NTCIR-8 Patent Mining Task at Toyohashi University of Technology

Yusuke Suzuki
Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho,
Toyohashi, Aichi, 441-8580, Japan
+81-532-44-6867
y-suzuki @smlab.tutkie.tut.ac.jp

Hirofumi Nonaka
Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho,
Toyohashi, Aichi, 441-8580, Japan
+81-532-44-6867
nonaka@smlab.tutkie.tut.ac.jp

Hiroki Sakaji
Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho,
Toyohashi, Aichi, 441-8580, Japan
+81-532-44-6867
sakaji@smlab.tutkie.tut.ac.jp

Akio Kobayashi
Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho,
Toyohashi, Aichi, 441-8580, Japan
+81-532-44-6867
kobayashi@smlab.tutkie.tut.ac.jp

Hiroyuki Sakai
Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho,
Toyohashi, Aichi, 441-8580, Japan
+81-532-44-6867
sakai@smlab.tutkie.tut.ac.jp

Shigeru Masuyama
Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho,
Toyohashi, Aichi, 441-8580, Japan
+81-532-44-6867
masuyama@tutkie.tut.ac.jp

ABSTRACT

Our group took part in the Patent Mining Task of the NTCIR-8. We proposed an extraction method of EFFECT and TECHNOLOGY expressions from a patent, respectively.

In order to extract TECHNOLOGY expressions, we developed a method that uses Support Vector Machine and delimiters collected by using entropy-based score.

On the other hand, our method for annotation of EFFECT tags is based on delimiters using entropy-based score.

We achieved accuracy of precision 0.55 and recall 0.27, F-measure 0.36, respectively.

Keywords

Entropy, Delimiter, Support Vector Machine, Patent Map

1. INTRODUCTION

A patent map, a visual representation of related patent information, is an effective strategic tool for analysis of technology trends. In particular, a technology-effect type patent map(Fig. 1) that uses the technology and the effect expression defined by NTCIR-8 PATMN Task(example shown in Fig. 2) is commonly used as patent examinations are based on technology terms and effects of the inventions. Currently, a technology-effect type patent map is manually made. Therefore, in order to reduce time consumption of this task, it is necessary to develop some methods that automatically generate technology-effect type patent maps.

However, many conventional patent analysis methods did not focus on generation of technology-effect type patent maps.

For example, in order to recognize progress of technologies as patent text is an ample resource to discover technological progress, Porter [1] employed a text-mining approach to generate a patent network as an analytical tool to recognize emerging technologies [2]. Studies on patent search aim an effective navigator to find the desired patent. Takaki [3] analyzed claim structures to improve the effectiveness of the search task. Itoh [4] improved the effectiveness of the technology survey task by using the different term distributions. Koster [5] investigated the effectiveness of the bag-of-words approach in classifying patents. Lai [6] used citation-based analysis to perform a patent classification. Chakrabarti [7] analyzed the diffusion of technical information in different organizations. Shinmori [8] aimed to improve the readability of patent claims, and proposed a method for analyzing the rhetorical structure. Uchida [9] develops a method for automation of patent map (that is not a technology-effect type patent map) generation using Concept-based Vector Space Model.

In order to extract the technology and the effect expression defined by NTCIR-8 PATMN Task, we proposed a method shown by the following steps.

(Step 1-1) Annotate long TECHNOLOGY tags by using delimiters extracted by using entropy-based score.

(Step 1-2) Annotate short TECHNOLOGY tags by using delimiters extracted by using entropy-based score.

(Step 2-1) Annotate VALUE tags by using delimiters that are extracted by using entropy-based score.

(Step 2-2) Annotate ATTRIBUTE tags by using heuristics based on Japanese particles.

This paper is organized as follows: We present our method in Section 2. Based on this method, we show the evaluation including the experimental results and related discussions in Section 3. Section 4 concludes this paper.

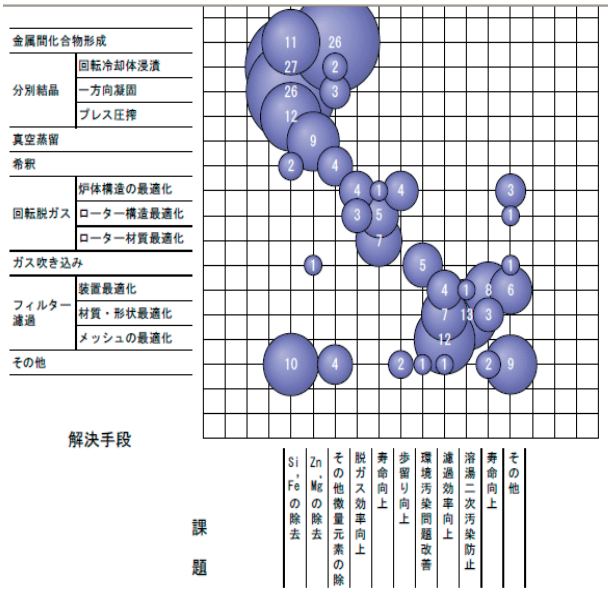


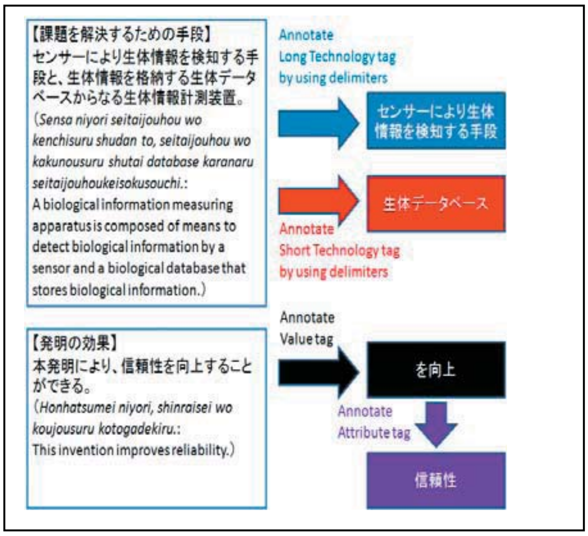
Fig. 1. Technology-effect type patent map

PM 磁束制御用コイルを設けて<Technology>閉ループフィードバック制御</Technology> を施すため、<Effect><Attribution>電力損失</Attributuion>を<Value>最小化</Value></Effect>できる。

Fig. 2. Example of a Technology tag and an Effect tag defined by NTCIR-8 PATMN task.

2. Our proposed Method

We developed a method for annotation of a technology tag and an effect tag, respectively. Overview of our method is shown in Fig. 3.


【課題を解決するための手段】
センサーにより生体情報を検知する手段と、生体情報を格納する生体データベースからなる生体情報計測装置。
(*Sensa niyori seitaijouhou wo kenchisuru shudan to, seitaijouhou wo kakunousuru shutai database karanaru seitaijouhoukeisokusouchi.:*
A biological information measuring apparatus is composed of means to detect biological information by a sensor and a biological database that stores biological information.)

【発明の効果】
本発明により、信頼性を向上することができる。
(*Honhatsumei niyori, shinraisei wo koujousuru kotogadekiru.:*
This invention improves reliability.)

Fig. 3. Overview of our method

2.1. Annotation of technology tag

2.1.1. Definitions of long TECHNOLOGY tag and short TECHNOLOGY tag

First, we begin with the following definitions to be used in this paper.

Definition 1 (Long technology tag): A long technology tag is an expression annotated with a technology tag consisting of more than one word such as “反射光を光検知器に集光する手段(*hanshako wo hikarikenchiki ni shukou suru shudan* denotes means focusing light to a photo-detector)”.

Definition 2 (Short technology tag): A short technology tag is an expression that is annotated with technology tag consisting of only one word such as “ヘリウムガス (*heliumu gasu* means helium gas”.

2.1.2. Tagging long TECHNOLOGY tags

Our method for tagging long TECHNOLOGY tags is as follows.

Step 1 Collect delimiters each of whose entropy is higher than its threshold value.

Step 2 Annotate technology tags that are enclosed in two delimiters.

At Step 1, we collect delimiters that frequently appear before/after a TECHNOLOGY tag such as “おいて、”, “と、”. Here, we divide delimiters into two groups. One of them appear before a TECHNOLOGY tag (we called the first delimiter group) and another group appear after a TECHNOLOGY tag (we called the second delimiter group). We select appropriate delimiters from a set of candidates that appears before/after a TECHNOLOGY tag. Therefore, we calculate entropy H(t) based on P(t) that is the probability that candidate word t appears before/after TECHNOLOGY tag on each document. If entropy H(t) is large, candidate word t appears uniformly before/after a TECHNOLOGY tag on each document and these candidates are appropriate. Entropy H(e) is calculated by the following Formula 1:

$$H(t) = -\sum_{t \in S(t)} P(t) \log_2 P(t) \qquad ・・・(1)$$

Note that the selection of the first delimiters and the second delimiters is the same as the above, where $S(t)$ is the set of words that appear after TECHNOLOGY tag.

We collect delimiters each of whose entropy is higher than the threshold value.

The threshold value of entropy is experimentally-determined by using F-measure on training data set.

The results of extracting delimiters are shown in Table 1.

Table 1. The result of extracting delimiters

Group name 1: first delimiter group 2: second delimiter group	Delimiters
1	は、(*to,*)
1	において、(*nioite,*)
1	ば、(*ba,*)
2	と、(*to,*)
2	とを(*towo*)
2	設け(*mouke*)
2	備え(*sonae*)
2	有し(*yushi*)
2	有する(*yusuru*)
2	用い(*mochii*)
2	介し(*kaishi*)
2	含み(*fukumi*)
2	含む(*fukumu*)

At Step 2, we tag TECHNOLOGY by using delimiter extracted at Step 1.

First, we extract expression *e* that is enclosed by the second delimiter group or a section between the beginning of sentence and a second group delimiter.

Second, we delete words that appears before the first delimiters from expression *e* (this expression is defined as *e'*)and we assign tag TECHNOLOGY before/after expression *e'*.

2.1.3. Tagging short TECHNOLOGY tags

Our method for tagging short TECHNOLOGY tags is as follows.

Step 1 Extract noun phrase sw_i from the solution to the problem tag on each patent document as candidates for a short TECHNOLOGY expression.

Step 2 Calculate the feature space for Support Vector Machine[10] that consists of (1)character proportion of alphabet and *katakana* in sw_i , (2) delimiters shown in Table 2. The weight of feature (2) on delimiter *j* is 1 when delimiter *j* is appeared after sw_i, otherwise 0.

Step 3 Extract a short TECHNOLOGY expression by using SVM.

Table 2. delimiters

Delimiters
と(*to*)
とを(*towo*)
設け(moke)
備え(*sonae*)
有し(*yushi*)
有する(*yusuru*)
用い(*mochii*)
介し(*kaishi*)
含む(*fukumu*)

2.2.Annotation of effect tags

2.2.1. Tagging VALUE tags

Our method for tagging VALUE tags is as follows.

Step 1 Collect an expression that is enclosed by VALUE tags and these expressions are candidates of VALUE expressions.

Step 2 Select an appropriate VALUE expression from the candidates by using entropy-based score.

The method for extracting content words as effective feature words is described as follows:

Score $W_p(t_i,S_p)$ of a candidate word t_i in a positive data set S_p is calculated using the following formula 2:

$$W_p(t_i,S_\mathrm{p}) = P(t_i,S_\mathrm{p})H(t_i,S_\mathrm{p}) \quad \cdot\cdot\cdot (2)$$

Here, we define "positive" as the case that a word is enclosed by VALUE tags.

Similarly, score $W_n(t_i,S_n)$ of a candidate word t_i in a negative data set S_n is calculated using the following formula 3:

$$W_n(t_i,S_\mathrm{n}) = P(t_i,S_\mathrm{n})H(t_i,S_\mathrm{n}) \quad \cdot\cdot\cdot (3)$$

Here, we define "negative" as the case when a word is not enclosed by VALUE tags.

$P(t_i, S_p)$ is the probability that t_i appears in S_p. The entropy $H(t_i, S_p)$ implies that a bias of probability distribution is calculated using the following Formula 4:

$$H(t_i,S_p) = \sum_{d \in S} P(t_i,d)\log P(t_i,d) \quad \cdot\cdot\cdot (4)$$

In our method, the value of $W_p(t_i,S_p)$ is compared with that of $W_n(t_i,S_n)$. If $W_p(t_i,S_p)$ is higher than $2W_n(t_i,S_n)$, which implies that t_i is biased toward S_p, t_i is selected as an appropriate VALUE expression.

A word that is certainly enclosed by VALUE tags such as "向上" may assign a high $W_p(t_i,S_n)$ and a low $W_n(t_i,S_n)$. Therefore, we compared value of the positive data set $W_p(t_i,S_n)$ with that of the negative data set $W_n(t_i,S_n)$.

2.2.2. Tagging ATTRIBUTE tags

After we extract VALUE expression, we tag ATTRIBUTE tags as follows.

Step 1 Extract *bunsetsu* w_{ai} that appears before a VALUE expression and the word is selected as candidate of ATTRIBUTE expression *Att.*

Step 2 Extract *bunsetsu* w_{aj} that appears before the candidate of ATTRIBUTE expression *Att*.

Step 3 Add *bunsetsu* w_{aj} to *Att* when *bunsetsu* w_{aj} includes Japanese particles "の", "を", "が", "や", "および", and "及び", otherwise select *Att* as an appropriate ATTRIBUTE expression.

Step 4 Repeat Steps 2 and 3.

3. Results and Disscussion

We apply our proposed method to an evaluation data set of NTCIR-8 formal run. The experimental results of annotating TECHNOLOGY tags, VALUE tags, ATTRIBUTE tags, EFFECT tags are shown in Tables 3, 4, 5, 6, respectively. Moreover, general performance of our method is shown in Table 7. Furthermore, we show the result of comparison with other participants in Table 8.

As a result, we achieved high precision on each task, in particular, task of annotating VALUE tags. On the other hand, recall of our method is a low value. One of the reason is that the collection of appropriate delimiters and heuristic rules are non-exhaustive.

Table 3. The result of TECHNOLOGY tag

Recall	Precision	F-measure
0.32	0.48	0.38

Table 4. The result of VALUE tag

Recall	Precision	F-measure
0.3	0.83	0.44

Table 5. The result of ATTRIBUTE tag

Recall	Precision	F-measure
0.18	0.49	0.26

Table 6. The result of EFFECT tag

Recall	Precision	F-measure
0.16	0.41	0.23

Table 7. The result of our method

Recall	Precision	F-measure
0.27	0.55	0.36

Table 8. Comparison of performance

Team	Recall	Precision	F-measure
HCU	0.43	0.55	0.48
HTC	0.22	0.38	0.28
ONT	0.21	0.37	0.27
TRL	0.44	0.51	0.47
Our proposed Method	0.27	0.55	0.36

4. Conclusion

In this paper, we proposed a method to annotate technology tags by using delimiters and machine learning. We also developed an annotation method for effect tags by using delimiters. In order to choose a delimiter, we use a score based on entropy. We evaluated our method and it attained 0.55 precision and 0.27 recall.

References

[1] A. Porter and D. Jhu. Technological mapping for management of technology. In Proceedings of International Symposium on Technology, 2001.

[2] W. Pottenger and T. Yang. Detecting emerging concepts in textual data mining. Computational Information Retrieval, pp.1-17, 2001.

[3] T. Takaki, A. Fujii, and T. Ishikawa. Associative document retrieval by query subtopic analysis and its application to invalidity patent search. In Proceedings of the 13th ACM International conference on Information and Knowledge Management (CIKM ’04) , pp.399-406, 2004.

[4] H. Itoh, H. Mano, and Y. Ogawa, Term distillation in patent retrieval. In Proceedings of the ACL-03 workshop on patent corpus processing, pp. 41-45, 2003.

[5] C. Koster, M. Seutter and J. Beney. Multi-Classification of Patent Applications with winnow. In Proceedings PSI 2003, pp.545–554, 2003.

[6] K. Lai, and S. Wu. Using the patent co-citation approach to establish a new patent classification system. Information Processing and Management, Vol. 41, pp.313–330, 2005.

[7] A. Chakrabarti, I. Dror, and N. Eakabuse. Interorganizational transfer of knowledge: An analysis of patent citations of a defense firm. IEEE Transactions on Engineering Management, Vol. 40 (1), pp. 91–94, 1993.

[8] A. Shinmori, M. Okumura, Y. Marukawa, and M. Iwayama. Patent claim processing for readability: structure analysis and term explanation. In Proceedings of the ACL-03 workshop on patent corpus processing, pp.56–65, 2003.

[9] H. Uchida, A. Mano and T. Yukawa, “Patent Map Generation Using Concept-Based Vector Space Model”, Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization,2004

[10] C. Cortes, V. Vapnik, “Support-vector networks”, Mach.Learn.20, pp. 273-297, 2001