Clustering Documents on Case Vectors Represented by Predicate-Argument Structures – Applied for Eliciting Technological Problems from Patents

Hitomi Yanaka, Yukio Ohsawa
Department of Systems Innovation, Faculty of Engineering,
University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
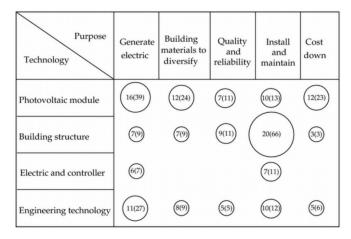Email: h2.yanaka@gmail.com, ohsawa@sys.t.u-tokyo.ac.jp

Abstract—Patent analysis is useful to understand the trends of technological problems and develop strategies for technologies. Here patent classification is a method to support the analysis. The purpose of this study is to propose a method for patent classification, with the use of hierarchical clustering based on the structural similarity of problems to be solved. The structural similarity can be calculated with case vectors based on predicate-argument structures of the contents of the patents. The interview survey indicated that this classification plays an essential role in analogical problem solving, by allowing visualization of similar technological problems.
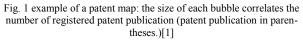
I. INTRODUCTION

IN COMPANIES, it is important to understand the trends of technological problems by analyzing patents. By this, they learn to improve strategies for the development of technologies. Classification symbols have been used for classifying patent documents in patent analysis. Patent examiners use them to search similar patent documents written in different technical words or languages. The classification symbols are updated manually by experts. For improving the efficiency, a method to classify patents automatically based on the semantic content is required.

A patent map, a model of patent visualization, 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 a patent matrix map [1]. As can be seen from Fig. 1, a patent map is mainly produced to grasp its technology trends by gathering related patent information of target technology fields. However, it cannot be identified at a glance which patent documents are contained in each cluster. Also, the target technology fields have been specified manually by experts.

Analogy is the process toward understanding and solving the problem from the relation between the knowledge base (sometimes called the source) and the target problem. Finding the relationship between the base and the target can

Fig. 1 example of a patent map: the size of each bubble correlates the number of registered patent publication (patent publication in parentheses.)[1]

be useful for patent analysis. According to the structure mapping theory [2], there are two kinds of similarities: superficial similarity and structural similarity. Superficial similarity is characterized by elements contained in the target concept and the base concept. Structural similarity is characterized by the primary or higher-order relationships between elements included in them. MAC / FAC 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 [3].

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

II. RELATED WORK

A. Extraction of Technological Problems by Clustering

A previous study proposed a method of extracting a topic by automatic classification of newspaper articles [4]. This method composes a document vector by using the frequency of each noun in the document as features and extracts the topic by performing hierarchical clustering. However, when the similarity between two documents is calculated by the frequency of each noun in the documents, a document including different words with the same structure can be regarded to be dissimilar. Therefore, using the frequency of each noun yields superficial similarity, not structural similarity. However, there is a possibility that a new topic of technological problems can be found by gathering technological problems with the same structure. Therefore, we considered both the superficial similarity and the structural similarity and represented the feature amount of patent document. This enables us to make clusters based on MAC / FAC Model, a human analogy model.

B. Conversion to a First-order Predicate Logic Formation

For supporting analogy, it is necessary to consider both structural and superficial similarities, not only for construction of sentence vectors but also for visualization. The previous study reports the usefulness of converting sentence to predicate-argument structures for the recognition of implicational relation and the analysis of dialogues [5]. This study applies the conversion result from a conversion system called Boxer [6, 7] to applied tasks of semantic analysis using analogy. However, the conversion system to predicate-argument structures has not been used for supporting idea creation. Furthermore, the conversion system from Japanese sentence has not been confirmed yet. Japanese grammar is different from English grammar and it is necessary to propose a method to convert into logical form particular in Japanese language.

Therefore, we proposed a method to support analogy with the use of predicate-argument structures. We expressed sentences of technological problems by combining important predicate-argument structures. Comparing technological problems written by predicate-argument structures with each other, patent researchers can find the structural similarity between the problems. With the use of predicate-argument structures, it enables the researchers to easily use analogy and to understand how each patent document approaches technological problems.

III. METHOD

The proposed method consists of six steps. Below let us describe details of each step.

Step.1 Summarization of Technological Problem

We extract important sentences about technological problems from patent documents by summarizing the content filled in the blank "problems to be solved by the invention." Here we use a basic model for summarizing text

by selecting the important sentences [8]. The important sentences are selected based on the scores until the length limit is reached. The length of the summary is fixed as around 100 words to recognize at a glance. The score of each sentence, $score(s)$, is calculated by the amount of information included in the sentence s as shown in Eq. (1).

$$score(s) = \sum_i log\, freq(w_i, s) pos(w_i, s) / \|s\| \quad (1)$$

In a Bag-of-Words model, the probability of a word w in a sentence s can be measured by its frequency as $freq(w, s)/\|s\|$, where $freq(w, s)$ indicates the frequency of w in s and $\|s\|$ indicates the total number of words in s. w_i indicates the i th word in s. $pos(w, s)$ is the weight of the position of w in s. We take a geometric sequence as a calculate function of $pos(w, s)$ based on the position hypothesis that earlier appearances of a word are more informative. A geometric sequence is defined as $f(w, s, i)$ in Eq.(2), based on the assumption that the degree of every appearance of a word is the sum of the degree of all the following appearances of it.

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

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 invention, i.e. the target technological problems. Important words of technological problems tend to appear at the beginning and repeatedly appear in the content. Therefore, the scoring method is adequate to summarize the content.

Step.2 Extraction of Predicate-argument Structures

Next, 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 simple format of the technological problem, predicate-argument structures are composed of nouns, case-marking particles, and verbs. MeCab[9] is used for morphological analysis and CaboCha[10] is used for dependency parsing. Both of them are appropriate for Japanese language.

In this research, the research object is confined to predicate-argument structures with adnominal case particles "-ga" (means subjective case), "-wo" (means accusative case), and "-ni" (means objective case). These three factors construct the framework of a sentence, especially if the sentences expressing an aimed function of a subject interacting with things to be given as objects. That is, nouns are regarded as indications of subjects and objects in the sentence. If consecutive nouns including prefix or suffix are contained in the sentence, they are regarded as a whole word. Numbers, pronouns, and syncategorematic words are excluded from the extent of research object.

Categorematic verbs are extracted as predicates. When the verbal auxiliary, "-nai" (means adding negative) followed behind the verb, we mention it behind the verb. The noun related to the nominal verb, "suru" (means "do") are extracted together as one verb.

From the result of dependency parsing, each verb written in original form is defined as a predicate, and relevant nouns are defined as values of its variables. The verbs, which appear in over 90% of the datasets, are meaningless verbs and are excluded from technical problems.

Step.3 Representation of Case Vector

To consider both the superficial similarity and the structural similarity, we also focus on a predicate-argument structure, which consists of nouns, verbs, and their relations. The predicate-argument structure represents case form, which is defined by a form of generative grammar such as subjective and objective, based on the semantic relationship of noun phrases to verbs. As one sentence contains at least one predicate-argument structure, the structure of the sentence can be represented by case vectors.

In Japanese language, a semantic relationship between a verb and a noun is confirmed by a case-marking particle. Furthermore, a verb tends to have a relationship with certain nouns to express its meaning. For instance, the verb "boil" tends to have a relationship with the nouns such as "human being" as the subject and has the relationship with the nouns such as "meals" as the object. Therefore, we represent a verb vector based on the verb's semantic role by using frequency of the nouns, which have a relationship with the verb such as subjects or objects.

Additionally, a case relationship represents the influence of a verb to/from a related noun. That is, the case relationship is useful for a semantic vector representation of a verb. We represent a verb vector by using the appearance of the nouns, which are in the same case relationship. The same case relationship means that they have the same case-marking particle. We define a case vector as a verb vector of each case-marking particle. For example, verb "Iku" (meaning "goes") in sentence "Prime Minister Abe ga Russia ni Iku." in Japanese, meaning "Prime Minister Abe goes to Russia," is put into:

Iku: (subject: Prime Minister Abe, Object: Russia).

This will be then put into a case vector where numerical values are filled as elements corresponding to "subject: Prime Minister Abe", "Object: Russia".

The case vector is constructed by a Bag-of-Words model. This model is the simplest vector representation for a sentence or a document. Each dimension of the case vector is correlated with a noun. Let us take a subjective case vector as an example. The value of the dimension is calculated as 1 if there is a subjective noun for a verb in the document and otherwise calculated as 0. Furthermore, the vector for the *i*-th verb is weighted by the *j*-th noun, which is related to the verb in the sentence. When the verb *i* has a predicate-argument structure with *m* nouns in the corpus, the weight of the *j*-th noun is calculated in Eq. (3).

$$w_i = (w_{i1}, w_{i2}, \dots, 2\,w_{ij}, \dots, w_{im}) \qquad (3)$$

In this study, to identify patterns in the relationships between nouns and concepts of verbs, a method of Latent Semantic Indexing (LSI) is applied to nouns/verbs matrix. LSI is an indexing and retrieval method to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text [11]. Additionally, the computational complexity decreases and the accuracy of a clustering algorithm can be improved with the use of LSI. This study is different from the previous study in applying LSI to nouns/verbs matrix.

If *n* is the number of verbs and *m* is the number of unique nouns in the corpus, *A* is the matrix about verbs V (V=1, …, *n*) which are related to nouns N (N= 1 , …, *m*) in the sentence *s* as shown in Eq. (4).

$$A = \begin{matrix} & V_1\ V_2\ V_n \\ \begin{matrix} N_1 \\ N_2 \\ N_m \end{matrix} & \begin{pmatrix} 0 & 2 & 0 \\ 2 & 0 & 0 \\ 0 & 0 & 4 \end{pmatrix} \end{matrix} \qquad (4)$$

This matrix *A* is decomposed by Singular Value Decomposition (SVD). When the rank of *A* is *r*, *T* is an *m* by *r* noun-concept vector matrix. *S* is an *r* by *r* singular values matrix. *D* is an *n* by *r* concept-verb vector matrix. Then, the matrix *A* is compressed to *k* dimensional space by restructuring the matrix *A'* using top *k* amounts of large and primary factor shown in Eq. (5). Dimension *k* is defined from preliminary experiment (In this study, k=200).

$$A' = T_k S_k D_k{}^T \qquad (5)$$

Step.4 Calculate Distance Between Patents

Next, hierarchical clustering by using average linkage is selected as a method of cluster analysis of technological problems of patents. The relationship of each cluster is visualized by hierarchical clustering. As the predicate-argument structure is the framework of the document, the distance between the documents can be treated as a total of the distance between the predicate-argument structures in the document. Therefore, we calculate a distance between two patent documents as the sum of distances between the case vectors in the documents. By this calculation, the distance between two documents can be determined according to their predicate-argument structures. For simplicity, in this study, we defined three kinds of case vectors: subjective case vector, accusative case vector, and objective case vector. Nouns, which prefer to relate to verbs are different on the type of case and the distance between case vectors should be calculated in the same type of case each other. Therefore, we calculate a distance of the same type of case vectors and add them together as shown in Eq. (6).

$$sdist = sdist_{sub} + sdist_{acc} + sdist_{obj} \quad (6)$$

$sdist$ means the distance between two documents, $sdist_{sub}$ means the distance between these documents in the subjective case.

The distance $cdist \ (V_1, V_2)$ between two case vectors, V_1 and V_2 is calculated by Euclidean distance divided by the

$$cdist \ (V_1, V_2) \ = \frac{1}{cdist_{MAX}} \sqrt{\sum_{k=1}^{n} (V_{1k} - V_{2k})^2} \quad (7)$$

maximum of all distances, as shown in Eq. (7).

V_{1k} and V_{2k} are values of V_1 and V_2 at a dimension k. Each distance takes a value within the range of $0 \leqq cdist \leqq 1$ so that the distance of each type of case should be treated as equivalent.

When the document p contains k subjective nouns and document q contains l subjective nouns, the distance $sdist_{sub} \ (p, q)$ between these documents in the subjective case is defined as Eq. (8).

$$sdist_{sub} \ (p, q) \ = \frac{1}{kl} \sum_{i \in k} \sum_{j \in l} cdist(V_i, V_j) \quad (8)$$

As shown in Eq. (8), the sum of distances of the subjective cases is divided by the number of the cases to consider the average of the distances as the distance of the case.

If one document contains one direct objective noun and the other does not, $sdist_{acc}$ cannot be calculated in Eq. (8). In this case, these two documents are completely different in accusative case and $dist_{acc}$ is regarded as the maximum $sdist_{acc}$, 1. Similarly, if two documents contain no indirect objective nouns, $dist_{obj}$ cannot be calculated in Eq. (8). In this case, these two documents are the same in objective case and $dist_{obj}$ is regarded as the minimum $sdist_{obj}$, 0.

To illustrate the above, there are two example sentences of calculating distances.
Example 1: Subject A / Verb B / Direct Object C
Example 2: Subject N / Verb O / Indirect Object P,
Subject Q / Verb R / Direct Object S

There are two pairs of the subjective case (A-N and A-Q) and one pair of the accusative case (C-S). Therefore, $sdist_{sub}(s_1, s_2)$ is expressed to be $\frac{1}{2} \ (cdist(V_A, V_N) + cdist(V_A, V_Q))$ and $sdist_{acc}(s_1, s_2)$ is expressed to be $cdist(V_C, V_S)$. In addition, these two documents are completely different in objective case and $sdist_{obj}(s_1, s_2)$ is 1. In this example, the formula to calculate the distance $sdist(s_1, s_2)$ between Example 1 (s_1) and Example 2 (s_2) is shown in Eq. (9).

$$sdist(s_1, s_2) = \frac{1}{2} \ (cdist(V_A, V_N) + cdist(V_A, V_Q))$$

$$+ cdist(V_C, V_S) + 1 \quad (9)$$

Step.5 Hierarchical Clustering of Patents

The hierarchical clustering method in the previous study[4] has a problem of fluctuating extracted topics because these topics subjected to the number of clusters and the number of clusters is decided in the hand. We here employed Upper Tail method [12, 13], clustering with an automatic decision of the number of clusters. This method is based on the stopping rule of Eq. (10).

$$\alpha_{j+1} > \alpha_{ave} + ks_\alpha \quad (10)$$

When the number of documents is n , the most appropriate number of clusters is j, and the number of clusters is n-j (j=0, 1,..., n-1), α_j is the minimum distance between two documents belonging to different clusters. α_{ave} is an average of a distribution of α_j. s_α is a square root of the unbiased estimate of variance. k is determined by the number of elements per cluster. We determine k=2, considering that the number of documents per cluster is estimated from 10 to 50. The previous study [13] shows normalizing α to be distributed by chi-square as shown in Eq. (11) achieved better accuracy and we perform the same procedures.

$$\alpha' = \Phi^{-1}(F_p(\alpha/s_\alpha \cdot p) \quad (11)$$

Φ^{-1} is the inverse function of the normal distribution. F_p is distribution function of chi-square when a degree of freedom is p (In this study, p=1).

Step.6 Visualization of Similar Technological Problems

In the end, 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, 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 and HTML.

IV. RESULTS AND DISCUSSION

A. Datasets

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 documents was 348. The dictionary data for constructing the case vector was the content of "problems to be solved by the invention" of seventy thousand patent documents relating to "Foods" (theme code: 4H).

First condition

Cluster 9

- the very hot water + "-wo" + use, ingredient + "-wo" + take in, liquid oxidation + "-wo" + move ahead, production dry processed foods + "-ni" + be suitable for + not, 2014217320
- flavor + "-wo" + maintain, spread + "-ni" + liquid seasoning, flavor + "-wo" + add, 2014103925
- production + "-ni" + require, everything + "-wo" + satisfy, 2014183758
- flavor + "-ni" + excel in, 2014103883
- ingredients for cooking + "-ni" + sprinkle, 2013172681
- upper part + "-ni" + attach, preventing device + "-wo" + attache, body of pot + "-ni" + attach, the opening + "-wo" + fill up, 20139772 / Reg. No. 4881413
- acid, separated and liquid seasoning + "-ga" + entwine, condition + "-wo" + maintain, texture + "-ga" + maintain, 201348569
- bouillon feel + "-ga" + dilute, not concentrated version + "-ni" + compare with, bouillon feel + "-wo" + feel, 2015181435

Second condition

Cluster 9

- acid, separated and liquid seasoning + "-ga" + entwine, condition + "-wo" + maintain, texture + "-ga" + maintain, 201348569
- flavor + "-wo" + maintain, liquid seasoning + "-ni" +spread, flavor + "-wo" + add, paste spice + "-wo" + provide, 2014103925
- acid taste + "-ga" + inhibit, richness + "-wo" + maintain, acid seasoning + "-wo" + provide, 2015171328

Third condition

Cluster 18

- yeast extract + "-ni" + become + not, 201479179
- umami + "-ga" + enhance, odor of extract + "-ga" + reduce, extract seasoning + "-wo" + provide, 201589349

Fig.　2 Part of clusterization of patent documents in three conditions

First condition used the distance determined by the case vector, second condition used the distance by the sentence vector with the verbs as feature amount and third condition used the distance by the sentence vector with the nouns as feature amount.

B. Evaluation of Summarization of Technological Problem

324 summaries were created from the datasets in this study. The reason why some summaries could not be created is that the contents of "problems to be solved" of such patent documents were too short to summarize. In order to evaluate the validity of the summarization, randomly selected 100 summaries extracted by the method were compared to summaries determined by discussion among patent experts for the same documents. The rate of concordance in this comparison was 81% and it shows the method is adequate to summarize patent documents.

C. Evaluation of Cluster Analysis and Visualization

To evaluate the utility of the proposed method, we performed cluster analysis of patent documents, using three kinds of distances and compared the results with each other. The first distance was determined by the proposed method, the case vector. The second distance was determined by the sentence vector using frequency of verbs as the feature. The third distance was determined by the sentence vector using frequency of nouns as the feature. We used TF/IDF to weight the features in the second and the third distances. TF means the frequency of words in each sentence and IDF means logarithm of inverse number of the frequency of words used in total sentences of the summary of the technological problem. Fig. 2 shows part of the result of the cluster analysis in three conditions, corresponding to the three kinds of distances above. The number of clusters under

each condition was determined as 44 (the first condition), 20 (the second condition), and 40 (the third condition).

We interviewed two patent experts engaged in the Intellectual Property Department of the food company for more than three years. We showed them the result of the cluster analysis in three conditions and asked two questions below.

Question 1. What kind of difference do you find by comparing the three types of clustering results? What similarities do you find in the technical problems in the same cluster?

Question 2. Can technological problems of patent document be grasped at the moment? Is there any use to develop technology strategy or intellectual property strategy?

The answers to Question 1 are shown below:

- At the first glance, each cluster in the first condition seems to lack in coherence. However, for example, the cluster No.9 seems to be aggregated by the same concept of verbs. The verbs such as "maintain", "take in" and "fill up" represent the concept of keeping and the verbs such as "sprinkle", "attach" and "add" represent the concept of adding.
- Each cluster at the second condition seems to be aggregated by the same verb roughly.
- As the cluster No.18 is aggregated by the same noun "extract" in the third condition, the same goes for others at the third condition.

These answers indicate that case vectors have a possibility to represent the concept of the verbs, which forms the

structural similarity of the sentences. On the other hand, as a result of the second and third condition, the sentence vector using frequency of words as the feature represents the superficial similarity based on the words. The third condition also indicates that the meaning of the sentence is more influenced by the meaning of nouns than that of verbs.

The answers to Question 2 are shown below:

- When technological problems are expressed by the predicate-argument structures, modifiers are removed. Therefore, removing terminology of confidential information with this method is useful for sharing confidential data with others. On the other hand, some technological problems are hard to read. For example, "advantage convenience+-ga+loses" is grammatically incorrect. The correct sentence is "advantage convenience is lost".
- The words which indicate not only the material of condiment but also the method for processing were extracted as technological problems.
- Some technological problems were extracted with accuracy, while others with unnecessary words.

The first answer indicates that the proposed method of visualization is useful in not only patent documents but also confidential documents. The answer also indicates that sentences of technological problems including passive verbs are a little difficult to read. This is because an auxiliary verb that represents the passive voice was not considered in this study and excluded. The method of expressing a sentence with a passive verb as predicate-argument structures will be an issue to be addressed in the future.

The second answer indicates that technological problems extracted by the proposed method are useful to find a point of view in patent analysis, whereas existing patent classification is mainly divided by the material of a condiment.

The third answer suggests that the interviewee could find the structural similarity of some technological problems and the usefulness of this method. However, as the structural similarity is not based on the commonality of words, the interviewee had difficulty to find the structural similarity of others. Therefore, it is necessary to get an evaluation of this method to more interviewees. Another reason for this result is the structural similarity of two documents is not always directly represented by predicate-argument structures including in the documents. Therefore, it is necessary to consider the extraction method of the abstract predicate-argument structure in the documents.

V. Conclusion

In this study, we proposed a novel method of text clustering by using the distance of case vectors derived from predicate-argument structures of patent documents. A method for calculating the distance between documents based on predicate-argument structures has not been approached. This study suggested the possibility to capture not only superficial similarity based on nouns in sentences but also structural similarity based on the meaning of verbs in sentences. Furthermore, the interview survey indicated

that the proposed visualization method of technical problems is useful to overview the problem, which patent researcher should search for. Therefore, this visualization method is thought to be effective to support to search patent documents relating to researcher's own technological problems.

In this research, we calculated the distance between patents as an average distance between case vectors included in the documents. However, as this method cannot consider the causal structures of documents, the similarity of documents could have been inaccurate. Therefore, an issue in the future is to evaluate the method for deriving similarity considering causal structures in documents.

References

[1] Chiu, Y.J., Ying, T., A Novel Method for Technology Forecasting and Developing R&D Strategy of Building Integrated Photovoltaic Technology Industry, Mathematical Problems in Engineering, 2012, 2012, pp.1-24, http://dx.doi.org/10.1155/2012/273530.

[2] Falkenheiner, B., Forbus, K., Gentner, D., The Structure Mapping Engine: Algorithm and Examples, Artificial Intelligence, 41, 1989, pp.1-63, http://dx.doi.org/10.1016/0004-3702(89)90077-5.

[3] Forbus, K., Gentner, D., and Law, K., MAC/FAC: A model of similarity-based retrieval, Cognitive Science, 19, 1994, pp.141- 205, http://dx.doi.org/10.1016/0364-0213(95)90016-0.

[4] Hashimoto, T., Murakami, K., Inui, K., Uchiumi, K., Ishikawa, M., Topic Extraction and Social Problem Detection Based on Document Clustering, SocioTechnica, Vol.5, 2008, pp.216-226, http://dx.doi.org/10.3392/sociotechnica.5.216

[5] Bos, J., Markert, K.. Recognising Textual Entailment with Logical Inference, Proceedings of the 2005 Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vol. 2012-NL-206, 2005, pp. 628–635, http://dx.doi.org/10.3115/1220575.1220654.

[6] Bos, J., Wide-Coverage Semantic Analysis with Boxer, Proceedings of the 2008 Conference on Semantics in Text Processing, 2008, pp. 277–286, http://dx.doi.org/10.3115/1626481.1626503.

[7] Bos, J., Clark, S., Steedman, M., Curran, J., Hockenmaier, J.. Wide-Coverage Semantic Representations from a CCG Parser. Proceedings of the 20th international conference on Computational Linguistics, 2004, pp.1240-1246, http://dx.doi.org/10.3115/1220355.1220535.

[8] Ouyang, Y., Li, W., Lu, Q., Zhang, R., A Study on Position Information in Document Summarization, Proceedings of the 23rd International Conference on Conputational Linguistics, Beijing, China, 23-27 August.2010, pp.919-927.

[9] MeCab: Yet another part-of-speech and morphological analyzer. http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html Accessed: 2016-05-06.

[10] Cabocha: Yet another Japanese dependency structure analyzer. http://taku910.github.io/cabocha/ Accessed: 2016-05-06.

[11] Dumais, S., Latent Semantic Analysis, Annual Review of Information Science and Technology, Vol.38, Issue.1, 2004, pp.188-230, http://dx.doi.org/10.1002/aris.1440380105.

[12] Mojena, R., Hierarchical grouping methods and stopping rules: an evaluation, The Computer Journal, Vol.20, 1977, pp.359-363, http://dx.doi.org/10.1093/comjnl/20.4.359.

[13] Shizu, A., Matsuda, S., Comparison of the Cluster Number Automatic Determination Method in a Cluster Analysis, Academia. Information sciences and engineering : journal of the Nanzan Academic Society, Vol.11, 2011, pp.17-34.