A Summarization Method of News Articles based on Time and Content Information

Ren, Yi

(出版者 / Publisher)
法政大学大学院情報科学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編 / 法政大学大学院紀要. 情報科学研究科編
(巻 / Volume)
9
(開始ページ / Start Page)
27
(終了ページ / End Page)
32
(発行年 / Year)
2014-03
(URL)
https://doi.org/10.15002/00010517

A Summarization Method of News Articles based on Time and Content Information

Yi Ren
Graduate School of Computer and Information Sciences
Hosei University
Tokyo, Japan
yi.ren.8b@stu.hosei.ac.jp

***Abstract*—In this paper, because the information in one news article is too little to make readers understand the whole view of the subject, and in order to help readers get a full view of it, we propose a method to summarize the content from news articles about one subject, which is by processing a number of news articles which have the same subject. We use news search engine to collect news articles and perform preprocessing upon them in order to calculate each article's vector by considering feature of words. We proposes a clustering method in consideration of similarity between articles and time property of news. We also perform the visualization of the result. We evaluated our method by performing conducting experiments using news articles being collected and verified the effectiveness of the method.**

Keywords—clustering, news theme, keyword extraction, topic detection.

I. INTRODUCTION

With the development of Internet, it becomes much easier to get information of news from Internet. Many news articles are posted on news websites. The theme of news article is called to be subject and each news may describes a certain side of the event. But we may not know the related things about the subject in advance. Although many news websites may list latest news and we can also search for related information about the subject, not only the information in one article is not enough, but also it may cost much time to read whole related information. So readers may not grasp the full view (related information) of the subject in a short time. That is the reason why we need a method to help readers understand the subject of news by reading the result of sorting out information from a number of related news articles. Our goal is to let readers grasp a full view of the subject.

Because the key point of this research is to extract information from a number of articles, we focus on how to identify topics from them. But at first, we should classify them into different categories which talk about different topics. Many conventional researches about how to classify document through counting similarity between articles or other method are proposed, but there is not effective method that can classify articles about one subject based on content yet.

The thesis paper of [01] purposes a method that can detect the "latest topic words" from incoming news articles to understand topical overview in them. Two aspects of topic in news stream are considered, one is accession of articles which play up same event, and the other is a continuation of follow up articles. The method extracts a few latest topic words from each cluster using the significance which is evaluated after applying clustering method to news collection. The key point of this research is counting the freshness of news articles based on time stamp and calculating the degree of topic by considering freshness, similarity and feature of genre clusters.

Two methods for classifying articles using feature vectors are proposed in [02]. The first method uses lexical co-occurrences while the second method uses the EDR electronic dictionary as a thesaurus to generate feature vectors.

The method proposed in [03] is performing hierarchic clustering upon document collection that includes several topics and building the structure based on similarity between clusters' words features. It can make readers get a full view through extracting keywords or entity name from clusters in a fast and effective way.

The research of [04] proposes a keyword extraction algorithm aiming at single document using word co-occurrence statistical information but not a corpus.

A method of discovering semantically related news topics for given time series data is proposed in [08]. As compared with most existing methods which find related news mainly based on the similarity among documents, they use both textual and temporal behavior of the news, and expect to find related topics which have some impact on the time series data.

However, there is not an effective method that can extract topics from news articles that have the same subject in consideration of time stamp yet.

We have published a thesis paper in the 27th Annual Conference of the Japanese Society for Artificial Intelligence which is the reference [9]. We have proposed the method of considering content and time property in that paper, but we had not conducted experiments to verify it effectiveness.

This research is based on these conventional researches and adopts the thinking of considering time stamp when extracting topic from news collection. We propose a method that adopt preprocessing, including morphological analysis and perform clustering in consideration of similarity and time stamp.

This paper is organized as follows. The next section we detailed algorithm and in the third section, we made evaluation

Supervisor: Prof. Mina Akaishi

and discussion. Finally, we summarize our contributions and conclude the paper.

II. METHOD PROCEDURE

A. Obtaining News Articles

This research is aim to expand the content related to the news subject, so we use news articles as the source sample. At the present stage, we manually collect news articles by using search engine, such as Google, which has the function of news search. Firstly, we conduct search by using subject as search keyword and get pages from the search results. Each search result is linked to a page in which a news article about the subject is posted. We extract the news article from each page, and we extract posted date, title and text from each article. Thus we turn every article to be one record. We will call time property as time stamp the in this paper. Finally, each record is written into text file with the format of "date (/n) title (/n) text". Experiment data is also prepared in this way.

We have not made this step automatic yet. We only collect experiment data manually from Internet using Google news research. Besides, this research is targeting news articles written in Japanese language.

B. Preprocessing

In this step, we firstly conduct morphological analysis. This analysis is conducted upon part of every record, including title and text. Morphological analysis is the analysis that uses dictionary (word list of information with part of speech) and knowledge (the rules of grammar) of grammar to break statements written in natural language into morphemes (which is the minimum unit with meaning in language.

Then we remove the word of type of particle, blank, mark and some other words that we do need when we analyze the content of the text. And we combine adjacent words into a new one if they are of the same type. For example, "7", "8" and "7", they are all nouns that represent the number, and we combine them into "787". When a word is a prefix or suffix, we combine it with the word after or before it. So what we really take into consideration is words of noun, verb, adjective, words of alphabet, numbers and so on.

We also get rid of the word of research keyword, because it will have an influence on the results such as counting the frequency of each word in every record in that the keyword frequency will be too big to have impact on the total result if we take it into consideration.

The next process is to deal with time stamp. We sort out each time stamp into the format of eight numbers ("year-month-day"), such as "20130101". The time stamp will be used in the step of D (Clustering) and F (Timeline).

Finally, we can obtain a list of words from the result of morphological analysis.

C. Calculating Vector of Record

In this step, we count vector of each record by applying the method of tf-idf. Tf-idf is the product of two statistics, term frequency and inverse document frequency. In this research, we do not adopt the method of counting vector only by term frequency because words with higher frequency may show that the word is closely related to the event but it cannot be said that they are representative of one side of the subject. Conversely, using tf-idf makes the importance of words that appear in many records decrease and the importance of words that only appear in some particular records increase. In this way, the words with high tf-idf value may be representative of some side of subject. In other words, the news articles about one subject may talk about different sides, and in this way, we expect to make the vectors can represent the relationship between news articles and different topics about the subject, and the higher the similarity between two vectors is, the more possibility that the two describe the same topic about the subject is.

The vector d_j of each record is defined as follows:

$$d_j = (x_{j1}, x_{j,2}, \cdots, x_{iV}) \ (i = 1,2, \cdots, N) \quad (1)$$

And the $x_{j,i}$ is calculated by multiplying tf and idf value of each word. N is the number of all words.

$$tf_{i,j} = n_{i,j} / \sum_k n_{k,j} \quad (2)$$

$$idf_i = \log N / \{d: d \ni t_i\} \quad (3)$$

$$x_{j,i} = tf_{i,j} \times idf_i \quad (4)$$

N is the number of all records (news articles). The $n_{i,j}$ in the formula 2 is the number of occurrence of word i in the record j. The $\{d: d \ni t_i\}$ in formula 3 is the number of record that contains word i.

D. Clustering

In this step, we use the k-means clustering. K-means clustering is a method of cluster analysis and aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

Because we do not know how many clusters we will partition the samples into, so in this research, we adopt a recursive k-means algorithm. We set a threshold which is the maximum number of elements in one cluster.

The clustering algorithm in this research is a little different from the common k-means algorithm.

The difference is reflected in two steps, which are the step of counting correlation coefficient and judging the cluster is too big or not. We propose an algorithm of counting mean in k-mean clustering method. The formula of (5) represents the method of counting correlation coefficient.

The figure below shows the basic flow of performing clustering of this research.

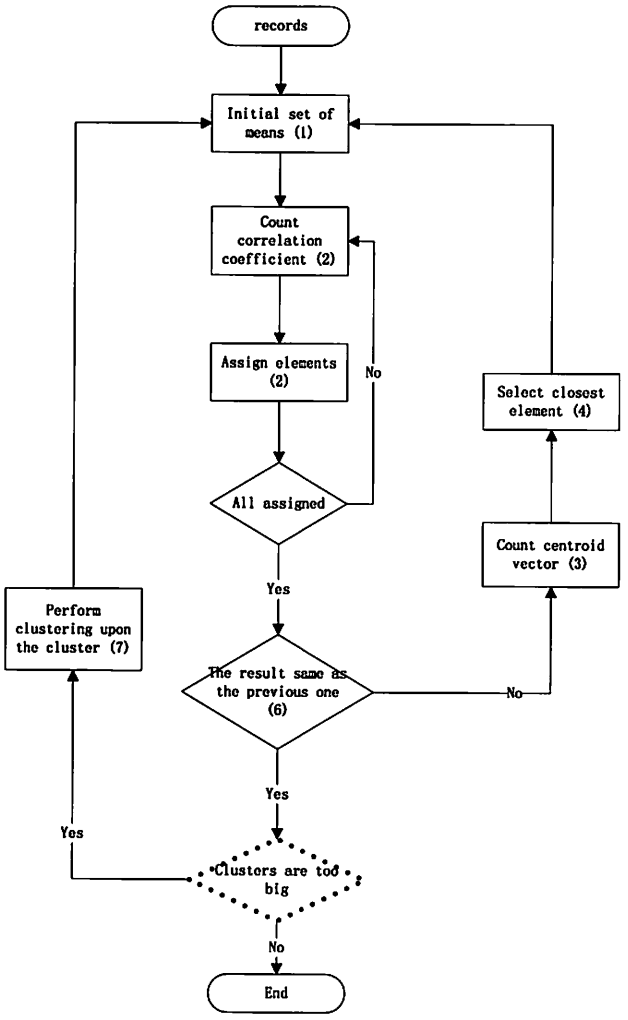

Fig. 1 Flow of clustering.

Step (1). Select an initial set of k means m_1, m_2, ..., m_k (in this research, k is set to the value of $\sqrt[4]{n}$, n is the number of articles). Because the result of k-means clustering varies from the initial set, and in order to value this method easier, we pick the first k words as the initial set.

Step (2). Assignment step: Assign each observation to the cluster whose mean is closest to it. We define a correlation coefficient in this step.

$$CC_{i,j} = Similarity(i,j) \times \frac{1}{TimeInterval(i,j)^2} \quad (5)$$

The Similarity (i, j) is the cosine similarity between two records, shown in the function below, which is calculated by using the vectors of the two.

$$\text{Similarity}_{i,j} = \frac{\sum_{k=0}^{V} x_{i,k} \times x_{j,k}}{\sqrt{\sum_{k=0}^{V} x_{i,k}^2} \times \sqrt{\sum_{k=0}^{V} x_{j,k}^2}} \quad (6)$$

The TimeInterval (i, j) is the time interval between dates of two records. The time interval is calculated by the formula below. The date (i) is the time stamp of record i.

$$\text{TimeInterval}(i,j) = \| date(i) - date(j) \| + 1 \quad (7)$$

The reason why we take time stamp into consider when we calculate the correlation coefficient is that the different sides of one subject are related to time stamp. In other words, the related events happen in some particular points in time, so we expect that the news articles that describe the same event or the same side may have the more closely time property. The more closely the two events happen, the more possibly they are talking about the same thing.

And we compute the correlation coefficient based on similarity and time stamp, so the bigger the value of $CC_{i,j}$ is, the more closely mean between two records is. So after we count each CC value between each record that have not been assigned and each mean, we assign the record to the mean that has the biggest value of CC.

Step (3). After assign all the records, calculate the centroid vector of every cluster.

$$\text{centorid(i)} = (x_1, x_2, \ldots, x_n) \quad (8)$$

i in the formula (8) above is the number of cluster, and n is the number of all words.

$$x_j = \frac{\sum_i^n x_{j,k}}{V} \ (k = 0, 1, \ldots V) \quad (9)$$

The V is the number of elements in each cluster.

Step (4). Select the element in each cluster that has the biggest value CC with the centroid as the new initial set of k means.

Step (5). Repeat the step 2 to assign all the elements that have not been assigned by step (4).

Step (6). If the clustering result turns out to be the same as the previous one, it can be deemed to be stable result. Then finish the clustering of this time.

Step (7). We set a threshold to detect whether the elements number in clusters is too big. If it is true, perform clustering upon the elements of the cluster using the step from 1 to 6. In this research, the threshold is set to three tenths of number of all words empirically.

In a conclusion, the K-means algorithm in this research has been altered using hierarchic clustering and the mean that is calculated by considering similarity and time interval.

E. Extracting Keywords

In this research, we use cl-keyword to express the general content of every cluster. Because we count similarity between vectors when clustering is been performed, when we distract cl-keyword from every cluster, we sum up all the vectors in each cluster and get the total vector as the following formula shows:

$$V_i = \left(\prod_{j=0}^{m}(1 + x_{j,0}), \prod_{j=0}^{m}(1 + x_{j,1}), \ldots, \prod_{j=0}^{m}(1 + x_{m,n})\right) \quad (10)$$

The n is number of words, m is number of elements in a cluster.

Then we select the five numbers of the largest value from V_i as cl-keyword.

F. Showing Clustering result

In this step, we implement the result of clustering visualization using Java language. The result of visualization can

represent the relationship among clusters in tree structure like the figure below.

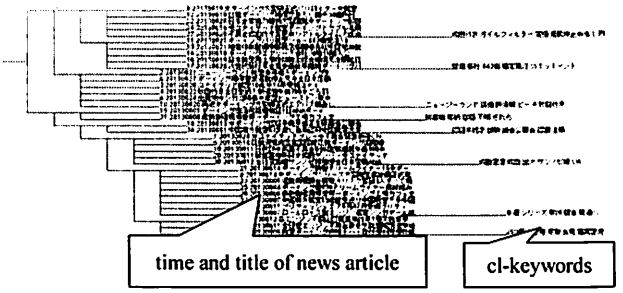

Fig. 2. Visualization of result of clustering.

G. Showing Timeline Result

As we described in part D, different topics about one subject appear in different time. In order to investigate the relationship between time and the appearance of topics, we use timeline to show the start time and end time of every cluster.

The purpose of this research is to extend the content about the subject in order to make reader have a better understanding of the subject in a news article. So the last step of the purposed method is to summarize content of every cluster to represent the general content of every group (cluster). In the present stage, we just list all the titles in every cluster. In the future, we will do more research on how to distract information from every cluster.

The result is shown in the figure below.

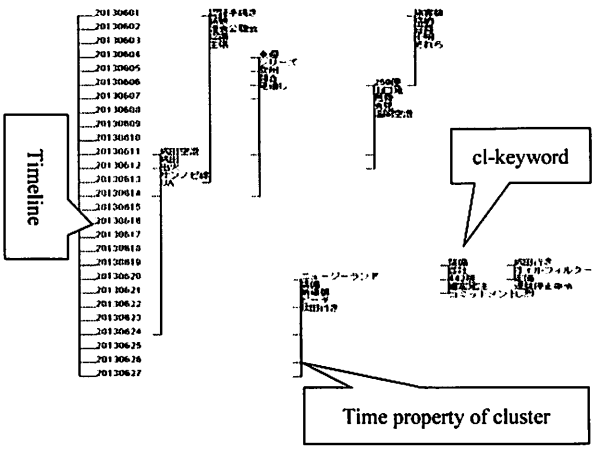

Fig. 3 Visualization of result of timeline.

The vertical line in the left side is the time line, representing the start time and end time of the whole sample. In this example, the start time is March 1st, 2013 and the end time is March 31th, 2013. The vertical lines on the right side of time line represent the start time, end time and keywords of every cluster. And we have sorted the result beforehand by the lasting time of every cluster. In this way, we can easily compare the experiment result of this research's result with other methods, so that we can vertify the effectiveness of this method.

This process is also implemented using Java language.

H. Summarizing

The purpose of this research is to extend the content about the subject in order to make reader have a better understanding of the subject in a news article. So the last step of the purposed method is to summarize content of every cluster to represent the general content of every group (cluster). In the present stage, we just list all the titles in every cluster. In the future, we will do more research on how to distract information from every cluster.

In conclusion, we propose a method that firstly collect news articles using news search engine and transform into records, then conduct preprocessing including morphological analysis upon these text. We perform clustering using the similarity between two vectors that every vector is corresponding to one record. We also consider the time interval between records because topics about the same side of a subject may appear in similar time. We implement the visualization of the result of clustering and time line. And at last we summarize the general content of every cluster.

III. Experiment

To evaluate the effectiveness of this method, we conducted experiments and made comparison with the result of conventional method. Specifically, we did comparison on the following results to evaluate this method.

- Whether the time property has an effect on the result of clustering or not.
- Whether keywords can make readers understand the general content of each cluster or not.
- Correct data rate.

We will call the proposed method "method 1" in the following parts. At the same time, we picked two methods as comparison method, namely method 2 and method 3. The method 2 uses the formula (11) to calculate the value of mean. And the method3 adopts the formula (12) to calculate the value.

$$Distance_{i,j} = Similarity(i,j) \quad (11)$$

$$Distance_{i,j} = Similarity(i,j) \times TimeInterval(i,j) \quad (12)$$

We collected three sets of experiment data using Google news search engine in three different times manually. We made use of the first few pages of the search results. We picked up all the news articles from the webpages. We used "787" as the search keyword. (787 is a new model airplane produced by Boeing, but in the past few months, there are many news that cover some events about the airplane because of cancellations and a lot of investigation into it.) Then we stored them into text files and numbered them as the sample 1, 2, 3.

1. Sample 1: 97 pieces of news articles collected in January 25th, 2013.
2. Sample 2: 43 pieces of news articles collected in March 31th, 2013.
3. Sample 3: 35 pieces of news articles collected in June 27th, 2013.

When we conducted experiments to verify whether the time property has an effect on the result of clustering or not, we made

statistics and analysis upon the results of the timeline.

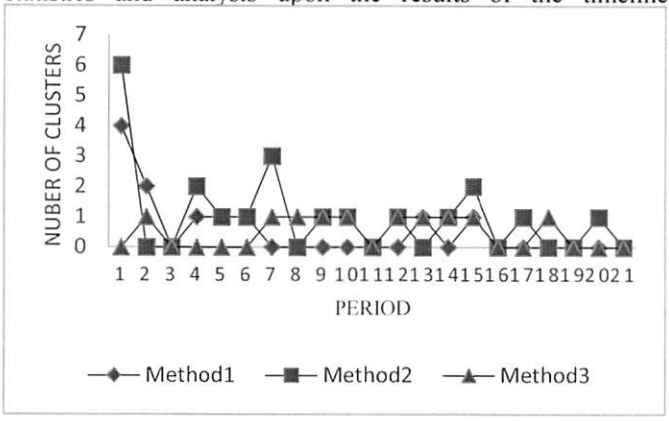

Fig. 4. Time properties of clusters. *(sample 1)*

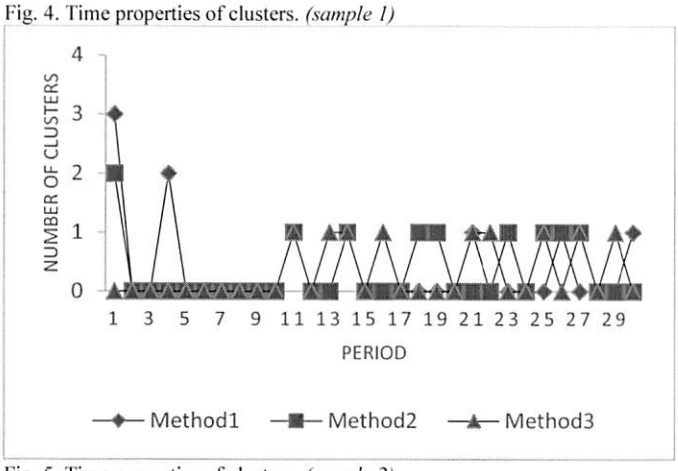

Fig. 5. Time properties of clusters. *(sample 2)*

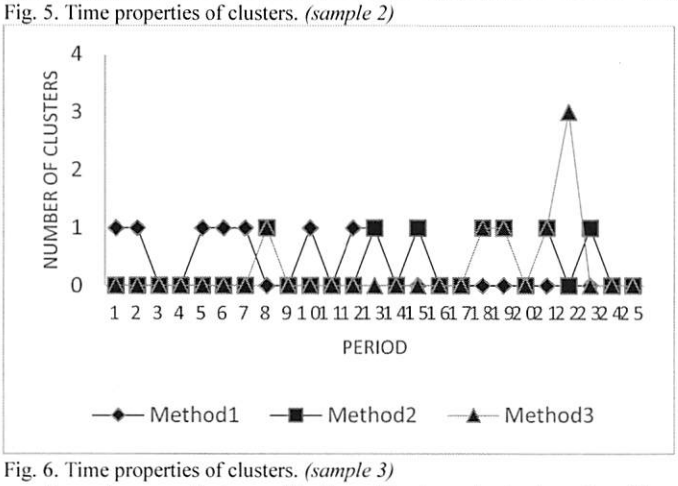

Fig. 6. Time properties of clusters. *(sample 3)*

It can be seen that an effective effect on clustering algorithm by time property is shown in this experiment. It meets the expected goal.

Then we conducted experiment to test if the keyword extraction can effectively extract keywords that can help readers understand the general content of clusters. Since the expressive of words is not objective but is determined by comprehension of readers, we conducted survey in the form of questionnaire. The questionnaire is targeted at readers. And letting making readers being subjected to this investigation, we told readers the information about Boeing 787 and because of some defects, events like suspend of flights, about it have appeared frequently in recent several months. What readers do is to make intuitive judgment that if these words of each cluster can make you understand the general content of articles and make choice from three options which are "cannot understand", "can understand a little" and "can understand". We assigned weights of 0, 1 and 2 to each degree when we counted the result.

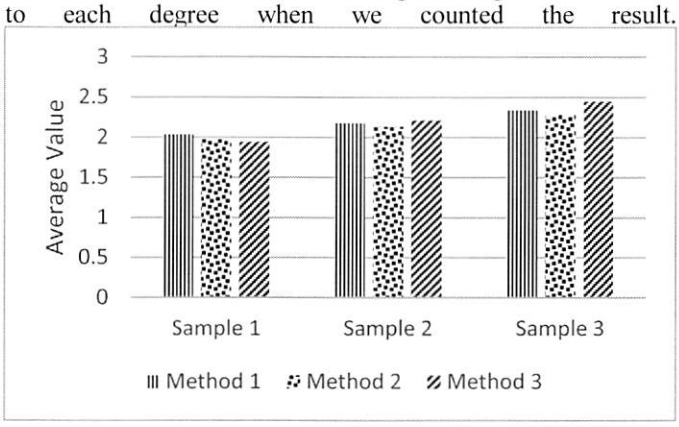

Fig. 7. Average value of questionnaire results

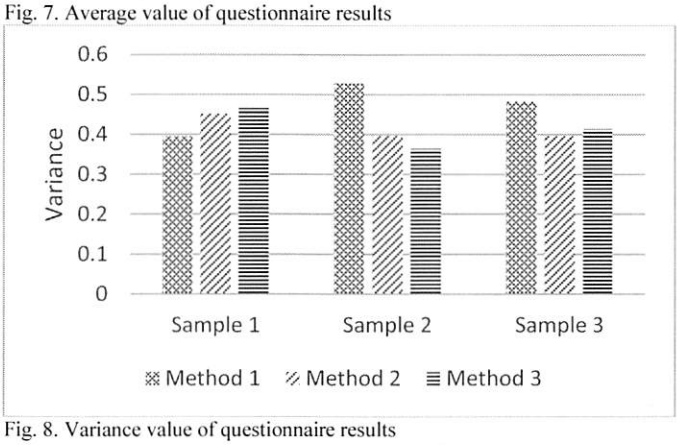

Fig. 8. Variance value of questionnaire results

After we summarized all the results of each candidate, it can be indicated that the minimum of average value is 1.975 and the maximum is 2.45. The variance is between 0.36 and 0.53. We can infer that the keyword extraction of this method turns to be effective and can bring reader a certain degree of understanding general content of cluster.

In fact, it is hard to create correct data because there may be not a clear line between news articles that talk about the same subject. And one piece of news may describe several sides about one subject. In order to verify the effectiveness of the proposed method, we created correct data in two different ways.

In one way, we created correct data based on the general content of news articles. By judging the meaning of title, we classified these articles into groups as the table shows. The row of expected data is what we obtained by classifying articles sample. Each word is the keyword of each group. The rows in right side mean the results of clustering of each method, we compared expected data with results. When two or more elements in one cluster also exist in one group in expected data, we defined these element to be correct data. When some groups in expected data only own one element, if the clustering can generate the cluster has the only element as in expected data, it is defined to be correct data. After we counted the number of correct data, we compared the correct data rate of each method.

We conducted an experiment using another way to create correct data. In this one, we extract words from all the titles whose frequency is not too big or too small. We sort the list of words by frequency and get rid of the first three words and the words that appeared only once. Then we use these words as a basis for classification. We define an element as correct data if it has the same words with other elements in the same cluster.

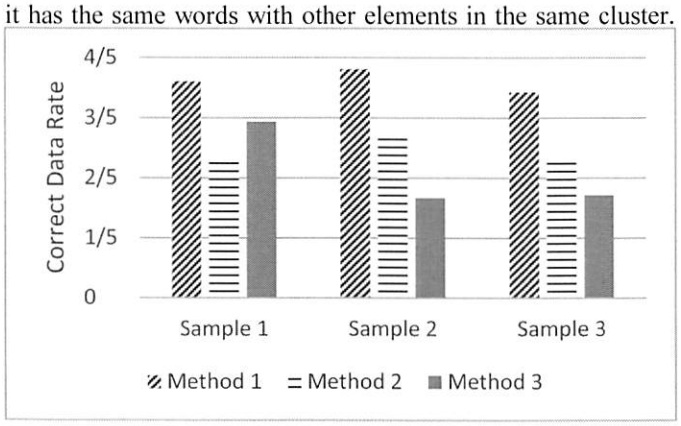

Fig. 9. Comparison of correct data rate.

After we compared the results of purposed method with the conventional method, we infer that our method has the following feature:

1. The scale of generated clusters is a little different from the conventional method. When it is a large sample, this method has a trend to generate fewer clusters than the conventional one. Conversely, when it is not a large sample, the division turns to be more detailed.
2. After performing clustering and extracting keywords of every clusters, the results of this method turn out to be more understandable for readers than the conventional one.
3. This method has the trend to generate more clusters that last shorter time (which mean the time difference between start time and end time of news articles in a cluster). And the average of duration time is much shorter than the conventional method.
4. This method is proven to have a better performance on detecting topics from a large number of news articles. And this also proves that it is possible to do clustering upon news articles based on similarity and time property.

IV. Conclusion

Because in nowadays, there are many news websites. However, it will cost some time for readers to know the more sides of the subject when reading about it. In order to help readers get know the whole information of subject in a short time, this research proposes a method to partition a number of news articles and extract keywords of every group. We can use similarity between articles when we perform clustering, but one news article may involve many sides of a subject. And from research results gotten in one time, there is a trend that some articles about the same sides appear in some certain time. Therefore the method that we proposed considers the time property of every news article.

The experiment result turns out that this method has a better performance in generating clusters and extracting keywords that readers can infer the general content of clusters.

In the future, we will focus on how to improve the method of summarizing the content from news articles and investigate how to extract main information from the result of summarization in order to present full view of subject to readers.

Acknowledgment

This research project would not have been possible without the support of many people. I wish to express my gratitude to my supervisor, Prof. Dr. Akaishi who was abundantly helpful and offered invaluable assistance, support and guidance. Deepest gratitude are also due to the members of the research laboratory, Dr. Sato, without whose knowledge and assistance this study would not have been successful. Special thanks also to all my graduate friends, Mr. Inoue and Mr. Kida, for sharing the literature and invaluable assistance. Also thanks to the collaborators who have made contribution to questionnaires in my study.

References

[1] Yoshihide Sato, Harumi Kawashima, Tsutomu Sasaki and Masahiro Oku, "Latest Topic Words Detection from Chronological News Stream," Information Processing Society of Japan Technical Reports, Japan, 2005(73), 2005, pp. 1-6.

[2] Yuasa Natsuki, Ueda Toru and Toqawa Fumio, "Classifying Articles Using Lexical Co-occurrence in Large Document Databases," Information Processing Society of Japan Technical Reports, Japan, 36(8), 1995, pp. 1819-1827.

[3] Hashimoto Taiichi, Murakami Koji, Inui Takashi, Utsumi Kazuo and Ishikawa Masamichi, "Topic Extraction and Social Problem Detection based on Document Clustering," Sociotechnology Research Network, Japan, 5, 2008, pp.216-226.

[4] Yutaka Matsuo and Mitsuru Ishizuka, "Keyword Extraction from a Document using Word Co-occurrence Statistical Information," the Japanese Society for Arificial Intelligence, Japan, 17-3D, 2002, pp. 217-223.

[5] Fukaya Ryo, Yamamura Tsuyoshi, Kudo Hiroaki, Matsumoto Tetsuya, Takeuchi Yoshinori and Ohnishi Noburu, "Measuring Similarity between Documents Using Term Frequency," the Institue of Electronics, Information and Communication Engineers, Japan, J87-D-II(2), 2004, pp. 661-672.

[6] Mikihiko Mori, "Extraction of Drifted Topics with Time from News Articles," the 22nd Annual Conference of the Japanese of the Japanese Society for Artifical Intelligence, Japan, 2F2-2, 2008.

[7] Fumiyo Fukumoto, Yoshimi Suzuki, "Term Weight Learnig for an Automatic Text Categorisation," Information Processing Society of Japan Technical Reports, Japan, Vol.40 No.4, 1999, pp. 1782-1791.

[8] Yimeng Zhang, Shumian He, Satoshi Oyama, Keishi Tajima, Katsumi Tanaka, "Discovery of Semantically Related Topics for Given Time Series Data," the Database Society of Japan, Japan, Vol.5 No.1, 2006, pp.133-136

[9] Yi Ren, Makoto Sato, Mina Akaishi, "Subject Expansion from Content of News Article," the 27th Annual Conference of the Japanese Society for Artificial Intelligence, Japan, .1F4-6.