Idea-Deriving Information Retrieval System

Tsuneaki Kato
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho,
Soraku-gun, Kyoto,
619-0237, Japan
+81-774-93-5370
kato@cslab.kecl.ntt.co.jp

Shigeo Shimada
NTT Advanced Tech. Corp
Higashitotsuka West Bldg.,
90-6, Kawakami-cho,
Totsuka-ku,Yokohama-shi,
Kanagawa, 244-0805, Japan
+81 45-826-6081
shigeo@totsuka.ntt-at.co.jp

Mutsumi Kumamoto
NTT Advanced Tech. Corp.
Higashitotsuka West Bldg.,
90-6, Kawakami-cho,
Totsuka-ku,Yokohama-shi,
Kanagawa, 244-0805, Japan
+81 45-826-6038
kumamoto@totsuka.ntt-at.co.jp

Kazumitsu Matsuzawa
NTT Service Integration Laboratories
3-9-11, Midori-Cho Musashino-Shi, Tokyo 180-8585 Japan
+81-82-211-5426
matuzawa@magnet.netlab.co.jp

ABSTRACT

This paper presents the information retrieval system that integrates concept-based retrieval, which focuses on the similarity of the meanings of words, with the character-string-matching-based retrieval. The system provides a word association function, a concept retrieval function, and a document classification function, which are expected to help the user to reach the target document quickly or to derive new ideas. This paper describes this system and evaluates the characteristics of various methods used in the system. The evaluation results show that integrating the concept base and character-string matching gives better results than either of them does singly.

Keywords

information retrieval, idea-deriving, word sense, concept base, word association, clustering.

1. INTRODUCTION

The widespread use of electronic documents, the WWW and electronic mail have given rise to the issues of how to retrieve, extract and manage information from a massive text information base. The natural language processing technologies, such as morphological analysis, syntax analysis, and statistical processing of occurrence frequencies of words in natural language texts, can process a large volume of text at high speed. The application of these technologies to solve the above issues is being studied [1,2]. More specifically, the areas of study include how: to construct knowledge of word meanings (concept base) of words expressed in high-dimensional vectors representing the word meaning in machine-readable dictionaries, or the associations between words found by analysis of general documents.

The handling of word meaning has two conflicting purposes: precise information retrieval and a divergent mode of retrieving as exhaustive information as possible. The latter purpose relates to the ambiguous query of the user who needs some assistance to derive related ideas. The proximity of words also has two aspects: the similarity of lexicographic senses of the words as found in dictionaries, and co-occurrences of the words in a specific time and a specific field. The example of the latter is the expected co-occurrences of the words, “soccer “and “Korea”, or “soccer” and “Japan” in sports news at the time of the World Soccer Match in 2002. In this case, a strong association between the words in each pair is represented, although they have no semantic association with each other. The differences between the two aspects of proximity lead to different results of information retrieval. We have developed an Idea-Deriving Information Retrieval System that enables the user to choose between "dictionary-based concept base," which reflects lexicographic senses, and "corpus-based concept base," which reflects association of words. This system’s aim is an extension of "the Associative Information Retrieval System" [4] with improvements reflecting the evaluation of the latter. This system aims at, not only retrieval of precise or related information, but also retrieval of diverging ideas through the cooperation between the user and the system. In other words, the system supports (1) accessing the desired information as quickly as possible, (2) choice between different retrieval methods (concept base, TFxIDF (Term Frequency times Inverse Document Frequency), and character-string matching) or combinations of these, and (3) knowing the subject matters of content through clustering of the retrieval results. This paper presents the Idea-Deriving Information Retrieval System.

2. IDEA-DERIVING INFORMATION RETRIEVAL SYSTEM

In the process of information retrieval, users often find it difficult to tell what it is that they really want to find. In fact, the user may need to exert an extra effort to pinpoint what the query really is. The same is true for expressing it. The Idea-Deriving Information Retrieval System (Figure 1) attempts to provide an environment for retrieving the information that best satisfies the user's intention by using a concept base to provide a word association function, a semantic retrieval function, and a document clustering function, which together assist the user in specifying a set of keywords that best reflect the user's intention, and offer concept-based information retrieval. This environment does not simply follow the usual sequence of information retrieval and the analysis of the result, but assists the user in deriving ideas by presenting information about the documents he/she is dealing with during the process of information retrieval. The ideas derived can be either the information the user intends to find or the diverging information that inspires the user to new ideas. Therefore, not only precision, but also the ability to recall all the information is important. The user who is knowledgeable about information retrieval will also find it useful that the system offers

Figure 1 Idea-deriving Information Retrieval System

the choice of, not only a concept base, but also character-string matching based information retrieval, such as TFxIDF. This section describes the concept base used in the conceptual retrieval, as well as word association and document clustering used in it.

2.1 Concept Bases

The concept base is a collection of words arranged in a high dimensional space. There are two kinds: a "Dictionary-based Concept Base", focusing on the intrinsic sense of each word, and a "Corpus-based Concept Base", focusing on the associative sense of each word.

In the corpus-based concept base, the words that co-occur in close proximity are placed in an orthogonal, high dimensional space after the number of dimensions is reduced. The words that co-occur in a similar manner throughout the document are expected to be placed close to each other in the space. The dictionary-based concept base expresses the co-occurrences of the entry word and the words that appear in the definition of the entry word in a dictionary. In this case, the number of dimensions can be reduced by clustering words in the similarity of their senses [6]. Therefore, the axes of the space are not orthogonal to each other, but the meaning of each axis is clear, which can be used to calculate the similarity of words with emphasis on certain aspects [2].

2.2 Word Association

Concept bases can be used for association of keywords. Namely, by specifying a keyword, words in close proximity to it can be presented. The user can have an overview of the association result, and select new keywords closer to his/her target information from the result and then add them to the previous keywords. The set of words obtained through the repetition of the above is expected to be most likely to point to the direction of the intrinsic query of the user. A good result can be expected by the concept-based retrieval using the set of keywords thus collected. For example, suppose one is interested in "multimedia." Applying word association yields the list of words arranged in the order of proximity shown in Figure 2 (a). If the user decides that he/she is also interested in "content," the word is added and word association is repeated. The result is shown in Figure 2 (b). If the user wants to study video related matters, in particular, in the content, he/she only has to add "graphics" to the keywords. The vector of the set of words obtained in this manner points to the direction of the intended information retrieval.

1: multimedia
2: software
3: LAN
4: video
5: interactive
6: computer
7: digital
8: media
9: communications technology
10: quark
11: character
12: content
13: database

(a) Word association with the keyword "multimedia"

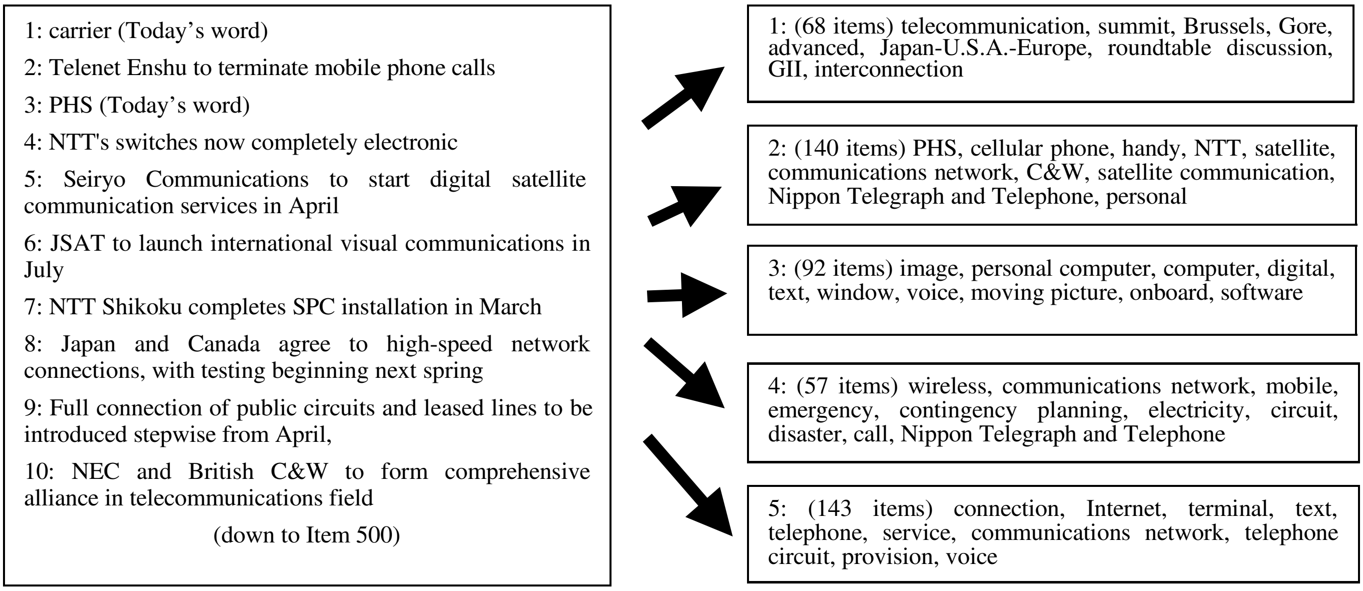
1: multimedia
2: content
3: interactive
4: LAN
5: computer
6: video
7: information
8: software
9: communications technology
10: Internet
11: character
12: graphics
13: transmission of

(b) Word association with the keyword "multimedia" and "contents"

Figure 2. Example of Word Association

2.3 Document Retrieval

The vector (document vector) of the target document is obtained in advance by applying the concept base, and it is placed in the same high dimensional space as that of the concept base. This vector is combined with the vector space model [3] to allow information retrieval based on the proximity between words and documents or between documents. This is different in nature from the information retrieval based on the rarity of word occurrences, such as TFxIDF. The document vector is created, first, by applying morphological analysis to dividing a document into words, and then by making the total of the word vectors obtained with the concept base into a document vector. In general, a word, which constitutes a document, has multiple meanings. However, when the vectors of multiple words in the document are put together, the meaning of each word conforming to the subject matter of the document becomes prominent, and the document vector points to the direction of the subject matter. In the process of information retrieval, the similarity between words or documents and between documents is determined by the cosine coefficient between the centroid vector of the keywords and the document vector, and results are listed in the order of similarity. The system also offers the information retrieval based on character-string matching such as TFxIDF. This is so that the user who is knowledgeable about the characteristics of different retrieval methods can choose the one most suitable for his case. The system also allows the integrated use of different retrieval methods. The characteristics of each method and the effect of integrating different methods are presented in Section 3.



1: carrier (Today's word)

2: Telenet Enshu to terminate mobile phone calls

3: PHS (Today's word)

4: NTT's switches now completely electronic

5: Seiryo Communications to start digital satellite communication services in April

6: JSAT to launch international visual communications in July

7: NTT Shikoku completes SPC installation in March

8: Japan and Canada agree to high-speed network connections, with testing beginning next spring

9: Full connection of public circuits and leased lines to be introduced stepwise from April,

10: NEC and British C&W to form comprehensive alliance in telecommunications field

(down to Item 500)

1: (68 items) telecommunication, summit, Brussels, Gore, advanced, Japan-U.S.A.-Europe, roundtable discussion, GII, interconnection

2: (140 items) PHS, cellular phone, handy, NTT, satellite, communications network, C&W, satellite communication, Nippon Telegraph and Telephone, personal

3: (92 items) image, personal computer, computer, digital, text, window, voice, moving picture, onboard, software

4: (57 items) wireless, communications network, mobile, emergency, contingency planning, electricity, circuit, disaster, call, Nippon Telegraph and Telephone

5: (143 items) connection, Internet, terminal, text, telephone, service, communications network, telephone circuit, provision, voice

The top 500 documents retrieved for "mobile communication" are clustered into five groups

Figure 3 Example of Document Clustering

2.4 Document Clustering

Since the results of information retrieval are listed in the order of similarity, the user can study whether he can find what he wants from the top of the list down. However, the needed information is not necessarily near the top. It is a time-consuming job to study each piece of information, one by one. One may tend to give up if there are more than hundreds of pieces. Clustering saves the user from the chore by clustering those similar to each other. At the same time, the system regards the center of gravity of the cluster as representing the characteristics of the cluster, and presents the association words near the center of gravity as characteristic words of the cluster. This helps the user to have an overview of the documents (Figure 3). Therefore, based on the information provided, the user can narrow the targets down or change the intention of the information retrieval.

3. Characteristics of Various Retrieval Methods

The Idea-Deriving Information Retrieval System provides not only concept-based retrieval, but also character-string-matching-based retrieval, such as TFxIDF. This is because different retrieval methods produce different results, and because we have come to think that one of the best ways to truly reflect one's intention in the information retrieval is to provide the user who is well-versed in information retrieval with various choices. The retrieval characteristics of the concept base and character-string matching tested with the test collection BMIR-J2 [7] (newspaper articles) of the Information Processing Society of Japan are described below. In the case of the concept base, in order to measure the pure characteristics, weighting with TFxIDF is not employed in creating the document vector.

3.1 Characteristics of Corpus-based Concept Base

The use of a word changes over time and in different areas. The corpus-based concept base is established on statistical analysis of what words co-occur with a specific word in actual documents, such as newspapers. Each word can be explained by the words that tend to co-occur. The words that have similar co-occurring words are considered to be similar to each other. Since no user dictionary needs to be created in advance, the corpus-based concept base can deal with the situation in which uncommon words are used in the target documents. Figure 4 shows the case where the corpus-based concept base is effective, in which information has been retrieved from articles on "three major domestic airlines." Since the proper nouns of airlines and related words are automatically registered in the concept base as being similar to each other, the corpus-based concept base produced the best result.

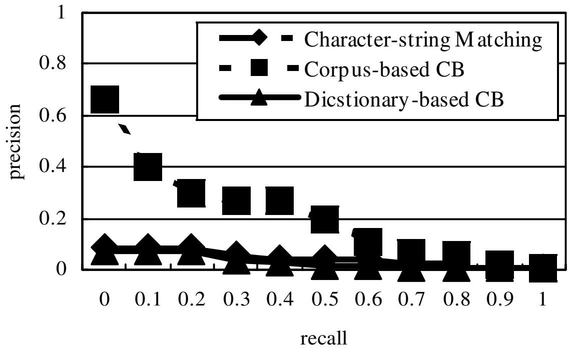


Figure 4. Case Where Corpus-based Concept Base Is Effective

Keyword = "three major domestic airlines": Corpus-based CB can effectively handle proper nouns (name of airlines)

3.2 Characteristics of Dictionary-based Concept Base

Each word has an intrinsic meaning. The dictionary-based concept base stores the intrinsic meaning of each word by using the entry word and its definition sentences in a dictionary. The words that are similar to each other intrinsically are determined to be similar to each other. Figure 5 shows the case where the dictionary-based concept base is effective. This case has to do with retrieving articles on "drinks." Newspaper articles do not often use the word "drinks." Specifically, they mention "beer" or "juice" because it goes without saying that these are "drinks." The dictionary-based concept base is thus effective in dealing with such common words. In contrast, the corpus-based concept base produced "plastic bottles" and "kitchen garbage" as associated words.

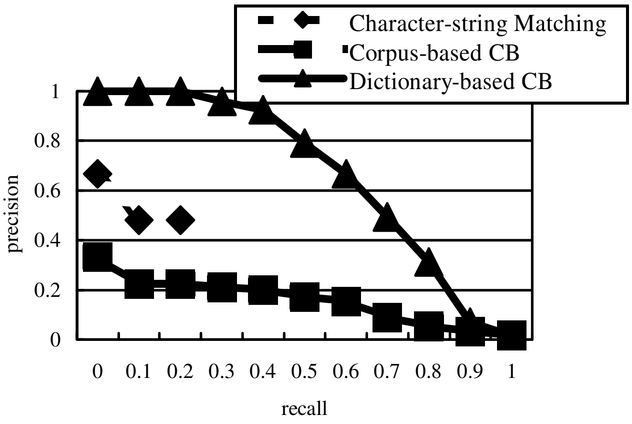


Figure 5. Case Where Dictionary-based CB Is Effective
Keyword = “drinks”: Dictionary-based CB is effective for common words, such as “beer” and “juice.”

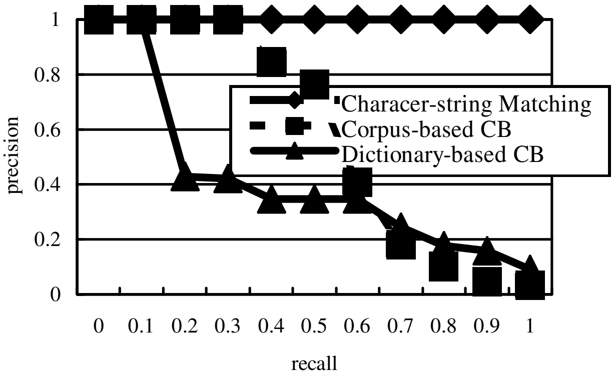


Figure 6. Case Where Character-string Matching Is Effective
Keyword = “agricultural chemicals”: the case where synonyms are not needed.

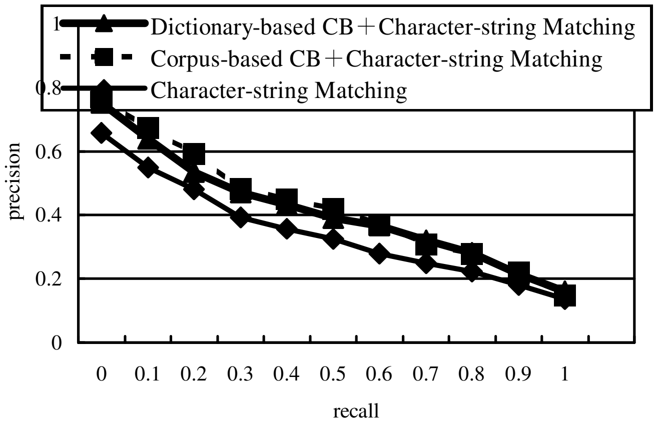


Figure 7. Precision of Integrated Methods

3.3 Characteristics of Character-String Matching

In the case of character-string matching, synonyms obtained from a thesaurus are sometimes included in the keywords. But, the following shows the results for the cases where no synonyms are used. Figure 6 shows the case where the character-string matching is effective. It is the case where the correct answer does include the exact character-string of the keyword. In contrast, the corpus-based concept base gave poor results because it associated "agricultural chemicals" with "agriculture," "medicine," and even "medical care" that derives from the association with "medicine."

3.4 Integration of Various Methods

The above results lead us to think that selective use or integrated use of the various methods can yield the highest retrieval efficiency. Figure 7 shows the case where the similarity is obtained by simply adding the similarity from the corpus-based concept base and that from the character-string matching. Even such a simple combination can give a better result that the individual methods do.

4. Demands by Short Sentences and the Determination of Similarity

The character-string matching is found to be the most powerful in cases where keywords are included in the target documents. However, as the number of keywords increase, it becomes difficult to create a query equation with combinations of *and*s and *or*s. The query equation once created becomes an asset and is difficult to change. In contrast, the concept base is disadvantageous when the number of keywords is small because each word has multiple meanings. As the number of keywords increases, however, the concept base is expected to benefit from a better focus on the subject matter and give a better result. This section presents the evaluation of the cases where the retrieval query includes a considerable number of keywords. The test collection workshop [8] of the National Center for Science Information Systems (NACSIS) aims to build a standard test collection. Since the authors have not completed the evaluation based on the NACSIS test corpus, we report the following evaluation instead. We used newspaper articles, for each of which there is a pair of the headline and the article itself. With the headline as the keywords, the articles are searched. Based on the order in which the correctly matching article appears, the various information retrieval methods are evaluated for the ability to determine similarity. The idea of evaluating the ability to determine similarity from the order of the probability of matching between the headline and the associated article came from the fact that the reader of a newspaper article decides whether to read the article from its headline. This is because the headline is believed to express the characteristics of the article. All the newspaper articles in BMIR-J2 are used. It was confirmed that, while the character-string matching often gives high similarity to the articles totally unrelated to the subject matter simply because they include the keywords, the concept base less often make such serious misjudgments. It was also interesting to find that the combining the concept base and the character-string matching produces a better result. Note that the evaluations in this section are not based on the precision-recall curves as presented in the previous section.

4.1 Evaluation of Individual method

The test collection of BMIR-J2 consisted of 5,080 articles from the Mainichi Newspaper. Document vectors were created using nouns and verbs, which can become the subject or predicate of a sentence. In order to find the characteristics of each of the dictionary-based concept base, corpus-based concept base and TFxIDF (without the range of synonyms expanded), the order in which the correctly matching article appeared for each headline was compared for each combination of two methods. The result is shown in Table 1. In the case of the concept base, in order to measure the pure characteristics, weighting with TFxIDF was not employed in creating the document vector. The comparison between the concept base and TFxIDF showed that, for 43.2% of the articles, the order where the correct article appeared in the case of concept base was higher than in the case of TFxIDF, and lower for 9.5% of articles. For the rest of the articles, i.e. 52.7%, the order of the correct article appearing was the same for both methods. The order of the correct article appearing in the case of TFxIDF was higher than in the case of the corpus-based concept base for 31.1% of the articles, and lower for 14.1%.

Next, individual cases were studied to find what kind of conditions each method is suited for.

[Example 1] The case where the orders of the correct articles of the dictionary-based concept base is higher than those of TFxIDF:

> For the headline, "A fire in a track spreads to nearby houses; A man was killed in the fire - Higashi-Osaka City [Osaka] (Article ID=00085170)," the dictionary-based concept base picked the correct article in the first place, while TFxIDF picked it in the 17th place. The reasons why TFxIDF gave a low order to the correct article may be that the article did not include the words, "fire" or "spread." The dictionary-based concept base, on the other hand, may have interpreted that the words, "burnt down" and "outbreak," in the article were highly similar to the words, "fire" and "spread" in the headline and gave the article the highest order. Incidentally, the first ranking article chosen by TFxIDF was an article whose headline was "Reducing house construction cost by 30%; "House Japan" plan boosted -- MITI by 2000 (Article ID=00751760). The word "house" was given a high TFxIDF value, which may have pulled this article up to the highest order, although the subject matter was different.

[Example 2] The case where the orders of the correct articles of the corpus-based concept base is higher than those of TFxIDF:

> For the headline, "[People's Square (Readers' Column)] Promote nuclear power generation with confidence - office worker, Mr. Tanaka (Article ID=000831250)," the corpus-based concept base picked the correct article in the first place, while TFxIDF picked it in the 63rd place. The reasons why TFxIDF gave a low order to the correct article were that the article did not include the words, "nuclear power generation." The corpus-based concept base, on the other hand, may have interpreted that the words, "atomic power plant," "electric power" and "power station," in the article were highly similar to the words, "nuclear power generation" in the headline and gave the article the highest order. Incidentally, the first ranking article chosen by TFxIDF was an article whose headline was "President, the so-and-so sales company, Tanaka found shot to death at home - Kyoto [Osaka] (Article ID=00267680)." The words "Tanaka" and "company" earned high TFxIDF values, which may have pulled this article up to the highest order although the subject matter was quite different.

[Example 3] The case where the order of the correct articles of TFxIDF is higher than those of the dictionary-based concept base:

> For the headline, "'Discovery' communicates with 'Mir'; a Russian astronaut is aboard [Osaka] (Article ID=00096280)," TFxIDF picked the correct article in the first place, while the dictionary-based concept base picked it in the 108th place. The reasons why the dictionary-based concept base gave a low order to the correct article was that it was made from an ordinary dictionary, and thus did not include such proper nouns as "Discovery," "Mir," and "Russia." But, such drawbacks can be made up for by using a number of different dictionaries.

[Example 4] The case where the order of the correct articles of TFxIDF is higher than those of the corpus-based concept base:

> For the headline, "[People's Square] What lies behind the incident of the fatal shooting a doctor: unemployed so-and-so age 74 (Article ID=000883500)," TFxIDF picked the correct article in the first place, while the corpus-based concept base picked it in the 108th place. Incidentally, the first ranking article chosen by the corpus-based concept base was an article

Table 1. Comparison of the methods in terms of the rank of the correct article

	Dictionary-based Concept Base	Corpus-based Concept Base	TFxIDF
Dictionary-based Concept Base	-	947 (18.6%)*1	484 (9.5%)
Corpus-based Concept Base	1988 (39.1%)	-	718 (14.1%)
TFxIDF	2195 (43.2%)	1581 (31.1%)	-

*1: indicates that the rank of the correct article picked by the dictionary-based concept base is higher than that by the corpus-based concept base in 947 cases (18.6%).

whose headline was "[Special Feature] 1994; the biggest news of the year; spreading of "Gun Contamination"; terrorist attacks against corporations continue (Article ID=01008300)." This article referred to the above shooting incident of a doctor and thus is thought to have earned a high mark in similarity to the incident.

The characteristics of each method described above are summarized in Table 2. The comparison between the dictionary-based concept base and the corpus-based concept base was omitted here because the result of the dictionary-based concept base is dependent on vocabulary and because we did not find other significant characteristics. As was presented above, it is clear that each method has its own unique characteristics. Therefore, it is appropriate to enable the user to make a choice among them, depending on the intention and target of information retrieval. Also, integrating the different methods may give a better solution. It is considered that TFxIDF determines similarity with respect to the importance of words, the dictionary-based concept base with respect to lexicographical similarity, and the corpus-based concept with respect to the association of words. The capability and characteristics of an integrated method will depend on the weight given to each individual method.

4.2 Evaluation of the Integrated Method

This section discusses a method integrating the dictionary-based concept base and TFxIDF, and a method integrating the corpus-based concept base and TFxIDF. Specifically, the similarity of the integrated method is calculated from the similarities of the two individual methods in the following manner:

$$S = a\,S_C + \;bS_T$$

where S_C is the similarity of a concept base (either the dictionary-based concept base or the corpus-based concept base) and S_T is the similarity of TFxIDF. Table 3 shows an example where a=1 and b=1. It was found that when the dictionary-based concept base was integrated with TFxIDF, the retrieval result was as good as that when TFxIDF alone was used. Also, it was found that when the corpus-based concept base was integrated with TFxIDF, the retrieval result was better than that when TFxIDF alone was used. In addition, we studied whether the integrated methods benefited from the integration using the above-mentioned examples.

[Example 1] For the article ID=0085170, the dictionary-based concept base picked the correct article in the first place and TFxIDF in the 17th place, as mentioned before. The integrated method of these picked the correct article in the first place. It was found that the characteristics of the dictionary-based concept base worded effectively and suppressed the TFxIDF's problem of picking articles on entirely different subject matters.

[Example 2] For the article ID=00831250, the corpus-based concept base picked the correct article in the first place and TFxIDF in the 63rd place, as mentioned before. The integrated method of these picked the correct article in the eighth place. The article picked in the first place by the integrated method was an article whose headline was "[People's Plaza] Study 'nuclear power generation' harder; a senior high school correspondence course student age 17 (Article ID=00851660)." It was found that, although the order of the correct article was lower than that of the corpus-based concept base, the characteristics of the corpus-based concept base were still in effect and suppressed the TFxIDF's problem of picking articles on entirely different subject matters.

[Example 3] For the article ID=0096280, the dictionary-based concept base picked the correct article in the 108th place and TFxIDF in the first place, as mentioned before. The integrated method of these picked the correct article in the first place. It was found that the characteristics of TFxIDF were effective and suppressed the dictionary-based concept base's problem of being insensitive to uncommon proper nouns.

[Example 4] For the article ID=00883500, the corpus-based concept base picked the correct article in the 108th place and TFxIDF in the first place, as mentioned before. The integrated method of these picked the correct article in the ninth place. The article picked in the first place by the integrated method was an article whose headline was "[Special Feature] 1994; the biggest news of the year; spreading of "Gun Contamination"; terrorist attacks against corporations continue (Article ID=01008300)," as did the corpus-based concept base alone. It was found that the order of the correct article was higher than

Table 2. Characteristics of Each Method

Methods	Characteristics
Dictionary-based Concept Base	This method is effective even if the document includes no exact keywords, as long as it includes synonyms. This method is effective for common words because of the nature of the dictionary. It is less effective for proper nouns and jargon but can be improved by using appropriate dictionaries.
Corpus-based Concept Base	This method is effective even if the document includes no exact keywords, as long as it includes synonyms or associated words. Since this method collects words from the target documents, it can deal with special words dependent on a specific field. Although it may not pinpoint the exact document as TFxIDF does, it rarely grossly miss the right one.
TFxIDF	This method is effective when the document includes the exact keywords, but may pick wrong documents that happen to include the same keywords. Since this method collects words from the target document, it can deal with special words dependent on a specific field. It will have to depend on thesaurus or some other means to cover synonyms and associated words.

Table 3 Comparison of the methods integrating Concept Base and TFxIDF in terms of the rank of correct article

	Dictionary-based Concept Base + TFxIDF	Corpus-based Concept Base + TFxIDF	TFxIDF
Dictionary-based Concept Base + TFxIDF	-	688 (13.5%)[*1]	678 (13.3%)
Corpus-based Concept Base + TFxIDF	827 (16.3%)	-	789 (15.5%)
TFxIDF	689 (13.5%)	668 (13.1%)	-

*1: indicates that the rank of the correct article picked by the dictionary-based concept base + TFxIDF is higher than that by the corpus-based concept base+ TFxIDF in 688 cases (13.5%).

in the case of the corpus-based concept base alone, indicating that the characteristics of TFxIDF prevailed without discrediting the characteristics of the corpus-based concept base.

The above examples show that the integration of various methods can benefit from the characteristics of each method. The comparisons of various methods for different headlines and articles are shown in Table 3.

5. CONCLUSION

This paper has presented the idea-deriving information retrieval system that views information retrieval as a cooperative work between the user and the system. It explained various functions available to the user: word association, document retrieval based on concept bases or character-string matching, and document classification. The idea-deriving information retrieval system makes information retrieval methods based on concept bases, which store knowledge about the meanings of words, as well as conventional methods available to choose from. The reason why we offer the choices is because the user should be able to use the results of different methods. This paper has also discussed the characteristics of each information retrieval method. Character-string matching is extremely effective in the case where the keywords exist in the target documents, while the concept base is effective even in the cases where the keywords do not exist in the target documents, but where synonyms or associated words exist. It has been found that the dictionary-based concept base is effective for information retrieval with common words. Similarly, it has been found that the corpus-based concept base is effective for information retrieval with proper nouns and jargon, which tend to be diverse and haphazard. Finally, it has been found that character-string matching is effective in the cases where keywords are to the point. It is also expected that, by simply adding the similarities of various methods, the characteristics of each method tend to compensate for the drawbacks of the other. Unfortunately, there is no way to automatically choose the most appropriate method for a particular information retrieval case. On the horizon of our approach lie the technologies for utilizing information and knowledge, such as information extraction, knowledge extraction, and knowledge management. One can envisage a series of processes: information retrieval results being fed to information extraction, and the information extraction result being fed to knowledge extraction. We believe that handling text information with concept bases paves the way for developing these technologies.

6. ACKNOWLEDGMENTS

The idea-deriving information retrieval system was developed based on the evaluation of the associative information retrieval system. The authors would like to thank the Information-technology Promotion Agency, Japan (IPA) for their assistance in the development of the associative information retrieval system (Creative Software Promotion Project).

7. REFERENCES

[1] Schutze, H. Dimensions of Meaning. Proceedings of Supecomputing 92, pp.787-796, 1992.

[2] Kasahara, K., Matsuzawa, K. and Ishikawa, T. A Method for Judgment of Semantic Similarity between Daily-used Words by Using Machine Readable Dictionaries. Transactions of IPSJ, Vol. 38, No. 7, pp.1272-1284, 1997 (in Japanese).

[3] Schutze, H. and Pedersen, J. O. Information Retrieval Based on Word Senses. 4th Annual Symposium on Document Analysis and Information Retrieval, pp.161-176, 1995.

[4] Iida, T., Matsuzawa, K., Ikegagi, T., Ishino, F. and Imai, K. Associative Information Retrieval System, IPSJ SIG Notes, 98-OS-77/98-DPS-87, pp.19-24, 1998 (in Japanese).

[5] Salton,G. and Allen, J. Text Retrieval Using the Vector Processing Model. 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.

[6] Ikehara, S. et al.. Goi-Taikei -- A Japanese Lexicon, the Semantic Attribute System (Vol. 1), Iwanami-shoten, 845p, 1998 (in Japanese).

[7] Kitani, T. et al. BMIR-J2 - A Test Collection for Evaluation of Japanese Information Retrieval Systems, IPSJ SIG Notes, 98-DBS-114-3, pp.15-22, 1998 (in Japanese).

[8] Kando, N. et al. NTCIR : NACSIS Test Collection Project. [Poster] the 20th Annual Collquium of BCS-IRSG, Autrans, France, March 25-27, 1997.