Clustering Documents Using Structural Similarity Based on Case Sets –Applied for Technological Problems from Patents

Conference Paper · August 2016

CITATIONS
0

READS
10

2 authors, including:

Hitomi Yanaka
The University of Tokyo
2 PUBLICATIONS **0** CITATIONS

SEE PROFILE

Clustering Documents Using Structural Similarity Based on Case Sets -Applied for Technological Problems from Patents-

Hitomi Yanaka and **Yukio Ohsawa**[1]

Abstract. The description of technological problems in patent documents is important to understand the motivation of the invented technology. Understanding the motivation helps us to analyze trends of the technologies contained in a set of patent documents. Here, we approach the classification of documents based on analogy of structures of the problem descriptions. The purpose of this study is to develop a method for patent classification, with the use of hierarchical clustering based on the structural similarity of problems to be solved by the patented invention. First, we present an approach for extracting predicate-argument structures in the contents of patents. Second, we propose the similarity function to measure the structural similarity between the case sets. The result of the questionnaire survey showed that the structural similarity between patent documents can be calculated with the use of the predicate-argument structures. Furthermore, the survey indicated that comprehension of document structures can be increased by reading the documents reconstructed by the predicate-argument structures.

1 INTRODUCTION

Patent documents provide relevant knowledge of previous inventions and their motivations. New inventors can take the advantage of those previous inventions and learn possible trends of technological problems for future inventions. Due to the variety and large amount of documents of previous inventions, the classification of those helps the search of a relevant document for a new context and problem. Currently, existing standard codes of classification systems are used to support patent document search. Patent examiners can search similar patent documents written in different technical words by using classification codes. There are different kinds of classification systems. For example, IPC (International Patent Classification) is a classification system to grant the classification code in the patent document in accordance with the hierarchy of technical content. Some classification systems are proposed in each region. For example, FI(File Index) is used in Japan, ECLA(European Classification) is used in Europe, and USPC(U.S. Patent Classification) is used in the U.S. Patent examiners use these different classification systems depending on their purposes. Furthermore, these classification codes have been updated manually by experts. For improving the robustness and efficiency of the patent classifications, a method to classify patents automatically based on the technological problems has been demanded.

[1] Department of Systems Innovation, Faculty of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan, email: h2.yanaka@gmail.com, ohsawa@sys.t.u-tokyo.ac.jp

A patent map is also a method of patent classification. Here patent information is collected for a specific purpose of use and depicted in a visual form of presentation such as a chart, matrix, graph, or table. Fig.1 shows an example of patent matrix map of fuel cell. Like this, the trend of target technology fields can be visualized by the patent map. However, details of patent documents contained in each bubble cannot be identified at a glance. In addition, target technology fields have been also specified manually by experts.

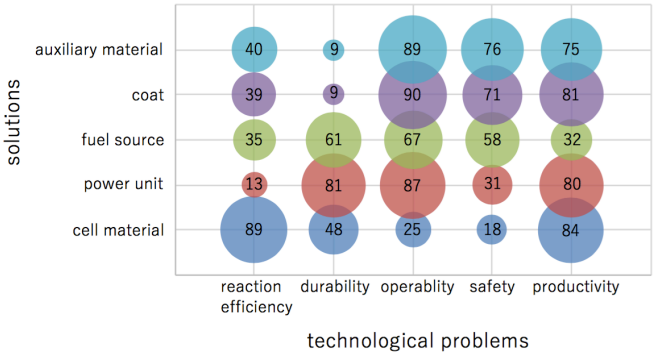

Figure 1. Example of matrix map of fuel cell: the size of pie chart shows the number of patents (in accordance with this technology and the effectiveness of patent). The figure in a pie shows the number of registered documents.

In patent analysis with these methods, analogy is expected to be useful for solving problems in technology development. The analogy is the process toward understanding the problem and solving it from the relation between the base (sometimes called the source) and the target. In patent analysis, the base is the patent data that is already familiar with, whereas the target is the technological problem that we want to understand and solve. In analogy, the correspondence between the target and the base, if noticed, triggers the solution of the target.

According to the structure mapping theory [6][3], there are two kinds of similarities: superficial similarity and structural similarity. The superficial similarity is characterized by elements contained in the target and the base. For example, let us take the following two sentences:

Sentence 1: Water is cold.

Sentence 2: Water is liquid.

Then, both sentences have the same word “water” and the superficial similarity is completed. Relation of the sentences is written below.

$cold\,(water) \rightarrow liquid\,(water)$

The structural similarity is characterized by the primary or higher-order relationships between elements included in them.

Sentence 3: A planet revolves around the sun.

Sentence 4: An electron revolves around the atom.

Then, both sentences have the same structure, "A revolves around B" and structural similarity is completed. Relation of the sentences is written below.

$revolve - around\,(planet, thesun)$

$\rightarrow revolve - around\,(electron, atom)$

MAC / FAC (Many Are Called, but Few Are Chosen) model has been proposed as a model to search common elements between the base and the target based on the superficial similarity, and to assess the validity of the reasoning by evaluating the structural similarity between the base and target [5]. If two patents of different fields of technology are similar in problem structures, they are different in the superficial similarity but similar in the structural similarity. In MAC / FAC model, the structural similarity cannot be found at first and it is difficult to recognize the relations between different fields of technology. According to the previous study about chance discovery[10], a hidden relation provides new knowledge. Therefore, if we can find a relation between patents based on the structural similarity, we can regard a chance discovery inference in which creative technology strategy can be generated. In addition, clustering and classifying patents based on the structural similarity help to find the relations between them. Therefore, we build the hypothesis that it is possible to produce a creative strategy of technology to see the visualization of patents classified by the structural similarity of technological problems.

In this study, we proposed a method for clustering and classifying patents as a method to support the creation of technical development strategy, based on the structural similarity of texts expressing technological problems to be solved by invention. In addition, we proposed a method of visualization of patents to be able to grasp the contents of the technological problems of each patent in a cluster at a glance.

2 RELEVANT STUDY

2.1 Knowledge Representation by Predicate-argument Structure

In English-speaking countries, some researchers suggested an analogy method, which converts a sentence to a logical form by syntactic analysis, in the process of question answering and recognition of implicational relations [9][12]. The previous study[9] showed that predicate logic formation helps to pinpoint exact answers for questions and justify answers on a state-of-the-art question answering system. Furthermore, the other previous study[12] showed that predicate logic formation indicates robust inference and improve to recognize textual inferences by machine learning.

Inspired by such previous work, we proposed a method to support analogy for humans with the use of predicate-argument structures. Predicate-argument structures represent what arguments are related to a predicate, and forms a basic unit for expressing the meaning of a sentence. We express sentences of technological problems by the combination of important predicate-argument structures. In this study, we focus on a first-order predicate logic formation to understand technological problems easily. As a simple and human-friendly format, the predicate-argument structures are composed of verbs, nouns, and the cases of nouns. By looking at technological problems written by the predicate-argument structures and comparing the structures with each other, patent examiners can find the structural similarity between the problems. Therefore, we have the hypothesis that with the use of predicate-argument structures, the examiners can easily use analogy and understand how each patent document approaches technological problems.

2.2 Measurement of Structural Similarity between Documents

The previous study [2] evaluated the method to calculate the structural similarity between sentences consisting of two words by conversion to tuples from an entity-relation graph. It considers structures of documents as an entity-relation graph, in which vertices correspond to entities and edges correspond to lexical-syntactic patterns that represent semantic relations between entities. Then, the previous study proposed numerous kernel functions to measure the degree of analogy between two tuples. This method does not assume a particular relation representation. However, a dependency relation between words can help to increase the accuracy of measuring the structural similarity between documents. Therefore, we use a dependency relation between the predicates and the variables of predicate-argument structures in documents to measure the structural similarity between the documents.

To identify relationships between predicate-argument structures, we focus on the case grammar theory. The case grammar theory is a theory that deals with the essential predicate-argument structure. It describes the logical form of a sentence in terms of a predicate and a series of case-labeled arguments such as agent, object, and location[4]. In Japanese, the case is represented by case particles, such as "ga" (which means subjective case), "wo" (which means accusative case), and "ni" (which means objective case). In previous study, wide case frames are automatically constructed from the web corpus.[7] The concept of case frames is based on the hypothesis that a couple of a verb and its closest case is explicitly expressed on the surface of text, and can be considered to play an important role in sentence meanings. Therefore, the couples of verbs and their closest cases are aggregated for each usage of the verbs, and basic case frames are generated. Then, the basic case frames are clustered to merge similar case frames in a thesaurus and wide case frames are generated.

In this study, based on the concept of case frames, we treat the structural similarity of documents as the similarity of the combination of cases in the documents. In order to measure the structural similarity, we calculate a distance between documents as a distance between the case sets in the documents. In the case set, we regard a combination of a case particle and a noun which has been used in predicate-argument structure in documents as one element. The next chapter describes the details.

3 METHOD

In this study, we used unexamined Japanese patent applications, which were published from 2013 to 2015 and contain the word "condiment" in the content of "problems to be solved by the invention" as datasets to make clusters and visualize. The number of the documents was 185. The proposed method consists of four steps shown in Fig.2. Below let us describe details of each step.

3.1 Summarization of Technological Problem

We extract the technological problem from the content of "problems to be solved by the invention." We approach this problem by summarizing the content. The summary of the content is constructed by

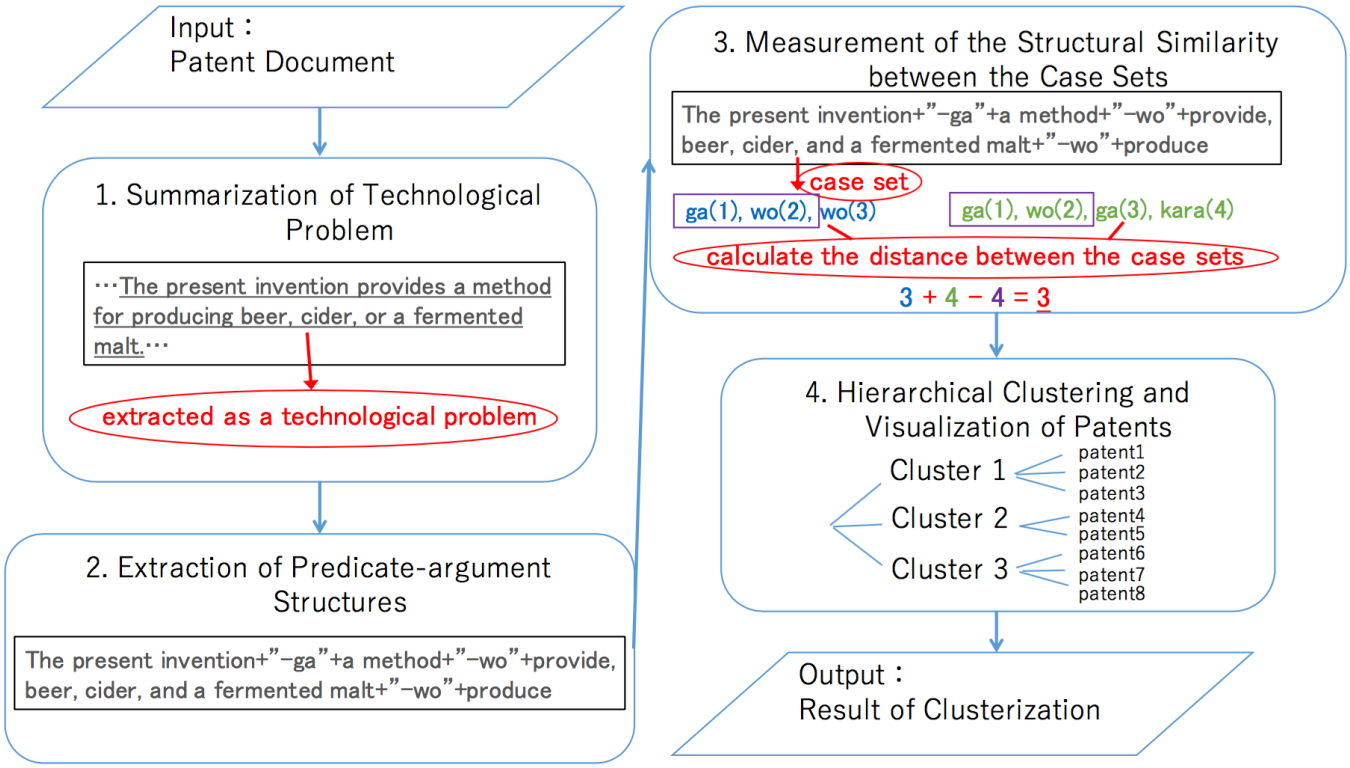

Figure 2. Flow chart of the proposed method

selecting the most important sentence from the content. In the previous study[11], the scoring method by the amount of information in the sentence is proposed to select the most important sentence. This method is based on the hypothesis that an earlier appearance of a word is more informative and scoring a sentence with a combination of appearance position and the frequency of words. The method to calculate the score of a sentence s, is described as Eq. (1).

$$score(s) = \frac{1}{||s||} \sum_i log freq(w_i) pos(w_i) \quad (1)$$

In Bag-of-Words model, the probability of a word w in a sentence s can be measured by its frequency as $freq(w)/||s||$, where $freq(w)$ indicates the frequency of w in s and $||s||$ indicates the total number of words in s. w indicates the i th word in s. $pos(w)$, which is the weight of the positions of w in a sentence s, is defined by the position of the appearances of word w in s. The word score $pos(w)$ is calculated by a geometric sequence, on the assumption that the score of every appearance of a word is the sum of the scores of all the following appearances of it. This sequence is defined as $f(w, s, i)$ in Eq.(2) for the i th appearance of word w in s.

$$f(w, s, i) = f(w, s, i+1) + f(w, s, i+2) + \cdots + f(w, s, n) \quad (2)$$

The content of "problems to be solved by the invention" represents the technological problem, starting from what previous methods could not solve, followed by details, concluding with the purpose of the patent. Important words tend to appear at the beginning and repeatedly in the content. Therefore, the scoring method is adequate to summarize the content. The sentence of the maximum $score(s)$ in Eq. (1) is here selected as the summary. The length of the summary is fixed as around 100 words to recognize at a glance.

3.2 Extraction of Predicate-argument Structures

The summary of a technological problem is expressed as the combination of predicate-argument structures. We use dependency parsing of the summary and extract predicate-argument structures. As a framework of the technological problem, predicate-argument structures are composed of nouns, case particles, and verbs. For morphological analysis, we use Juman[8] and for dependency parsing, we use KNP[7]. Both of Juman and KNP are appropriate for Japanese language. If consecutive nouns including prefix or suffix are contained in the sentence, we regard them as a whole word. Numbers, pronouns, and syncategorematic words are excluded from the extent of research object. We extract categorematic verbs as predicates. When the verbal auxiliary, "nai" (means adding negative) followed after the verb, we mention it behind the verb. We analyze the result of dependency parsing then define each verb written in root as a predicate, and relevant nouns as values of variables of the predicate.If a certain technological problem contains two predicate-argument structure A, B and both of them contains two variables, the technological problem is represented by this format below. Each word is connected by the symbol " + " and each predicate-argument structure is connected by the symbol " , ".

$$noun_{1A} + case_{1A} + noun_{2A} + case_{2A} + verb_A,$$
$$noun_{1B} + case_{1B} + noun_{2B} + case_{2B} + verb_B$$

3.3 Measurement of Structural Similarity between Case Sets

More than one predicate-argument structure must be included in a document. As the predicate-argument structure is the framework of the document, the distance between the documents assumes to be the

same as a total of the distances between their predicate-argument structures. Therefore, we calculate a distance between two patent documents as the distance between the combinations of cases in the documents. By this calculation, the distance between two documents can be determined according to their predicate-argument structures.

The case set is defined as a combination of case particles and variables in a sentence. If a certain sentence i contains n variables, the case set of the sentence set_i is represented as Eq.(3). Here, a variable in the predicate-argument structure val_n is digitized by an appearance order in the sentence. When the same noun appears more than once, the first appearance order is adopted. $case_n$ is defined as the type of case particles related to val_n in the predicate-argument structure. In Japanese, there are 9 case particles: *wo, ni, ga, de, no, yori, he, ya, kara, to*.

$$set_i = \{case_1(val_1), case_2(val_2), \cdots, case_n(val_n)\} \quad (3)$$

The distance of two documents i, j is calculated by the number of elements of the difference set of the case sets as follows:

$$dist_{i,j} = set_i \setminus set_j \quad (4)$$

Let us consider the distance of two Japanese sentences represented by the format of predicate-argument structures below.

1. A+“ga”+B+“ni”+C, A+“ga”+D+“wo”+E+“ni”+F, D+“ga”+G+“ni”+H+“wo”+I
2. J+“ga”+K+“ni”+L, J+“ga”+M+“kara”+N+“he”+O

In the first sentence, there are 3 predicate-argument structures and 8 variables. The first predicate-argument structure and the second predicate-argument structure contain the same variable “A”. The second predicate-argument structure and the third predicate-argument structure contain the same variable “D”. The case set of the first sentence is written as Eq.(5).

$$\{\text{“}ga''(1), \text{“}ni''(2), \text{“}ga''(1), \text{“}wo''(3), \text{“}ni''(4), \text{“}ga''(4), \text{“}ni''(5), \text{“}wo''(6)\} \quad (5)$$

In the second sentence, there are 2 predicate-argument structures and 5 variables. The first predicate-argument structure and the second predicate-argument structure contain the same valuable “J”. The case set of the second sentence is written as Eq.(6).

$$\{\text{“}ga''(1), \text{“}ni''(2), \text{“}ga''(1), \text{“}kara''(3), \text{“}he''(4)\} \quad (6)$$

Therefore, the difference set of these two case sets is calculated as Eq.(7) and the structural distance of the two sentences is calculated as 7.

$$\{\text{“}wo''(3), \text{“}ni''(4), \text{“}ga''(4), \text{“}ni''(5), \text{“}wo''(6), \text{“}kara''(3), \text{“}he''(4)\} \quad (7)$$

3.4 Hierarchical Clustering and Visualization of Patents

Hierarchical clustering by using average linkage is selected as a method of cluster analysis of technological problems of patents. The reason why hierarchical clustering is selected is that the relationship between clusters is visualized by it. Then, we propose a method of visualization of technological problems. The format of output data is JSON, which is a lightweight data-interchange format. The embedded structure of the data is composed of the number of the cluster, the technological problem solved by each patent document, and the patent number in sequence. In visualization, we use D3.js, which is a JavaScript library for visualizing data with JSON, HTML, and CSS.

4 QUESTIONNAIRE SURVEY

4.1 Questionnaire Hypothesis

We carried out a questionnaire survey to evaluate the validity of clustering and usefulness of visualization. We have three hypotheses as follows:

1. The distance between the case sets in documents reflects the structural distance between documents.
   In this study, we regard the structural distance between documents as the distance between the case sets in documents. To verify the validity of this hypothesis, we should confirm that the cluster is clustered based on the structural similarity.
2. Representing technological problems as predicate-argument structures helps to understand technological problems easily and accurately.
   Shown in the previous study[9], predicate-argument structures can help to find exact answers for questions. Therefore, participants can understand the structures of technological problems more easily by reading the documents formed with predicate-argument structures than by reading the plain texts. Also, the awareness of the causal structures of technological problems can be increased by the formation of predicate-argument structures.
3. Representing technological problems as predicate-argument structures helps to support analogy and to find common problems in different fields of technology.
   In relation to the second hypothesis, participants can find common problems in different fields of technology more easily by reading the documents formed with predicate-argument structures.

4.2 Questionnaire Details

To verify these hypotheses above, a questionnaire survey conducted to 18 participants in twenties in June, 2016. All relevant consent and human subject approvals were obtained for this experiment. We selected two clusters from the result of the cluster analysis and made a questionnaire form with each cluster. The questionnaire 1 was made with the cluster which contains three documents and the questionnaire 2 was made with the cluster which contains five documents. Half of all participants were given the questionnaire 1 and the other half were given the questionnaire 2. The reason why we selected different size of the clusters is that we have a hypothesis that it is more difficult to capture the structural similarity of documents if a number of documents increases and then people prefer to refer predicate-argument structures to answer the analogy question.

We presented the participants both the original documents and the result of the cluster analysis. Then, we asked 3 questions for the group of original documents and the result of the cluster analysis. The 3 questions are shown below.

Part 1. Comparing the two documents, as seen from the structure of each document, choose a word that corresponds to the word in one document from the words in the other. An example question is shown below.

Question: Comparing No.2015104322 with No.2015112075 in Fig.3, please choose a word that corresponds to the ”taste ” of No.2015104322 from the words in No.2015112075.

Choices: flavor, liquid condiment, denseness, smell, flavor, vegetable, rich, problem

The correct answer: flavor

This part contains two multiple choice questions. The correct answer is determined by the discussion among the author and two patent

experts engaged in the intellectual property department of the food company for more than three years.

Part 2. Please answer the technological problems which may be common in the cluster. (example answer: All of the sentences refers "a method of manufacturing a composition about the food" as a common issue.) If you cannot find common technical problems, please answer "nothing."
This part is free description type of the question. To make it easier to answer this part, an example answer is described in it.

Part 3. Which type of question is easier to answer, plain texts or predicate-argument structures?

The part 2 is a question to verify the hypothesis 1 and hypothesis 3. The part 1 and 3 are questions to verify the hypothesis 2.

5 RESULT AND DISCUSSION

5.1 Result of the Clusterization

Fig.3 shows the overall and one cluster of the result of the cluster analysis. The number of clusters was determined as 30. The number of elements in each cluster was in the range of 1 - 136. The number was inclined to one element and the number of clusters which contain only one element was 21. However, this tendency can be seen because clustering is based on the structural similarity of technological problems and many documents with different structures are included in datasets. Therefore, as mentioned at 4.1, the validity of the result is evaluated by the results of the questionnaire.

5.2 Result of the Questionnaire Survey

Table.1 shows the percentage of correct answers between plain texts and predicate-argument structures at Part 1. The percentage was calculated by the average of two questions.

Table 1. Percentage of correct answers between plain texts and predicate-argument structures

	PT	PS	Chi-square test
Questionnaire1	27.8	38.9	$p < 0.05$
Questionnaire2	50.0	77.8	$p < 0.05$

PT:plain texts, PS:predicate-argument structures

Shown in Table.1, the percentage of correct answers was increased by reading the documents formed with predicate-argument structures. In addition, the difference in the percentage of correct answers was significant in the chi-square test. (significant level $p < 0.05$) This indicates that the use of predicate-argument structures is effective for humans to understand the structure of documents.

Table.2 and Table.3 show common technological problems answered at Part 2 in Questionnaire 1 and Questionnaire 2.

Table 2. Common technological problems answered at Questionnaire 1, Part 2

	Common problem	Number
PT	improvement of taste	5
	property change of material	2
	production of condiment	1
	analysis of condiment	1
PS	improvement of taste	8
	property change of material	1

PT:plain texts, PS:predicate-argument structures

Table 3. Common technological problems answered at Questionnaire 2, Part 2

	Common problem	Number
PT	improvement of taste	3
	preservation method	2
	prevent deterioration of liquid	2
	solving problems of food additives	1
	production of condiment	1
PS	improvement of taste	6
	solving problems of food additives	2
	production of condiment	1

PT:plain texts, PS:predicate-argument structures

The technological problem of each patent in the clusters used in the questionnaires is an improvement in the taste of different condiments with different materials. Hence, the improvement of the taste can be considered as a common structure of technological problems. Shown in Table 2 and Table 3, the improvement of the taste was answered as a common technological problem in both of the questionnaires. This result suggests that the distance between the case sets in documents reflect the structural distance between documents. In addition, comparing plain texts with predicate-argument structures, the answers were less varied in predicate-argument structures. This indicates that the common structure of technological problems should be more easily grasped from documents formed by predicate-argument structures. Furthermore, comparing the two questionnaires, the answers were more varied in the questionnaire 2. This implies that the common structure of technological problems should be more easily grasped if the number of sentences for comparison is smaller.

Table.4 shows the percentage of selected answers at Part 3.

Table 4. Percentage of selected answers

	PT	PS	Chi-square test
Questionnaire1	55.6	44.4	non-significant
Questionnaire2	55.6	44.4	non-significant

PT:plain texts, PS:predicate-argument structures

Shown in Table.4, the readability was no difference between predicate-argument structures and plain texts. Furthermore, the preference of predicate-argument structure was no difference by the degree of the difficulty of the analogical problem. There are two reasons for this. First, some participants pointed out they didn't understand the meaning of the symbol " + " and " , ". It indicates that some participants couldn't grasp the structure of documents from predicate-argument structures. Because the questionnaire did not contain the explanation about how to form predicate-argument structures, the explanation should be added to it. Second, some participants said conjunctions in plain texts helped to understand the structure of documents. Therefore, the relations between predicate-argument structures should be represented in them.

6 CONCLUSION

This study proposes a novel method for text clustering. Compared with the previous method of frequent term based text clustering[1], extracting predicate-argument structures from patent documents and calculating the distance between patent documents based on case sets in the predicate-argument structures is a new approach. In the questionnaire survey, some participants found the same common problems from the cluster. This indicates that the proposed method can

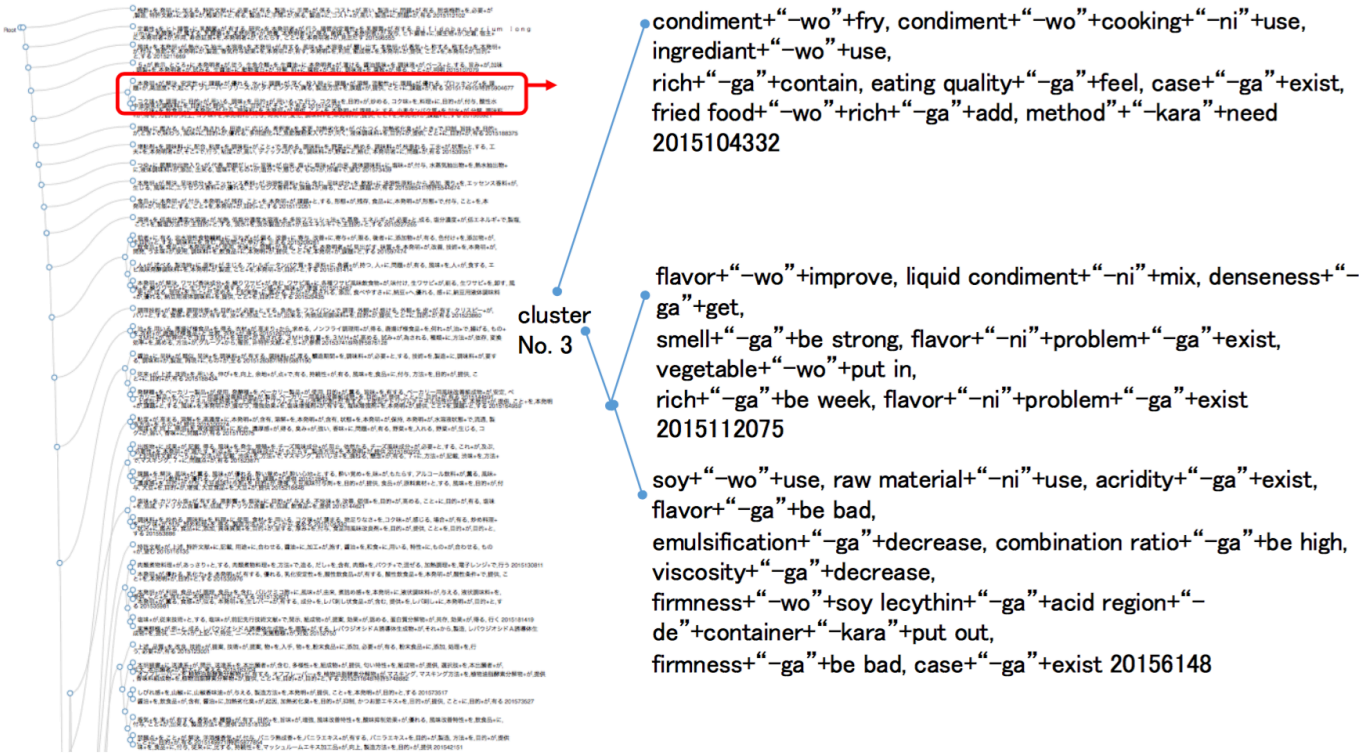

Figure 3. Overall and one cluster of the result of the cluster analysis

be a method of document clustering based on the structural similarity. However, relations between clauses are not reflected in the structural similarity of documents because this proposed method cannot consider relations between clauses such as subordinate clause and parallel clause. Therefore, an issue in the future is to design the distance between the documents, considering both relations between predicate-argument structures and relations between clauses.

Furthermore, the questionnaire survey also indicates that the percentage of correct answers can be increased by reading the documents formed by predicate-argument structures. Therefore, this visualization method is thought to be effective to support to search some patent documents relating to inventors' own technological problems. However, the survey also indicates that the readability is no difference between predicate-argument structures and plain texts. It is due to the lack of explanation of the questionnaire and there is a need for improvement of the questionnaire survey. Also, it is necessary to improve the format of predicate-argument structure easily to understand the structure for humans.

ACKNOWLEDGEMENTS

This work was supported by CREST, Japan Science and Technology Agency. This paper has been accomplished under collaborative research project with Toppan Forms, Tokyo, Japan.

REFERENCES

[1] F. Beil and M. Ester, 'Frequent term-based text clustering', *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.436–442, (2002).

[2] D. Bollegala, M. Kusumoto, Y. Yoshida, and Kawarabayashi K., 'Mining for analogous tuples from an entity-relation graph', *Proceedings of the 23rd international joint conference on Artificial Intelligence*, pp.2064–2077, (2013).

[3] B. Falkenheiner, K. Forbus, and D. Gentner, 'The structure mapping engine: Algorithm and examples', *Artificial Intelligence*, **41**, pp.1–63, (1989).

[4] C. Fillmore, *The Case of Case*, Universals in Linguistic Theory, New York, 1968.

[5] K. Forbus, D. Gentner, and K. Law, 'Mac/fac: A model of similarity-based retrieval', *Cognitive Science*, **19**, pp.141–205, (1994).

[6] D. Gentner, 'Structure-mapping: A theoretical framework for analogy', *Cognitive Science*, **7**, pp.155–170, (1983).

[7] D. Kawahara and S. Kurohashi, 'A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis', *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp.176–183, (2006).

[8] S. Kurohashi and M. Nagao, 'Japanese morphological analysis system jumanversion 3.61', *Proceedings of the 20th National Conference on American Association for Artificial Intelligence*, (1999).

[9] D. Moldovan, C. Christine, Sanda H., and Steve M., 'Cogex: A logic prover for question answering', *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.87–93, (2003).

[10] Y. Ohsawa and P. McBurney, *Chance Discovery*, Advanced Information Processing, Berlin: Springer, 2003.

[11] Y. Ouyang, W. Li, Q. Lu, and R. Zhang, 'A study on position information in document summarization', *Proceedings of the 23rd International Conference on Conputational Linguistics*, pp.919–927, (2010).

[12] R. Raina, A. Ng, and C. Manning, 'Robust textual inference via learning and abductive reasoning', *Proceedings of the 20th National Conference on American Association for Artificial Intelligence*, pp.1099–1105, (2005).