A RETRIEVAL METHOD OF SIMILAR QUESTION ARTICLES FROM WEB BULLETIN BOARD

Yohei Sakurai, Soichiro Miyazaki, Masanori Akiyoshi
Osaka University
Yamadaoka 2-1, Suita, Osaka 565-0871, Japan

Keywords: Web bulletin board, Natural sentence input, Question articles, Cosine similarity index.

Abstract: This paper proposes a method for retrieving similar question articles from Web bulletin boards, which basically use the cosine similarity index derived from a user's query sentence and article question sentences. Since these sentences are mostly short, it is difficult to distinguish whether article question sentences are similar to a user's query sentence or not simply by applying the conventional cosine similarity index. In an attempt to overcome this problem, our method modifies the elements of the word vectors used in the cosine similarity index, which are derived from a sentence structure from the viewpoints of common words and non-common words between a user's query sentence and article question sentences. Experimental results indicate that our proposed method is effective.

1 INTRODUCTION

Web bulletin boards have been used in several domains, where chat-style sentences are published as they are, and often including useful information such as consumers' genuine opinions in the contributed articles. However, users who want to obtain information cannot at present efficiently inspect articles containing that information in bulletin boards; there are simply too many articles on Web bulletin boards.

Users generally retrieve by keyword or narrowly searched articles, judging whether an article includes necessary information by reading its title or first sentence. Users must sometimes inspect articles including irrelevant information, too. As a result, it takes a lot of time for them to judge whether the articles are indeed useful.

In this research we propose a method for retrieving similar question articles to a query by natural sentence input to improve retrieval accuracy. Recently, as information retrieval technology (Ohtsuka, 2004; Kishida, 1997; Mochihashi, 2004) has improved, various methods for judging similarity of sentences have been developed. As a natural sentence input, A question-answer system (Sasaki, 2002; Tamura, 2005) that retrieves the answer of the input in the document has also been exammed as a natural style of sentence input. However, these studies only deal with formal sentences like those in newspapers.

A question article on a Web bulletin board is a form by which a question asked in the first contribution, and the answers are given in following contributions. Similarity judgments by conventional methods that only match words are, however, insufficient because the retrieval query sentence and the article question sentence are short.

Consequently, in this we research consider not only matching of the words in the retrieval query sentence and the article question sentence, but also the structure of the sentence. As a result, the accuracy of similarity judgment is improved.

In Section 2, we describe the problems with the retrieval method using natural sentence input. In Section 3, we described a the question sentence retrieval method that applies co-occurrence information about non-common words and concrete procedures to solve the problem. In Section 4, the proposed method and a conventional method are compared by retrieving practical data in order to evaluate the effectiveness of the proposed retrieval method. Section 5 provides a summary.

2 ARTICLE RETRIEVAL FROM A WEB BULLETIN BOARD

2.1 Question Article Retrieval

A question sentence might be included in the first contribution of an article on a Web bulletin board, and users judge whether the article includes the required information by reading it. In this paper we research propose a retrieval method that judges similarity to

Sakurai Y., Miyazaki S. and Akiyoshi M. (2006).
A RETRIEVAL METHOD OF SIMILAR QUESTION ARTICLES FROM WEB BULLETIN BOARD.
In *Proceedings of the First International Conference on Software and Data Technologies*, pages 238-243
DOI: 10.5220/0001315202380243

the user's input retrieval query sentence by using the question sentence in the first contribution. Figure 1 shows an outline of the proposed retrieval method.

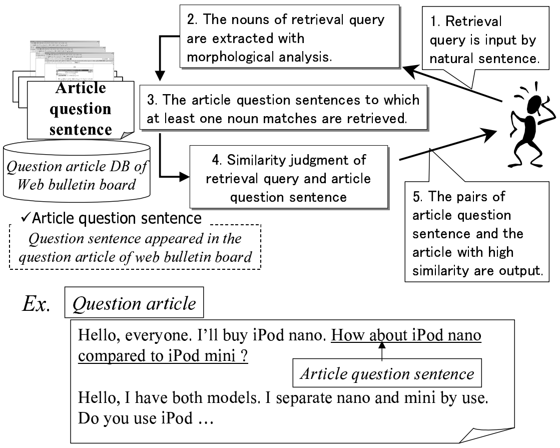
2. The nouns of retrieval query are extracted with morphological analysis.
1. Retrieval query is input by natural sentence.
Article question sentence
3. The article question sentences to which at least one noun matches are retrieved.
Question article DB of Web bulletin board
4. Similarity judgment of retrieval query and article question sentence
5. The pairs of article question sentence and the article with high similarity are output.
✓Article question sentence
Question sentence appeared in the question article of web bulletin board
Ex.
Question article
Hello, everyone. I'll buy iPod nano. How about iPod nano compared to iPod mini ?
Article question sentence
Hello, I have both models. I separate nano and mini by use. Do you use iPod …

Figure 1: Overview of question article retrieval by natural sentence input.

The question article database of a Web bulletin board is composed of a pair comprising the question article and the article question sentence, which is a set of sentences extracted from the first contribution (Skowron, 2005; Li, 2002). The flow of the retrieval procedure is shown below.

Step 1 A user inputs the retrieval query sentence.

Step 2 The input question sentence is analyzed in the morpheme, and the nouns are extracted.

Step 3 The articles, i.e. candidate articles, are retrieved from the question article data base by using a set of extracted nouns.

Step 4 It is judged whether the question sentence in the candidate article is similar to the retrieval query sentence.

Step 5 As a result of the similarity judgment, users receive the article question sentence and question articles in order of their similarity to the question, with the most similar at the top.

The system proposed here extracts the article question sentence from the first contribution as preprocessing, then judges whether the retrieval query sentence is similar to the article question sentences.

2.2 Features of a Targeted Bulletin Board

Since the contribution articles by consumers are published as they are and they are not wellformed with respect to sentences, Web bulletin boards have certain special characteristics. The description characteristics of question sentences are shown below.

Description characteristic 1 Though questions may have the same content, the sequence of their words may be different.

Description characteristic 2 Though the questions may have the same content, their length may be different.

Ex. 1 How convenient is it to write mail with the SH901ic?

Ex. 2 On the SH901ic, is it possible to input T9 input and the bell when a character is input ?

In "description characteristic 1," it is necessary to judge similarity without depending on the sequence of the words. In the examples of "description characteristic 2," both questions are about convenience when mail is written. The first question (Ex. 1) is a vague one, while Ex. 2 asks a question about character input. It is necessary to judge the sentence without depending on the sequence of words and the length of the sentence (number of words). Here we judge similarity by the modified cosine similarity index. Next, the problem of solving similarity judgments of the retrieval query sentence and the article question sentence is described.

2.3 Judgment by Cosine Similarity Index

Figure 2 shows the calculationdefinition of the conventional cosine similarity index. The cosine similarity index considers the retrieval query sentence and the article question sentence to be a set of words, and puts them into the word vector of *n* dimensions. The similarity index between sentences is calculated at the angle of the word vector.

The cosine similarity index tends to be high when there are a lot of words common to both the retrieval query sentence and the article question sentence. However, common words are in fact few because the sentence length is generally short on Web bulletin boards. Therefore, a mostly low cosine similarity index is calculated, and judgement is often erroneous. In the example of Fig 2, it is difficult to judge whether the article question sentence is similar by this cosine similarity index.

In Fig 2, "kisyu," "henkou," and "miniSD" in a Japanese article question sentence are common words for the retrieval query sentence. However, it is difficult to judge similarity from the cosine similarity index in cases where only the term frequency of words is considered.

Therefore, we introduce not only the words but also the structure of the sentence for this similarity cal-

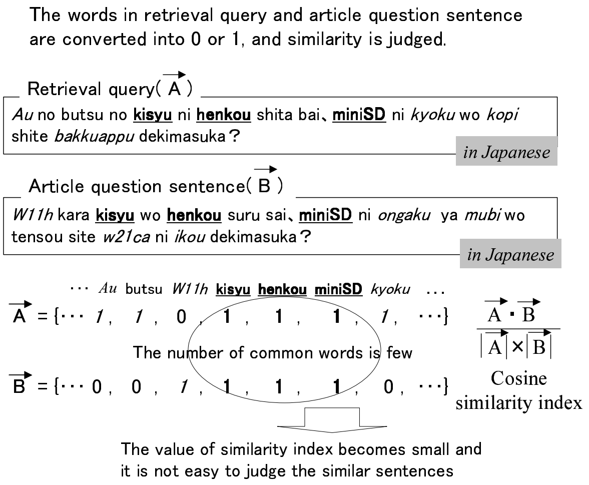
The words in retrieval query and article question sentence are converted into 0 or 1, and similarity is judged.
Retrieval query(A)
Au no butsu no kisyu ni henkou shita bai、miniSD ni kyoku wo kopi shite bakkuappu dekimasuka？
in Japanese
Article question sentence(B)
W11h kara kisyu wo henkou suru sai、miniSD ni ongaku ya mubi wo tensou site w21ca ni ikou dekimasuka？
in Japanese
··· Au butsu W11h kisyu henkou miniSD kyoku ...
A = {··· 1, 1, 0, 1, 1, 1, 1, ···}
The number of common words is few
B = {··· 0, 0, 1, 1, 1, 1, 0, ···}
A・B / |A|×|B|
Cosine similarity index
The value of similarity index becomes small and it is not easy to judge the similar sentences

Figure 2: Example of difficulty in similarity judgment that uses cosine similarity index.

culation. The values of the word vector of the retrieval query sentence and the article question sentence are, consequently, modified, improving similarity judgment by the cosine similarity index.

3 QUESTION ARTICLE RETRIEVAL METHOD BY CO-OCCURRENCE INFORMATION

3.1 Approach for the Problem of Few Matching Words

Here, we propose similarity judgment based on the modified cosine similarity index. Erroneous similarity judgement is caused by few matching words between the retrieval query sentence and article question sentence. The value of the word vector needs to be modified in consideration of structural similarity among sentences.

In similarity judgment of the retrieval query sentence and the article question sentence, it is necessary to judge whether the query sentence is similar to the article question sentence with the same common words. In Fig 3, though the article question sentence has the same common words, article question sentence 1 is similar to the retrieval query sentence, whereas sentence 2 is not similar in Japanese. Since it is difficult to judge similarity by only common words in both sentences, the system pays attention to non-common words.

Figure 4 shows an example of modifying the value of the vector in consideration of the sentence structure. The upper part of figure 4 shows the structural similarity of the sentence that the person has judged. Phrases with similar structure often have the same meaning, so in this study we use the partial structure of sentences composed of pairs comprising common words and non-common words of the retrieval query sentence and the article question sentence. In the lower part of Fig 4, it is judged that the sentence structure “meka” and “kisyu” of the retrieval query sentence is similar to the sentence structure “w11h” and “kisyu” in the article question sentence. The similarity index is improved by modifying the vector element of non-common words included in the structure of the retrieval query sentence to .

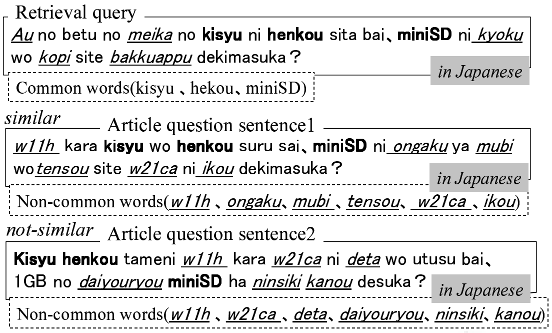
Retrieval query
Au no betu no meika no kisyu ni henkou sita bai、miniSD ni kyoku wo kopi site bakkuappu dekimasuka？
in Japanese
Common words(kisyu 、hekou、miniSD)
similar
Article question sentence1
w11h kara kisyu wo henkou suru sai、miniSD ni ongaku ya mubi wo tensou site w21ca ni ikou dekimasuka？
in Japanese
Non-common words(w11h 、ongaku、mubi 、tensou、 w21ca 、ikou)
not-similar
Article question sentence2
Kisyu henkou tameni w11h kara w21ca ni deta wo utusu bai、1GB no daiyouryou miniSD ha ninsiki kanou desuka？
in Japanese
Non-common words(w11h 、w21ca 、deta、daiyouryou、ninsiki、kanou)

Figure 3: Example of similarity judgment of article question sentence in case common words are same.

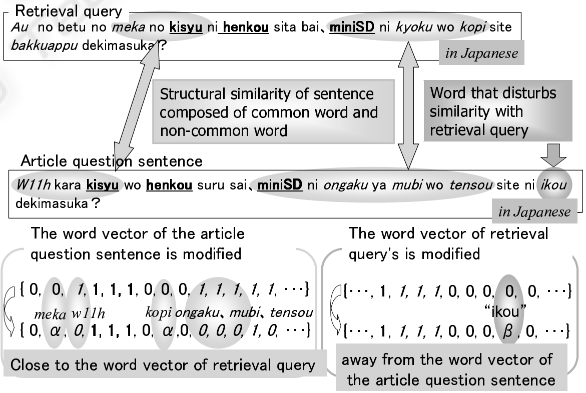
Retrieval query
Au no betu no meka no kisyu ni henkou sita bai、miniSD ni kyoku wo kopi site bakkuappu dekimasuka？
in Japanese
Structural similarity of sentence composed of common word and non-common word
Word that disturbs similarity with retrieval query
Article question sentence
W11h kara kisyu wo henkou suru sai、miniSD ni ongaku ya mubi wo tensou site ni ikou dekimasuka？
in Japanese
The word vector of the article question sentence is modified
{ 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, ···}
meka w11h kopi ongaku、mubi、tensou
{ 0, α, 0, 1, 1, 1, 0, α, 0, 0, 0, 0, 1, 0, ···}
Close to the word vector of retrieval query
The word vector of retrieval query's is modified
{···, 1, 1, 1, 1, 0, 0, 0, 0, 0, ···}
“ikou”
{···, 1, 1, 1, 1, 0, 0, 0, β, 0, ···}
away from the word vector of the article question sentence

Figure 4: Example of modifying word vector based on sentence structure.

Conversely, there are words that cannot apply to word matches or structural similarities of sentences. These words are considered to suppress the similarity of sentences. As shown in Fig 4, the non-common word “ikou” in the article question sentence is not included in the structure of the sentence composed of common words and non-common words. Non-common words in the article question sentence are assumed not to be related to words in the retrieval query sentence. In this case, the similarity index is reduced by modifying the word vector element of the retrieval query sentence to (< 0).

By using the structural feature, some elements of the word vector are modified to augment or suppress the similarity. Then, we have to solve the following problems of modifying the word vector.

- A comparison method for sentences composed of common words and non-common words of the retrieval query sentence and the article question sentence
- A calculation method for the value when the word vector is modified

3.2 Comparison of the Structures of Sentences that Consists of Common Words and Non-Common Words

Here we consider the structures of sentence composed of common words and non-common words. The common and non-common words are all nouns.

The structure of a sentence is decided by using the dependency analysis tool " Cabocha. "[1] The structure of the sentence the person used for the similarity judgment in Fig 4 is decided with a dependency analysis tool. The definition of the structure in the sentence, including common and non-common words is as follows when the dependency analysis is used.

- A clause including common words qualifies a clause with non-common words.
- A clause including non-common words qualifies a clause with common words.

The sample tree structure output, which is the dependency analysis result of the retrieval query sentence, is shown in Fig 5. The structures of the sentence determined from the above-mentioned definition are two of the pairs "meka"and "kisyu," and "miniSD"and "kopi".

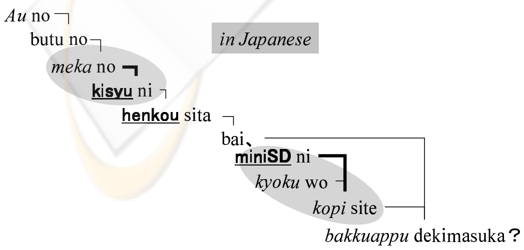

Figure 5: Dependency analysis result of retrieval query.

The structure of the retrieval query sentence is decided as shown above, however, the dependency analysis is not applicable because the style of the article question sentence is not wellformed. Therefore, it is impossible to determine the sentence structure of the article question sentence with a dependency analysis tool.

Common words are considered to be included in the article question sentence. If a word related to non-common words of the retrieval query sentence is included in the article question sentence, it is assumed that the structure of the retrieval query sentence is included in the article question sentence. If so, the structural similarity of the sentences is compared using the relation between words with the co-cooccurrences.The co-occurrences of a word is when the two words exist simaltaneously in the same sentence. The co-occurrence dictionary for the word is automatically produced beforehand from all the article question sentences.

Figure 6 shows a comparison of the structure of the retrieval query sentence and the article question sentence. In this figure, it is noted that "meka," and "kisyu" in the structure of the retrieval query sentence are decided by the dependency analysis, and it is judged that the partial structure of the sentence is similar because non-common the word "w11h" cooccurs with "meka" in the article question sentence. Moreover, two or more words co-occur, such as the non-common words "ongaku," "mubi," and "tensou" in the article question sentence, with "kopi" of the non-common words ub the retrieval query sentence. In this case, the three partial structures "miniSD" and "ongaku," "miniSD" and "mubi," and "miniSD" and "tensou" in the article question are assumed to be similar to "miniSD" and "kopi" in the retrieval query sentence.

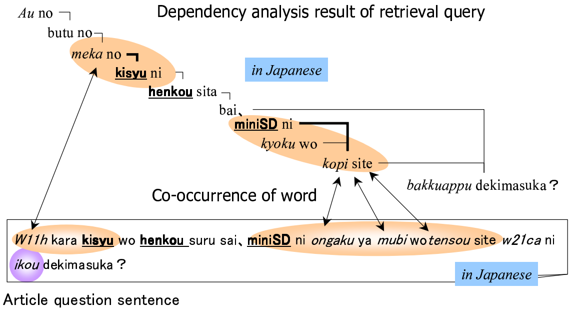
Dependency analysis result of retrieval query
Au no
butu no
meka no
kisyu ni
henkou sita
bai、
miniSD ni
kyoku wo
kopi site
bakkuappu dekimasuka？
in Japanese
Co-occurrence of word
W11h kara kisyu wo henkou suru sai、miniSD ni ongaku ya mubi wo tensou site w21ca ni ikou dekimasuka？
in Japanese
Article question sentence

Figure 6: Comparison of sentence structure including sentence of common words and non-common words by cooccurrence of word.

If non-common words of the article question sentence did not co-occur with all non-common words of the retrieval query sentence in the comparison of the sentence structure, we consider the word that sup-

[1] http://chasen.org/ taku/software/cabocha/

presses the similarity of the retrieval query sentence and the article question sentence.

3.3 Modification of the Word Vector

3.3.1 Modification of the Word Vector by the Structural Similarity of Sentences

The word vector of the article question sentence is modified close to the word vector of the retrieval query sentence when the sentence structure of the retrieval query sentence are similar to the structure of the article question sentence.

In our research, the co-occurrence of a word is used to find a sentence structure that is similar to the structure of the retrieval query sentence in an article question sentence. It is thought that the higher the level of co-occurrence, the more similar the sentence structure. Therefore, the co-occurrence index is used as an index when the value of the vector is modified. Co-occurrence index $C_{i,j}$ of non-common words W_i in the retrieval query sentence and non-common words W_j in the article question sentence are defined by the following formula.

$$\begin{aligned} C_{i,j} &= \frac{S_{i,j}}{S_{all}} \\ (S_{i,j} &= Co-occurrence\ sentences\ number \\ &\quad of\ word\ W_i\ and\ word\ W_j, \\ S_{all} &= All\ sentences\ number) \end{aligned}$$

Figure 6 shows that two or more non-common words in an article question sentence might co-occur with non-common words of the retrieval query sentences. Normalization is necessary to modify the vector when the vector is modified with the word's co-occurrence. Then, value of the word vector is normalized with the maximum 1. The modification expression for the word vector is defined as follows: , where α_i is a modification value. This value is used when the word W_i vector in the retrieval query sentence is corrected. X_i is the average of the co-occurrence index of non-common word W_i in the retrieval query sentence and non-common words in the article question sentence. The bigger the value of X_i, the higher the value of α_i.

$$\begin{aligned} \alpha_i &= \frac{1}{1+e^{0.5-X_i}} \\ X_i &= \frac{\sum_j C_{i,j}}{Q*n} \end{aligned}$$

$$\begin{aligned} Q &: Number\ of\ article\ question\ sentences \\ n &: Number\ of\ non-common\ words\ of \\ &\quad article\ question\ sentences\ that \\ &\quad co-occur\ to\ W_i \end{aligned}$$

3.3.2 Modification of the Word Vector by a Word That Suppresses the Similarity of Sentences

A word that does not co-occur with non-common words of the retrieval query sentence is considered to suppress the similarity of the sentences with respect to non-common words of the article question sentence. The value of the word vector in the retrieval query sentence is modified so that it diverges from the word vector in the article question sentence. The more common words are in the retrieval query sentence and the article question sentence, the more similar both sentences tend to be. Thus, it is thought that the smaller the number of common words is, the higher the possibility that the retrieval query sentence is not similar to the article question sentence. The expression that modifies the word vector of the retrieval query sentence is defined as follows: , where β_k in non-common words of the article question sentence is a modification value of the vector corresponding to word W_k.

$$\beta_k = -\frac{1}{Number\ of\ common\ words}$$

4 EVALUATION EXPERIMENT

To confirm the effectiveness of the proposed method, we retrieved by using the retrieval query sentence 1,162 article question sentences from Web bulletin boards in order. All the retrieval results were sorted using results of the cosine similarity index. The data for correct question article sentences were constructed manually in advance. Since in this paper we aim to improve a user's retrieval accuracy, we performed the evaluation in according to the number of correct question article sentences in the top 10 retrieval results, and the distribution of the top 20 correct question article sentences.

4.1 Comparison of Top 10 Correct Question Article Sentences

Tables 1 below shows the results for the number of correct question article sentences in the top 10 retrieval results. The result was better than the conventional cosine similarity index obtained for all the queries in the top 10 correct question article sentences, thus confirming the effectiveness of this method. We obtained the same result as the conventional cosine similarity index in query (7) because there were few correct question article sentences.

Furthermore, there were few correct question article sentences in the top 10 for the number of correct

Table 1: Number of correct question article sentences @@@@@ in the top 10 retrieval.

Query number	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Correct sentences	10	24	7	10	15	37	2
Proposed method	4	2	3	1	2	5	1
Conventional method	1	0	1	1	1	3	1

question article sentences. It is preferable that correct question article sentences are all ranked highly. This is a problem to solve in the near future.

4.2 Comparison of Emergence Distribution of Correct Question Article Sentences in the Top 20

The emergence distribution of the correct question article sentences for the cosine similarity index in the top 20 is shown in the following figures 7 and 8. The arrows in the figures indicate a change in the correct question article sentences' order in the top 20. The order of all the correct question article sentences under the conventional cosine similarity index has improved in Fig 7, while in Fig 8, the correct question article sentences in the top 20 and under the conventional cosine similarity index move within the top 20. The cosine similarity index thus improves, and the correct question article sentences move to a higher rank. The effectiveness of this method is confirmed from both results. Consequently, we found that the incorrect question article sentences in subordinate positions occasionally exceeded the similarity index of high-ranking correct question article sentences when using the conventional cosine similarity index. Our future work is to more effectively suppress the similarity index in the question sentence of incorrect articles.

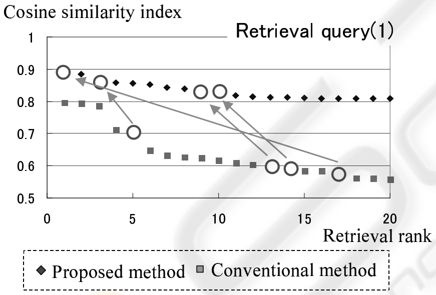

Figure 7: Correct question article sentences' distribution in the top 20 of query (1).

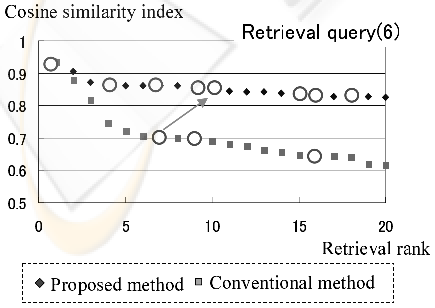

Figure 8: Correct question article sentences' distribution in the top 20 of query (6).

5 CONCLUSION

In this research we proposed a retrieval method to present articles that include similar question sentences to retrieval query sentences input by bulletin board users. We found that in similarity judgments of the article question sentence in the first contribution of the question article and the retrieval query sentence, not only word match but also the structure of the sentence are considered. The value of the word vector of the retrieval query sentence and the article question sentence is modified, making possible a similarity judgment by the cosine similarity index.

The proposed technique was applied to 1162 question sentences from articles on Web bulletin boards with manual query retrieval. The results were evaluated using the number of correct question article sentences in the top 10 and the distribution of correct question article sentences in the top 20 retrieval results sorted by the results of the cosine similarity index. The results confirmed that the proposed method is better than the conventional one for all retrieval query sentences.

REFERENCES

Ohtsuka, T. (2004). *An Evaluation Method of Web Search Engines based on User's Sense*. NTCIR Workshop 4 Meeting Working Notes, Supplement Volume 1 : WEB Task.

Mochihashi, D. (2004). *Learning Nonstructural Distance Metric by Minimum Cluster Distortions*. EMNLP-2004, pp.341-348.

Kishida, K. (1997). *International publication patterns in social sciences: a quantitative analysis of the IBSS file*. Scientometrics Vol.40, No.2, pp.277-298.

Sasaki, Y. (2002). *NTT's QA Systems for NTCIR QAC-1*. working notes, NTCIR Workshop 3, Tokyo.

Tamura, T. (2005). *Classification of Multiple-Sentence Questions*. In Proceedings of the 2nd IJCNLP-05.

Skowron, M. (2005). *Effectiveness of Combined Features for Machine Learning Based Question Classification*. Special Issue on Question Answering and Text Summarization, Journal of Natural Language Processing, Vol.6, pp. 63-83, 2005.

Li, X. (2005). *Learning Question Classifiers*. COLING 2002, pp.556-562, 2002.