

Issues of the Morphological Analysis in Comparison with the Compound Noun Extraction Analysis for a Patent Document

Kyoko Yanagihori, Kazuhiko Tsuda

Graduate School of Business Sciences, University of Tsukuba

Keywords:

text mining
patent search
compound noun
information retrieval
similarity calculation
morphological analysis

ABSTRACT

Compound nouns are frequently encountered in the claims of a patent application. We compared the use of compound noun analysis to morphological analysis as a search method for similar documents in patent applications. This paper focused on the claims written in the Jepson format with consideration to Japanese language claims. Our analysis indicated that the co-occurrence frequency between morphemes and compound nouns in claims is significantly different, where the recurrence of compound nouns is significantly less than morphemes. Although this proved to be a useful feature in precision searches, it was necessary to extend the meaning of compound nouns to include a wider range of similar documents. This was accomplished with the construction of a preliminary semantic dictionary. An important feature discovered during the analysis was that the position of a compound noun in a claim affects the meaning of the noun, thus affecting the search results.

*Copyright © 2013 Information Systems International Conference.
All rights reserved.*

Corresponding Author:

Kyoko Yanagihori,
Graduate School of Business Sciences, University of Tsukuba
Email: kyoko@gssm.otsuka.tsukuba.ac.jp

1. INTRODUCTION

The patent claim is the most important step in the patent application process. A claim has a defined scope of rights pertaining to an invention. Therefore, if the contents of a claim is not well understood, a prior art search may not proceed efficiently. Prior art search is an investigation conducted by an applicant prior to patent application to determine whether another similar patent application exists. Despite the possibility of a similar patent, it is possible to miss it during the prior art search because of insufficient research. Therefore, it is necessary to understand what is contained within a claim to conduct an effective search. For example, it is difficult to understand claims written in Japanese because of the following issues:

- 1) Claims are written in one sentence with no punctuation.
- 2) Demonstrative pronouns are not used in a claim; i.e., the same words appear repeatedly.
- 3) Intentional and deliberate use of compound nouns is frequent in claims.
- 4) In addition, there are no spaces between words in Japanese sentences.

In this study, we find similar documents by applying compound noun analysis rather than morphological analysis to show the validity of noun analysis. We clarify the problems encountered in similar patent searches using compound nouns and describe the solution to these problems.

Research has recently been conducted on patent document analysis because of its importance to the Japanese industry and business interests [1]. The research is introduced for evaluation with a variety of methods using test collections, including the patent document, and provides a summary of results from the patent search.

A study of patent documents using text mining is conducted in languages other than Japanese. It is easier to find a similar claim in copyright infringement cases in Chinese patent documents by clustering the document structure and creating a self-organizing map [2]. The use of semantic analysis and parsing to build patent processing tools increases the overall productivity of patent attorneys or patent analysts involved in claims infringement and validity and quality analysis [3]. The evaluation of the similarity between two claims

on the basis of syntactic and semantic matching of the natural language text using neuro-linguistic programming (NLP) is a valuable tool, which combines symbolic grammar formalisms with data-intensive methods while enhancing analysis robustness to analyze patent claims. This methodology can be used in any patent-related application, such as machine translation, improving readability of patent claims, information retrieval, extraction, summarization, and generation [4]. Another method parses and annotates every sentence in a claim and extracts the desired semantic information from the text using regular expression pattern matching techniques [5]. One of the methods to assist a patent search involves an analysis of a large amount of patent literature, summary extraction, co-occurrence word extraction, and the title based on data clustering [6]. In the above method, along with creating a map of patented technology, the authors assert that the method can be used in prior art search and patent classification. Finally, a summary of existing search tools is discussed in ref. [7].

The majority of proposed methods and tools for patent search assistance involve performing semantic analysis and other analyses on the structure of claims. However, we have not found any previous study on the relationship between the position of a word and words extracted from the claims.

1.1. Overview of patent applications

There are many ways to write a claim. The Jepson format is one possible representation of a claim. In Japan, many claims are written in this format. In this format, a claim is described by a known part (preamble), characterizing part (body), and subject part (theme). Figure 1 shows an example of the Jepson format.

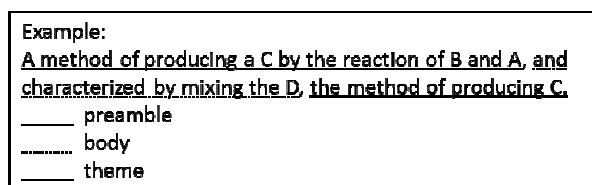


Figure 1. Example of a Jepson claim

Patent applicants use an International Patent Classification (IPC) number to classify an invention according to different categories. The IPC number is used internationally, and is used to classify a patent according to its technical contents. The classification is performed depending on the body of knowledge relevant to a possible invention and is divided into eight sections. “Sections” represent the highest level of hierarchy of the classification.

SECTION A: HUMAN NECESSITIES

SECTION B: PERFORMING OPERATIONS; TRANSPORTING

SECTION C: CHEMISTRY; METALLURGY

SECTION D: TEXTILES; PAPER

SECTION E: FIXED CONSTRUCTIONS

SECTION F: MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING

SECTION G: PHYSICS

SECTION H: ELECTRICITY

Each section is divided into levels that represent inherited attributes in the hierarchical level of the classifications, and are referred to as a subclass, main group, and subgroup followed by lower class, if required. Figure 2 shows the levels of the IPC classification number H04M11/00.

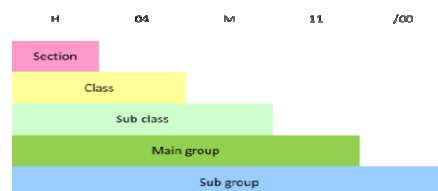


Figure 2. Hierarchy of the IPC number H04M11/00

A patent search unit is also present in the patent office of each country, and each office has access to a search system that is provided by the World Intellectual Property Organization (an example of a search screen is shown in Figure 3). We can search for related words and IPC numbers as search terms, but the original patent documents are not available for display.

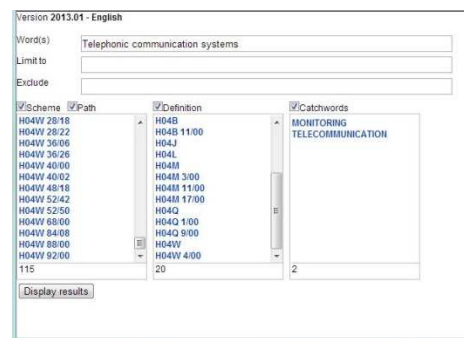


Figure 3. Search screen example of the World Intellectual Property Organization

2. RESEARCH METHOD

2.1. Comparison of morphological and compound noun analyses

We extracted patent claims from the Industrial Property Digital Library (IPDL: <http://www.ipdl.inpit.go.jp/homepg.ipdl>). In particular, we extracted 1200 claims from the Japanese Unexamined Patent Application Publication. These were collected in sets of 300 from the IPC section fields H, A, F, and C. Morphological analysis is a technique that separates the components of speech for each sentence. We extracted only nouns in this study. For compound noun analysis, we use “*TermExtract*” [8], [9] for compound extraction and “*termmi*” [10] to evaluate similarities in a document. Here we assume that the documents in the same IPC section are relevant. We define similarity by assuming that cosine similarity between similar sentences is not zero. The effectiveness of this assumption is proven in the results achieved by applying it to morphological and compound noun analyses. Individual words were weighted using the term frequency–inverse document frequency (tf–idf) method [11]. Figure 4 shows the average precision and recall of the claims in each IPC section using compound noun and morphological analyses.

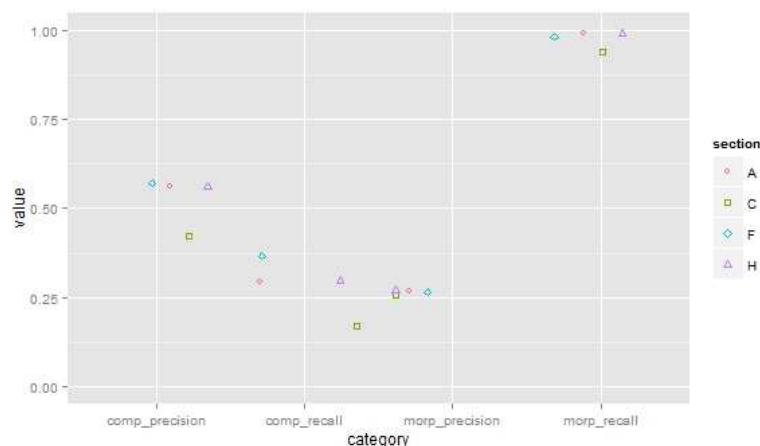


Figure 4. Average precision and recall of claims in each IPC section

2.2. Tests to locate similar documents in IPDL

As an example of a search using the IPDL database, we search for documents similar to a specific patent document. In this case, we use the invention number 2010-094649, and a document for invention 2007-90218 is cited as a similar document by a patent examiner.

Original document

[Application number] 2010-094649

[Title of invention] bubbler

Cited document (Document similar to the original)

[Application number] 2007-90218

[Title of invention] Organic wastewater treatment facilities and organic wastewater treatment method

We apply the IPDL search. When we independently search using the key word “air supply device” and the IPC number of the original document, neither search produces similar documents, as shown in Table 1. This indicates the need for a more refined or rigorous search method during a prior art search.

Table 1. No relevant documents are found in IPDL even when using a keyword search (results in the bottom line)

search method	search term	IPC number	synonyms of search term
	空気供給装置 Air supply device	(B01/F3/12)or (B01/D21/02)or (C02F1/60)	超微細気泡発生装置 Ultra fine bubble generation device
Number of results	424	1031	13
presence or absence of relevant documents	-	-	○

2.3. Creating a semantic extension dictionary of compound nouns

When a patent examiner has judged final rejection of a patent application, other documents are cited as similar documents. On the basis of this process, we created a semantic dictionary using rejected patent information. We select a rejected patent application (IPC H04M11/00) as an example. We extract compound nouns that appear in each document (the rejected document and similar documents). The compound nouns are compared using the algorithm shown in Figure 5.

```

Given an input compound noun(X,Y)
X=m1n2...nl (l=noun count)
Y=n1m2...ml (l=noun count)

Compared an ml(i:1→m) with nl (i:1→n)

STEP1 LOOP n1=nl i:1→m
IF n1=nl
  X=Y /* possibility of similar X and Y */
ELSE IF n1≡m1 ∨ n1≡m2
  X=Y /* possibility of semantic similar X and Y */
ELSE
  n1<>m1
  X≠Y /* possibility of non-similar X and Y */
END OF LOOP

STEP2 LOOP m1=n1 i:1→n
IF m1=nl
  X=Y /* Possibility of similar X and Y */
ELSE IF m1≡n1 ∨ m1≡n2
  X=Y /* possibility of semantic similar X and Y */
ELSE
  m1<>n1
  X≠Y /* possibility of non-similar X and Y */
END OF LOOP

```

Figure 5. Algorithm for compound noun comparison

3. RESULTS AND ANALYSIS

3.1. Availability of compound noun

A morphological search is affected by the frequency of the appearance of a word within a document. Therefore, the recall value using morphological analysis is higher than that using compound noun analysis. Morphological analysis results in a range of similar nouns that is significantly large to perform a precision search.

On the basis of the analysis of the 1200 Japanese claims, the “useless” category in Table 2 indicates that morphological analysis produces many documents that are incorrectly considered to be similar to the application document.

Table 2. Results of average of recall, precision, and useless claims

section	morp_recall	comp_recall	morp_precision	comp_precision	morp_useless %	comp_useless %
H	0.9925	0.2967	0.2697	0.5600	72.4932	33.4967
A	0.9925	0.8623	0.2612	0.7428	73.6230	22.8779
C	0.9359	0.1732	0.2547	0.4231	74.0891	44.0067
F	0.9818	0.3661	0.2637	0.5704	73.4974	35.0428

In contrast, the use of compound nouns does not have a problem with precision but recall is low. Thus, it is necessary to process compound nouns in order to increase recall. This is accomplished through the use of a semantic dictionary to define synonymous relationships among compound nouns.

We examine the compound nouns extracted from each of the 1200 claims previously discussed and focus on the word “device” as used in the terms “ultra-fine bubble generation device” and “air supply device.” By considering the two compound nouns as synonymous, it is possible to find relevant documents. Thus, it is necessary to build a dictionary for compound nouns to determine whether a concept is synonymous or has semantic similarity to another concept. This has the effect of increasing the range of potential similar documents.

3.2. Creating a semantic extension dictionary of compound nouns

A semantic dictionary that attempts to provide synonymous relationships among compound nouns is shown in part in Figure 6.

ID	compound_ORIGINAL	number of noun	composition	appearance position	synonyms	number of noun	composition	appearance position	match1	category	comment1	match2	category	comment2	match3	category	comment3
G	外部デバイス	2	japson	body	外部端末	2	itemization	body	デバイス端末	同義							
S	携帯端末手段	3	japson	body	携帯手段	3	japson	body	携帯手段	同義							
K	監視データ発生時刻	4	japson	body	発生時刻発生時刻	4	japson	body	監視データ発生データ	※		発生時刻発生時刻	概念一致				
K	携帯電話用制御装置	3	japson	body	制御装置	3	enumeration	body	携帯電話制御	類似	携帯電話と制御の区別はどのよう?						
K	コンテンツデータ	2	itemization	body	広告情報	2	enumeration	body	コンテンツ/広告	概念一致	コンテンツ/広告						
R	利用開始情報	3	japson	body	サービス開始時刻	4	japson	body	利用/サービス提供	※	利用/サービス提供の区別はどのよう?	開始時刻/終了時刻	※	時刻データ			
M	文字列画像	3	japson	body	文字列記号帳	4	japson	body	画像/記号帳	概念一致		文字列/文字行	同義				
N	ネットワークノード	2	japson	body	サービス提供機器	3	japson	body									
D	音声ファイル化力	6	itemization	body	ノードデータ	2	japson	body	1/2/3/4/5/6	概念一致		ファイル/データ	概念一致				
T	転送機能	2	enumeration	body	発信側制御手段	4	japson	body	転送/通信								
D	電圧制御手段	4	itemization	body	電圧制御手段	4	japson	preamble	電圧/電圧	概念一致							
Y	優先度制御手段	4	japson	body	FAX通信手段	6	japson	body	制御/制御	類似	優先度/優先度の優先						
R	リサイズ手段	2	itemization	body	再符号化	3	itemization	body	リサイズ/再符号化	※							
D	サーバ管理手段	4	japson	body	個人管理データベース	5	japson	body	記録型/データベース	概念一致	リサイズ/優先度の優先						

Figure 6. Part of the semantic dictionary

The dictionary is based on concept match, broader concept, and subordinate concept. In this result, this is 65% of matches. However, in order to match more precisely, estimation method from the syntax analysis is required. It is considered necessary match of more than 80% in order to practical use. Therefore, In the future, the size of the semantic dictionary, there is a need to expand as much as possible to each IPC classifications.

3.3. Relationship of position of the compound noun

Without considering the position of the compound noun, it was found that the retrieval accuracy decreases. Using a keyword search, we investigated whether “air supply device” appeared in any position in the Jepson format. In 424 of these claims found in the IPDL search, a feature of the keyword appeared in the preamble 81.25% and in the central part 23.68% of the claim document. In other words, even when using the same compound noun, there are cases where the meaning of the noun changes depending on where it appears within a document.

4. CONCLUSION

Compound noun analysis is useful to capture the characteristics of text for precision searches. The frequency of compound nouns that appear in the same document is less than morphemes, and their similarity is lower when searching for similar documents during the evaluation of the frequency of occurrence of words. The number of characters in a Japanese claim is approximately 300 words. In order to perform an effective search by using limited number of characters in a claim, it is important to determine the position of a compound noun in a claim. Thereafter, because the probability of co-occurrence of compound nouns is smaller than that of morphemes, it is necessary to extend search parameters to some extent. Thus, a semantic dictionary is constructed for this purpose. In the future, we will continue to build the compound noun semantic dictionary and prove its value by performing similar document searches according to IPC classifications.

ACKNOWLEDGEMENTS



We appreciate the permission to use the term “*TermExtract*” for the extraction system jointly developed by the Mori Laboratory of Yokohama National University and Nakagawa laboratory of the University of Tokyo. We also express the deepest appreciation to Mr. A. Maeda of the National Institute of

Informatics, who informed us of “*termmi*.” Finally, special thanks to the members of the Tsuda seminar group.

REFERENCES

- [1] A. Fujii, M. Iwayama, N. Kando, "Introduction to the special issue on patent processing," *Information Processing & Management*, vol. 43, No. 5, pp. 1149-1153, 2007.
- [2] S. H. Huang, H. R. Ke, W. P. Yang, "Structure clustering for Chinese patent documents," *Expert Systems with Applications*, vol. 34, No. 4, pp. 2290-2297, 2008.
- [3] K. V. Indukuri, A. A. Ambekar, A. Sureka, "Similarity analysis of patent claims using natural language processing techniques," in *IEEE International Conference on Computational Intelligence and Multimedia Applications*, 2007, pp. 169-175.
- [4] S. Sheremetyeva, "Natural language analysis of patent claims," in *Proceedings of the ACL-2003 workshop on Patent corpus processing*, Sapporo, Japan, Association for Computational Linguistics, 2003, vol. 20, pp. 66-73.
- [5] S. Y. Yang, S. Y. Lin, S. N. Lin, C. F. Lee, S. L. Cheng, V. W. Soo, "Automatic extraction of semantic relations from patent claims," *International Journal of Electronic Business*, vol. 6, No. 1, pp. 45-54, 2008.
- [6] Y. H. Tseng, C. J. Lin, Y. I. Lin, "Text mining techniques for patent analysis," *Information Processing & Management*, vol. 43, No. 5, pp. 1216-1247, 2007.
- [7] M. Fattori, G. Pedrazzi, R. Turra, "Text mining applied to patent mapping: a practical business case," *World Patent Information*, vol. 25, No. 4, pp. 335-342, 2003.
- [8] TermExtract[homepage on the Internet].University of Tokyo;2003[updated 2013 June 12; cited 2013 June 14]. Available from: <http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html>
- [9] H. Nakagawa, T. Mori, "A Simple but Powerful Automatic Term Extraction Method", *Proceedings of the Computerm2: 2nd International Workshop on Computational Terminology*, Taipei, Chinese Taipei, COLING-2002 WORKSHOP, 2002, pp. 29-35.
- [10] termmi[homepage on the Internet].University of Tokyo;2003[updated 2013 June 12; cited 2013 June 14]. Available from: <http://ns.dl.itc.u-tokyo.ac.jp/termmi.html>
- [11] K. Kita, K. Tsuda, M. Shishibori. *Information retrieval algorithms*. Tokyo: Kyoritsu Shuppan; 2002.

BIBLIOGRAPHY OF AUTHORS

	<p>Kyoko Yanagihori</p> <p>She received her Master of Intellectual Property from Tokyo University of Science Graduate School of Innovation Studies in 2011. She is currently studying the application of natural language processing to the patent system. She is a member of the Information Processing Society of Japan and Japan Society of Directories.</p>
	<p>Kazuhiko Tsuda</p> <p>He received the B.S. and Ph.D. degrees in Engineering from Tokushima University in 1986 and 1994, respectively. He was with Mitsubishi Electric Corporation during 1986-1990, and with Sumitomo Metal Industries Ltd. during 1991-1998. He is presently with the Graduate School of Business Science, University of Tsukuba, Tokyo, Japan as an associate professor during 1998-2005, and as a professor since 2005. His research interests include natural language processing, database, information retrieval and human-computer interaction. He is a member of The Information Processing Society of Japan and The Institute of Electronics, Information and Communication Engineers.</p>