of these words. Comparing with other patent documents, these words would be connected with the context with novelty of the invention and its technical contents.

According to it, we make some rule template to label <Technology> tag. The rule template used in the formal run is as followed:

Basic rule 1： A hypothesis is that all "Technology" phrases are noun phrase. For instance, the structures of phrases may be: article + adjective + noun or noun or adjective + noun plural, that is a phrase is limited to start from a non-verb.

Basic rule 2： Filtering the phrase "a plurality of" out of topic text when it appears before the sentences.

Rule of includ (include/s, including)

In the followed rules, the characters of A, B, C are defined as the labeled content by rule model for this task.

Rule 1: includ A;B;C

Rule 2: includ A, B, and C

Rule 3: includ A and B

Rule 4: includ (a) A and (b) B or includ (1) A and (2) B

Rule 5: plurality of A are included

Rule 6: plurality of A included in B

Rule of compris (comprise/s, compring)

Rule 1: compris A;B;C

Rule 2: compris A, B, and C

Rule 3: comprising A and B

Rule 4: comprising A, B, and C

Rule of "by means of"

rule 1: by means of A

4. Experiments

Before formal run, we test the CRF model and Rule model by training data and evaluation data of dry run. The test is based on two ideas:

(1) Testify utility of our approach by integrating CRF model and Rule model;

(2) Comparing difference of test results with patent document and research paper.

4.1 Test Collections

For patent mining task at NTCIR-8, sets of topics with manually assigned "TECHNOLOGY", "EFFECT", "ATTRIBUTE", and "VALUE" tags are necessary for training and evaluation. Therefore, NTCIR made a human subject to assign these tags to the following two kinds of texts for English subtask of Technical Trend Map Creation:

- Five hundred English research papers (abstracts)
- Five hundred English patents (abstracts)

The data sets for English Subtask of Technical Trend Map Creation in Patent Mining task are as followed:

Table 2. Data collection for English Subtask of Technical Trend Map Creation in Patent Mining task at NTCIR-8

Data Set	Source	Number
Group1	Patent training data for dry run	250
Group2	Paper training data for dry run	250
Group3	Patent evaluation data for dry run	50
Group4	Paper evaluation data for dry run	50
Group5	Patent evaluation data for formal run	200
Group6	Paper evaluation data for formal run	200

[1] http://sourceforge.net/projects/crfpp/files/

4.2 Evaluation Results

4.2.1 Test results with data sets of dry run

For testing our approach, we design five systems by combining different data sets, the system description is shown in table 3.

Table 3. Experimental systems with data sets from dry run

System	Training Data	Method
TEST-1	Group1+ Group2	Only using "Feature 2+ Feature 3" in CRF Model
TEST-2	Group1	Only using "Feature 2+ Feature 3" in CRF Model
TEST-3	Group1+ Group2	CRF model
TEST-4	Group1	CRF Model
TEST-5	Group1+ Group3	CRF Model + Rule Model

We use the tagged evaluation data of dry run and evaluation tool provided by NTCIR-8 to test the above system, evaluation results are shown in table 4. The more detailed evaluation information is shown in appendix 1.

Table 4. Experimental result with evaluation tool from dry run

System	MAP (Recall)	MAP (Precision)	MAP (F)
TEST-1	0.130	0.343	0.189
TEST-2	0.163	0.366	0.226
TEST-3	0.134	0.349	0.194
TEST-4	0.167	0.362	0.228
TEST-5	0.239	0.438	0.309

4.2.2 Evaluation results of formal run

We summit the formal run results, and table 5 and table 6 show the official evaluation results of formal run provided by NTCIR-8.

Table 5. Formal run evaluation result for patent

System	MAP (Recall)	MAP (Precision)	MAP (F)
ISTIC-1	0.224	0.432	0.295
ISTIC-2	0.221	0.447	0.295
ISTIC-3	0.102	0.436	0.165
ISTIC-1-1	0.239	0.438	0.309
ISTIC-2-1	0.223	0.423	0.292

Table 6. Formal run evaluation result for paper

System	MAP (Recall)	MAP (Precision)	MAP (F)
ISTIC-1	0.064	0.405	0.11
ISTIC-2	0.043	0.324	0.076
ISTIC-3	0.058	0.425	0.102

4.3 Discussion

Compare with the evaluation results, we make the followed discussions:

(1) For our evaluation results, the systems only using patent as training set are better than those using patent and research paper data sets. It may be the reason that we focus on patent data more than research paper, especially rule model all come from the analysis results for patent documents.

(2) The quality of labeling tags manually makes a greater impact on our CRF model.

(3) When adding "article" feature into CRF model, the results are slightly improved.

(4) The results of < ATTRIBUTE > and < VALUE > tags are not satisfactory, and the main reason may be that the key features of these two tags are not obvious and have little relations with part of speech.

(5) The effect of using improved method for formal run integrating CRF model and Rule model is better than the method only using CRF model for dry run.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Technology R&D Program (No. 2006BAH03B03) and Research Fund Project of Institute of Scientific and Technical Information of China (No. 2009KP01-7-1)

REFERENCES

[1] Hidetsugu Nanba, Atsushi Fujii, Makoto Iwayama, Taiichi Hashimoto. Overview of the Patent Mining Task at the NTCIR-8 Workshop. Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2010.

[2] Tseng Y H，Lin C I，LiII Y L．Text Mining for Patent Map Analysis ． IACIS Pacific 2005, Conference Proceedings．2005.

[3] L. Wanner et al. Towards content-oriented patent document processing, World Patent Information, 2008, 30, 21–33.

[4] Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Proc. 18th International Conf. on Machine Learning*, Morgan Kaufmann, San Francisco, CA (2001), 282–289.

[5] Daisuke ISHIKAWA, Hidehiro ISHIZUKA, Norihiko UDA and Yuzuru FUJIWARA, Extraction and Integration of Causal Relationships in Patent Documents: Summary and A Subsequent Activity, Japan Society of Information and Knowledge, Vol.15, No.3, 2005.

Appendix

Table 7. Experience result with evaluation tool from dry run

System	Object Tag	Recall	Precision	F

TEST-1	Title_technology	0.444	0.667	0.533
	Abstract_technology	0.131	0.3	0.182
	Abstract_attribute	0.03	0.333	0.056
	Abstract_value	0.143	0.667	0.235
	Abstract_effect	0.133	0	0
TEST-2	Title_technology	0.556	0.714	0.625
	Abstract_technology	0.16	0.311	0.212
	Abstract_attribute	0.061	0.5	0.108
	Abstract_value	0.179	0.833	0.294
	Abstract_effect	0.167	0	0
TEST-3	Title_technology	0.444	0.667	0.533
	Abstract_technology	0.136	0.308	0.189
	Abstract_attribute	0.03	0.333	0.056
	Abstract_value	0.143	0.667	0.235
	Abstract_effect	0.133	0	0
TEST-4	Title_technology	0.556	0.714	0.625
	Abstract_technology	0.16	0.303	0.21
	Abstract_attribute	0.061	0.5	0.108
	Abstract_value	0.214	0.857	0.343
	Abstract_effect	0.2	0	0
TEST-5	Title_technology	0.359	0.304	0.329
	Abstract_technology	0.294	0.429	0.349
	Abstract_attribute	0.061	0.464	0.108
	Abstract_value	0.172	0.63	0.27
	Abstract_effect	0	0	0