Clustering Web Pages Considering the Position of Each Word and the Search Term

Ryutaro Akiyama, Katsutoshi Kanamori, and Hayato Ohwada

***Abstract*—Users can easily find the pages they are seeking by clustering the web pages of search results obtained using a search engine. The vector space method is often used to cluster web pages. However, the method that has been conventionally used has low clustering accuracy and high computational cost. In this study, we propose a method to solve these problems. We assume that the words that appear near the search term have a high degree of importance. We then try to solve the problems by considering the distance to the search term in the text. We conducted verification experiments in Japan for Japanese search terms. The results confirmed that the proposed method considering the distance to the search term in the text has higher clustering accuracy and lower computational cost than the conventional method.**

***Index Terms*—Information retrieval, search engine, web clustering, web mining.**

I. Introduction

Search engines such as Google and Yahoo! are generally used to search for web pages on the Internet. However, pages with the information sought by the user do not appear in higher rank in the search results [1]. Some words have semantic ambiguity and polysemy, so even if users enter the search term with some intent, pages that differ from the intent may appear in the search results.

To resolve this, clustering web pages of search results obtained by the search engine has been proposed [2]-[6]. One of the methods of the clustering web pages, a method that vectorizes web pages using the vector space method [7], calculates their similarity, clusters them, and classifies them for each content has been proposed [8]-[12]. The vector space method is used mainly in text mining [13]-[15] and information retrieval [16], [17]. The method represents the contents of a document by a feature vector based on the degree of importance of words used in it and calculates the similarity with other documents. Highly similar documents belonging to the same group are considered. Then, by replacing documents with the text in web pages, Web pages can be clustered.

Fig. 1 presents an example of web page clustering. If the search term "speed" is entered in a search engine, pages for “speed test” of line speed, “speed” of card game, and “speed” of a movie are mixed in the search results. Clustering classifies them by content, and users can easily find the desired pages.

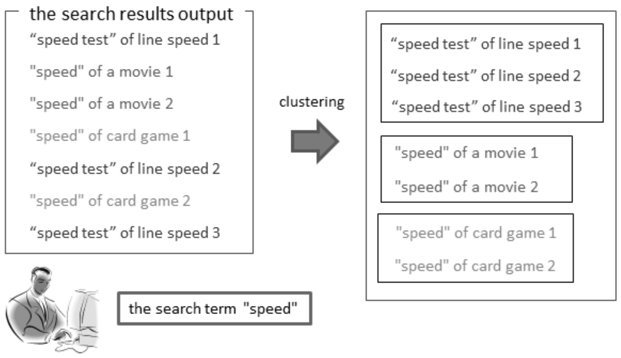


Fig. 1. Example of web page clustering.

However, the vector space method that has been conventionally used has three problems. First, clustering accuracy is low. It is necessary to calculate the importance of words when using this method, and the tf-idf method that is commonly used lacks accuracy [10]. We use using co-occurrence information of words [18], [19] as reference and by using a new indicator of importance, we seek to improve clustering accuracy. Second, the computational cost is high [4], [9], [20]. When using all words that appear at least once on every page with the vector space method, the number of words is enormous. Users require a fast response in their web search. By reducing the number of words [20], we hope to reduce the computational cost. Third, the user cannot determine the contents of the cluster without checking some of its pages. Thus, we propose a method that automatically presents words that represent the characteristics of each cluster [9], [16].

In this study, we propose a method to solve these three problems. First, we attempt to resolve the low clustering accuracy by using a new indicator of importance that consists of the distance to the search term in the text and tf-idf. We believe that this new indicator is effective based on the hypothesis that words that appear closer to the search term represent more clearly the meaning of the search term used in the page. Second, we attempt to resolve the problem of high computational cost by using only words that are close to the search term with the vector space method. Third, we attempt to solve the problem that users must spend time determining the contents of each cluster. For this problem, it is said a method that automatically presents words that represent the characteristics of each cluster after clustering is effective [9], we use the same method. In this study, we present words that appear on more than half of all pages in each cluster and have high importance as words that represent the characteristics of

Manuscript received May 7, 2013; revised July 22, 2013.

Ryutaro Akiyama was with the Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science, Japan. He is now with the Department of Industrial and Management Systems Engineering, Creative Science and Engineering, Graduate School, Waseda University, Japan (e-mail: r-akiyama@akane.waseda.jp).

Katsutoshi Kanamori and Hayato Ohwada are with the Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science, Japan (e-mail: katsu@rs.tus.ac.jp, ohwada@rs.tus.ac.jp).

DOI: 10.7763/IJEEEE.2013.V3.275

each cluster.

Section II describes the proposed method. Section III discusses verification experiments to confirm the effectiveness of the proposed method. Section 4 presents our conclusions.

II. The Proposed Method

In this section, we describe the conventional vector space method of web page clustering [8], and the proposed method that solves the three problems of this conventional method.

This study solves the problems of the conventional method mainly by using the distance to the search term in the text. We define the distance to the search term as follows. The distance to the search term is calculated for each sentence. The words in the sentence that contains the search term are distance “0”, those in the next sentence are distance “1”, and those in the sentence after that are distance “2”. The distance for the word that appears in more than one sentence is the shortest of the distances. An example of distance to the search term "speed" is presented below. The following itemization is obtained by dividing a document into sentences. Numbers are indicated at the beginning of each sentence.

1) In November 1994, Fox Video released Speed on VHS and LaserDisc formats for the very first time.
2) Rental and video sales did very well and helped the film's domestic gross.
3) The original VHS cassette was only available in standard format at the time and in 1996 Fox Video re-released a VHS version of the film in widescreen allowing the viewer to see the film in a similar format to its theatrical release.
4) In 1998, 20th Century Fox Home Entertainment released Speed on DVD for the very first time.
5) The DVD was in a widescreen format but, other than the film's theatrical trailer, the DVD contained no extras aside from the film.
6) In 2002, Fox released a special collector's edition of the film with many extras and a remastered format of the film.
7) Fox re-released this edition several times throughout the years with different covering and finally, in November 2006, Speed was released on a Blu-ray Disc format with over five hours of special features.

The search term “speed” appears in sentences 1, 4, and 7. Therefore, words in sentences 1, 4, and 7 are distance “0”. Words appearing in sentences 2, 3, 5, and 6 are distance “1”. For example, “VHS” in sentence 1, “DVD” in sentence 4, and “Blu-ray” in sentence 7 are distance “0”. Rental" in sentence 2 and “widescreen” in sentence 5 are distance “1”. Here, “VHS” is distance “0” because “VHS” appears in sentence 1, which is distance “0”, and in sentence 3, which is distance “1”. The shortest distance (“0”) is the distance for "VHS."

A. Words Used for the Vector Space Method

The conventional method gets the text of each web page, splits the text into words and identifies their parts of speech using morphological analysis, and uses all the necessary parts of speech for the vector space method.

However, the proposed method selects only words for which the distance to the search term is x or less from all necessary parts of speech and uses them for the vector space method. In Section III, we will determine an appropriate value for x and change x accordingly. This method will reduce the number of words based on the hypothesis that only words that appear close to the search term can represent the meaning of the search term used in the page. Thus, it is expected that clustering accuracy will not decline and that computational cost will be reduced.

B. Importance of Words

When vectorizing web pages using the vector space method, it is necessary to calculate the importance of words. The conventional method calculates the importance of words using the tf-idf method. Tf assigns greater importance to a word that appears more frequently in a document. For example, (1) calculates *tf (D,T)* for word *T* in document *D*.

$$tf(D,T) = \frac{the\ number\ of\ word\ T\ in\ document\ D}{the\ number\ of\ all\ words\ in\ document\ D} \qquad (1)$$

Idf assigns greater importance to a word that appears in fewer documents. For example, (2) calculates *idf (T)*for word *T*. Here, *N* is the number of all documents, and *df (T)* is the number of documents in which word *T* appears.

$$idf(T) = \log\frac{N}{df(T)} + 1 \qquad (2)$$

Tf-idf consists of tf and idf. (3) calculates *tf-idf(D,T)* for word *T* in document *D*.

$$tf\text{-}idf(D,T) = tf(D,T) \times idf(T) \qquad (3)$$

However, the proposed method uses a new indicator of importance that consists of the distance to the search term in the text and tf-idf. (4) calculates the importance used in the proposed method. Here, *R* is the distance to the search term, and *y* is the degree of importance for the distance. In section 3, we will determine an appropriate value for *y*, and change it accordingly.

$$v(D,T) = tf\text{-}idf(D,T) \times (\tfrac{1}{R+1})^{y} \qquad (4)$$

We expect that clustering accuracy using this method will be higher than that using the conventional method based on the hypothesis that words that appear closer to the search term represent more clearly the meaning of the search term used in the page and are important.

C. Extraction of Words that Represent the Cluster

With the conventional method of simply clustering web pages using the vector space method, a user must determine the contents of a cluster by checking some pages in the cluster. For this problem, it is said a method that automatically presents words that represent the characteristics of each cluster after clustering is effective [9]. We make it easier for users to understand the characteristics of each cluster by using the same method. This study presents three words that appear on more than half of all pages in each cluster and have the highest importance as words that represent the characteristics of each cluster. Fig. 2 presents an example of words that represent the characteristics of each cluster in search term "speed." The characteristic of cluster1 is represented by “test,” “line,” and “Internet,” and the characteristic of cluster2 is represented by “movie,” “Jan,” and “Bont.”

Cluster 1: test line Internet

Cluster 2: movie Jan Bont

Cluster 3: card game player

Cluster 4: the others

Fig. 2. Example of words that represent each cluster.

III. Experiments and Consideration

This section describes two verification experiments conducted to confirm the effectiveness of the proposed method and discusses the experiment results. The first experiment was a comparative experiment for indicators of the importance of words. The second was a comparative experiment for words used for the vector space method. We conducted these two experiments in Japan for Japanese search terms.

A. Comparative Experiment for Indicators of the Importance of Words

1) Experiment method

Using (3), which is used in the conventional method that uses the tf-idf method, and (4), which is used in the proposed method, we calculated the importance of words, clustered web pages, and compared the evaluations for the clustering using each formula. The results verified the effectiveness of the proposed method. The experiment using (4) was conducted by changing the value of *y*, which is the degree of importance for distance: the higher the value, the more important the distance. We used all nouns in web pages for the vector space method in this experiment.

In this experiment, we used the Purity/Inverse Purity indicator to evaluate the clustering. This is an evaluation indicator approving that what should belong to the same cluster separate into multiple clusters. The Purity/Inverse Purity indicator is defined as follows. The cluster set of result is $C=\{C_1,\ldots,C_i,\ldots,C_M\}$, and the cluster set of correct results is $L=\{L_1,\ldots,L_j,\ldots,L_K\}$. (5) calculates $Precision(C_i, L_j)$ of the cluster of result C_i for an arbitrary cluster of correct result L_j.

$$Precision\left(C_i, L_j\right) = \frac{|C_i \cap L_j|}{|C_i|} \tag{5}$$

Purity, a weighted average of maximum Precision, is calculated using (6), where N is the number of all search results.

$$P = \sum_{i=1}^{M} \frac{|C_i|}{N} max_j Precision\left(C_i, L_j\right) \tag{6}$$

(7) calculates $Recall(C_i, L_j)$ of the cluster of result C_i for an arbitrary cluster of correct result L_j.

$$Recall\left(C_i,\, L_j\right) = \frac{|C_i \cap L_j|}{|L_j|} \tag{7}$$

Inverse Purity, a weighted average of maximum Recall, is calculated using (8).

$$IP = \sum_{j=1}^{K} \frac{|L_j|}{N} max_i Recall\left(C_i, L_j\right) \tag{8}$$

We use the F-measure, which is a harmonic mean, for comprehensive evaluation. (9) indicates the F-value calculated by the F-measure.

$$F = \frac{1}{\alpha\frac{1}{P}+(1-\alpha)\frac{1}{IP}} \tag{9}$$

In this experiment, the value of α was 0.5.

We calculated the F-value in each threshold for similarity of clustering, changing the threshold in the range of 0.4 to 0.1, and used the F-value to evaluate the clustering. We used the Japanese search terms “スピード (“speed” in English),” “トレーナー (“trainer” and “sweatshirt” in English),” and “ワンピース (“onepiece” of comics and “one-piece dress” in English),” each of which has multiple meanings. We clustered 100 web pages of search results obtained for each search term. First, we created cluster sets of correct results for each search term by hand.

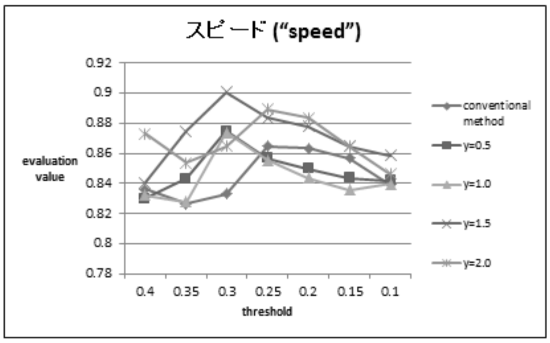


Fig. 3. Results of the importance experiment for search term “スピード (“speed”)”.

2) Results and considerations

Fig. 3 indicates that for search term “スピード,” *y* = 1.5 has the highest evaluation and that *y* = 1.5 has a higher evaluation than that with the conventional method for all thresholds. Likewise, Fig. 5 indicates that for search term “ワンピース,” *y* = 1.5 has the highest evaluation and that *y* = 1.5 has a higher evaluation than that with the conventional method for all thresholds. Fig. 4 indicates that for search term “トレーナー,” though the conventional method has the highest evaluation for some thresholds, *y* = 0.5 has the highest evaluation for the threshold near 0.1, in which the evaluation is maximum. Thus, the proposed method is effective when determining the value of *y* well. However, no optimal value for *y* can be determined because *y* = 1.5 is the best for search term “スピード” and “ワンピース”, and *y* = 0.5 is the best for search term “トレーナー”. In this study, we focused on y = 1.0. The highest evaluation (*y* = 1.0) is higher than that of the conventional method for all search terms.

While the threshold decreases, the evaluation should increase and after it reaches the highest point, the evaluation should decrease. For search term “トレーナー”, after it reached the highest point at the threshold 0.1, the evaluation decreased with the decrease of the threshold. However, for search term “ワンピース”, the evaluation did not reach the highest point and was increasing with the decrease of the threshold. This indicates that for search term “ワンピース”, web pages of search results were not classified for each content well. The cluster set of correct results for search term “ワンピース” has a cluster that has a large number of elements. Due to (3) and (4) use idf, the importance of words that represent a cluster that has a large number of elements did not become high, and therefore this problem occurred.

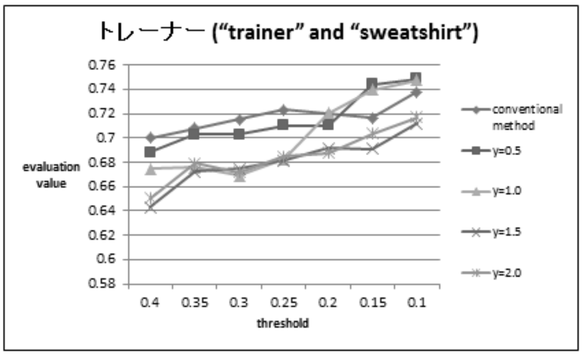


Fig. 4. Results of the importance experiment for search term “トレーナー (“trainer” and “sweatshirt”)”.

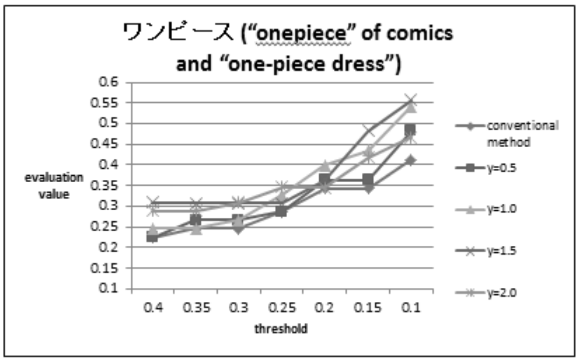


Fig. 5. Results of the importance experiment for search term “ワンピース (“onepiece” of comics and “one-piece dress”)”.

B. Comparative Experiment for Words Used for the Vector Space Method

1) Experiment method

We used all the nouns in the web pages (below, all words) and only words for which the distance to the search term is *x* or less selected from all the nouns in the web pages. We clustered the web pages and compared the evaluation and the processing time for clustering using each word with the vector space method. We verified the effectiveness of the proposed method by comparison. This experiment was conducted by changing the value for *x*, which is distance. We used (4) in which *y* is 1.0 as an indicator of importance. In this experiment, as in the first experiment, we calculated the F-value at each threshold for similarity of clustering, varying the threshold from 0.4 to 0.1, and compared the F-value to evaluate the clustering. We used the Japanese search terms “スピード (“speed”)”, “トレーナー (“trainer” and “sweatshirt”)”, and “ワンピース (“onepiece” of comics and “one-piece dress”)”, and clustered 100 web pages of search results obtained for each search term.

2) Results and considerations

Fig. 7 indicates that for search term “トレーナー”, clustering using words for which the distance is 5 or less is evaluated the highest, and clustering using words for which the distance is 5 or less is evaluated more highly than clustering using all words at all thresholds. Moreover, Fig. 9 indicates that the processing time for clustering using words for which the distance is 5 or less is one fourth that for clustering using all words. Fig. 6 indicates that for search term “スピード”, clustering using words for which the distance is 5 or less is more highly evaluated than clustering using all words at most thresholds. In addition, the highest evaluation of clustering using words for which the distance is 5 or less is higher than that of clustering using all words, and the processing time is one fourth that for clustering using all words. Fig. 8 indicates that for search term “ワンピース”, clustering using words for which the distance is 5 or less is more highly evaluated than that using all words at most thresholds, and the processing time is one eighth that for clustering using all words. These results confirm the effectiveness of using only words for which the distance is 5 or less.

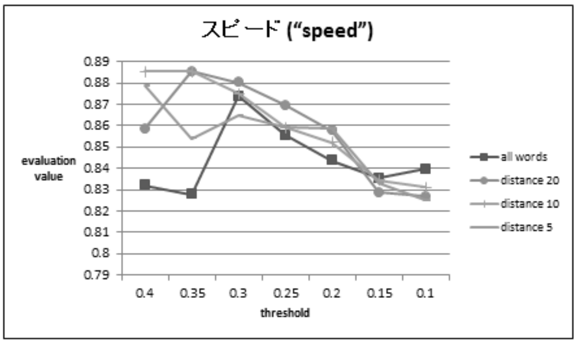


Fig. 6. Results of the word experiment for search term “スピード (“speed”)”.

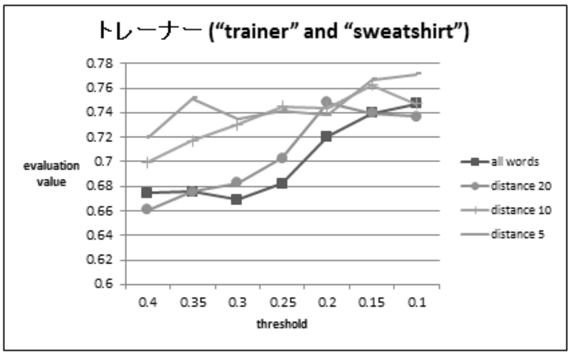


Fig. 7. Results of the word experiment for search term “トレーナー (“trainer” and “sweatshirt”)”.

As with the first experiment, for search term “ワンピース”, the evaluation did not reach the highest point and was increasing with the decrease of the threshold.

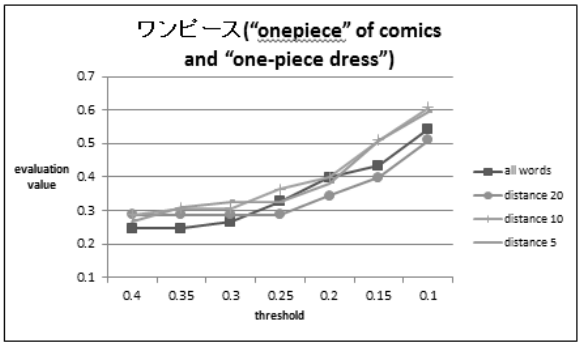


Fig. 8. Results of the word experiment for search term “ワンピース (“onepiece” of comics and “one-piece dress”)”.

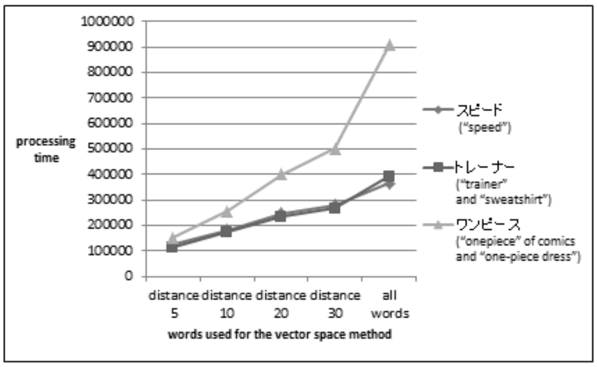


Fig. 9. Processing time for clustering using each word.

IV. Conclusions

This study proposed a method to solve the problems of web page clustering using the vector space method. First, we attempted to resolve the low clustering accuracy by using a new indicator of importance that consists of the distance to the search term in the text and tf-idf. Second, we attempted to resolve the high computational cost by using only words that are close to the search term for the vector space method. Third, we attempted to solve the problem that users must spend time determining the contents of each cluster by presenting three words that appear on more than half of all pages in each cluster and have the highest importance.

The verification experiments indicated that the method that uses only words for which the distance to the search term is 5 or less and calculates the importance of words using (4) in which y is 1.0 is more highly evaluated and has shorter processing time for clustering than the conventional method. Web pages of search results for search term that has a cluster that has a large number of elements were not classified for each content well.

As future work, we will attempt to determine more appropriate values for x and y by conducting more experiments for other search terms, compare the proposed method and other methods, resolve a new problem indicated by the experiment, and conduct experiments for other languages.

References

[1] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama, "Evaluation methods for web retrieval tasks considering hyperlink structure," *IEICE Transactions on Information and Systems*, vol. 86, no. 9, 2003, pp. 1804-1813.

[2] C. Carpineto, S. Osiński, G. Romano, and D. Weiss, "A survey of web clustering engines," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, pp. 17, 2009.

[3] Özel and S. Ayşe, "A web page classification system based on a genetic algorithm using tagged-terms as features," *Expert Systems with Applications*, vol. 38, no. 4, pp. 3407-3415, 2011.

[4] C. M. Chen, H. M. Lee, and Y. J. Chang, "Two novel feature selection approaches for web page classification," *Expert systems with Applications*, vol. 36, no. 1, pp. 260-272.

[5] Mangai, J. Alamelu, and V. S. Kumar, "A novel approach for web page classification using optimum," *IJCSNS*, vol. 11, no. 5, pp. 252, 2011.

[6] Z. Xu, F. Yan, J. Qin, and H. Zhu, "A web page classification algorithm based on link information," *Distributed Computing and Applications to Business, Engineering and Science (DCABES), Tenth International Symposium on. IEEE*, 2011.

[7] Salton, Gerard, and M. J. McGill, Introduction to modern information retrieval, 1986.

[8] N. Otani. "Application of vector space model in information retrieval," Musashino industrial University Faculty of Environment and Information Studies research paper, pp. 3-6, in Japanese, 2004.

[9] T. Miyoshi and H. Joichi, "The improvement of web page retrieval by page grouping using fuzzy reasoning," *International Journal of Innovative Computing, Information & Control*, vol. 2, no. 1, pp. 237-247, 2006.

[10] T. Miyoshi and H. Joichi, "Comparison with fuzzy reasoning and modified tf-idf in page grouping for the result of web retrieval," *International Journal of Innovative Computing, Information and Control*, vol. 3, no. 2, pp. 307-317, 2007.

[11] H. J. Zeng, Q. C. He, Z. Chen, W. Y. Ma, and J. Ma, "Learning to cluster web search results," in *Proc. The 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2004.

[12] Y. H. Zhao and X. F. Shi, "The application of vector space model in the information retrieval system," *Software Engineering and Knowledge Engineering: Theory and Practice*, Berlin Heidelberg: Springer 2012. pp. 43-49.

[13] L. P. Jing, M. K. Ng, and J. Z. Huang, "Knowledge-based vector space model for text clustering," *Knowledge and information systems*, vol. 25, no. 1, pp. 35-55, 2010.

[14] M. Danesh and S. Hossein, "Text document clustering using semantic neighbors," *Journal of Software Engineering*, vol. 5, no. 4, pp. 136-144, 2011.

[15] X. Tian and Y. Du, "Improve VSM text classification by title vector based document representation method," *Computer Science & Education (ICCSE), 2011 6th International Conference on IEEE*, 2011.

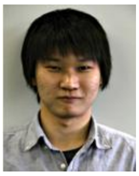
[16] H. Mochida, S. Omachi, and H. Aso, "Web page retrieval system by automatic detection of topic words," *IEEJ Transactions on Electronics, Information and Systems*, vol. 127, no. 12, pp. 2142-2147, 2007.

[17] C. Fautsch and J. Savoy, "Adapting the tf idf vector-space model to domain specific information retrieval," in *Proc. the 2010 ACM Symposium on Applied Computing*, ACM, 2010.

[18] Wartena, Christian, R. Brussee, and W. Slakhorst, "Keyword extraction using word co-occurrence," *Database and Expert Systems Applications (DEXA), 2010 Workshop on. IEEE*, 2010.

[19] C. Peng, N. Feng, and H. Ma, "Document clustering algorithm based on word co-occurrence," *Computer Engineering*, vol. 38, no. 2, 2012.

[20] Y. C. Liu, X. L. Wang, and B. Q. Liu, "A feature selection algorithm for document clustering based on word co-occurrence frequency," Machine Learning and Cybernetics, IEEE, in *Proc. 2004 International Conference*, vol. 5, 2004.



Ryutaro Akiyama is a student at the Department of Industrial and Management Systems Engineering, Graduate School, Waseda University, Japan. He received Bachelor Engineering from Tokyo University of Science, Japan 2013. His research interest is in text mining.

Katsutoshi Kanamori is an assistant Professor of Industrial Administration, Tokyo University of Science, Japan. He received Doctor of Science from Tokyo University of Science 2009. He has been working on Artificial Intelligence and Formulation of Creativity.

Hayato Ohwada is a professor of Industrial Administration and Director of Division of Next Generation Data Mining Technology, Tokyo University of Science, Japan. He received Doctor Engineering from Tokyo University of Science 1988. He has been working on Machine Learning and Inductive Logic Programming.