A Character-Based Indexing and Word-Based Ranking Method for Japanese Text Retrieval

Toshikazu Fukushima and Susumu Akamine
Human Media Research Laboratories, NEC Corporation
8916-47, Takayama-Cho, Ikoma, Nara 630-0101, Japan
TEL: +81-743-72-3756, 3760
E-mail: {fuku, akamine}@HML.CL.nec.co.jp

ABSTRACT

This paper describes a Japanese text retrieval system that we applied to the Japanese ad-hoc IR task in the NTCIR Workshop. A character-based indexing and word-based ranking method was implemented on this system. The system generates an index for <title>, <abstract>, and <keyword> parts in documents. It parses only <description> parts in search topics as queries. Its ranking strategy is very simple. It uses the vector space based on short units of Japanese words. It deletes stop words in a query, and calculates the TF*IDF score for each document. Its average precision score for the training set of search topics is 0.36. Experimental results show the effectiveness of using the short units of words and deleting stop words in a query.

Keywords

Japanese text retrieval, character-based indexing, word-based ranking.

1. INTRODUCTION

There are two approaches to Japanese information retrieval. One is character-based, and the other is word-based [1][2]. We have been developing a character-based indexing and word-based ranking method [3][4][5], and implemented it on our system for the NTCIR Workshop.

As Japanese text has no explicit word boundaries, morphological analysis is required for word segmentation. However, word segmentation error in the analysis is not avoidable. This causes failures in retrieving relevant documents. Therefore, we prefer the character-based indexing method to the word-based one.

Ranking methods have been researched and improved, based on the word-based approach. Words are more expressible units for document topics than characters. Therefore, we adopted the word-based ranking method, while we use the character-based indexing. That is to say, the character-based indexing process selects a set of documents which include query keywords, and the word-based ranking process sorts the documents.

2. CHARACTER-BASED INDEXING

A character-based indexing method uses a character or characters (i.e., n-gram) as an indexing key unit [3][6][7]. It stores each key (a character or characters) with a list of all positions where it appears in a document collection. When a query keyword is given, each key unit (a character or characters) in it is compared with the ones in the index, and the position lists for all the key units in the query keyword are extracted from the index. Then, the extracted position values are compared with each other to find the same sequences of the key units as in the query keyword. This process gives a set of documents and position lists where the query keywords appear.

The results of this type of search are the same as the results of a string search like the UNIX grep command; search speed is much faster because the index is used. When using the character-based index, search response time depends mainly on the time taken to extract the position lists from the index. The time for looking at the keys is relatively negligible. Therefore, we developed our own indexing method called CBFI (Character-Based Flexible Indexing) [3] in order to improve the search response time by reducing the extraction time for frequent keys. Our method uses the following five techniques: (1) variable key length depending on character classes, (2) reduced contexts for each key, (3) a sub-index for frequent keys, (4) character position data compression, (5) parameter adaptation to character statistics in sample text. The details of these techniques used to shorten the response time are given elsewhere (see [3]), because the NTCIR Workshop focuses on retrieval accuracy, not on retrieval speed.

3. WORD-BASED RANKING

We adopted a very simple ranking strategy. It uses the vector space based on short units of Japanese words. It deletes stop words in a query, and calculates the TF*IDF score [11] for each document. Keyword expansion techniques are not yet applied.

3.1 TF*IDF Score Calculation

The character-based index search gives the position lists where the query keywords appear. The number of times the query keywords appear in each document and the number of documents in which they appear can be calculated from this information [4]. This means that the word-based TF*IDF score can be calculated even in the case of using the character-based index.

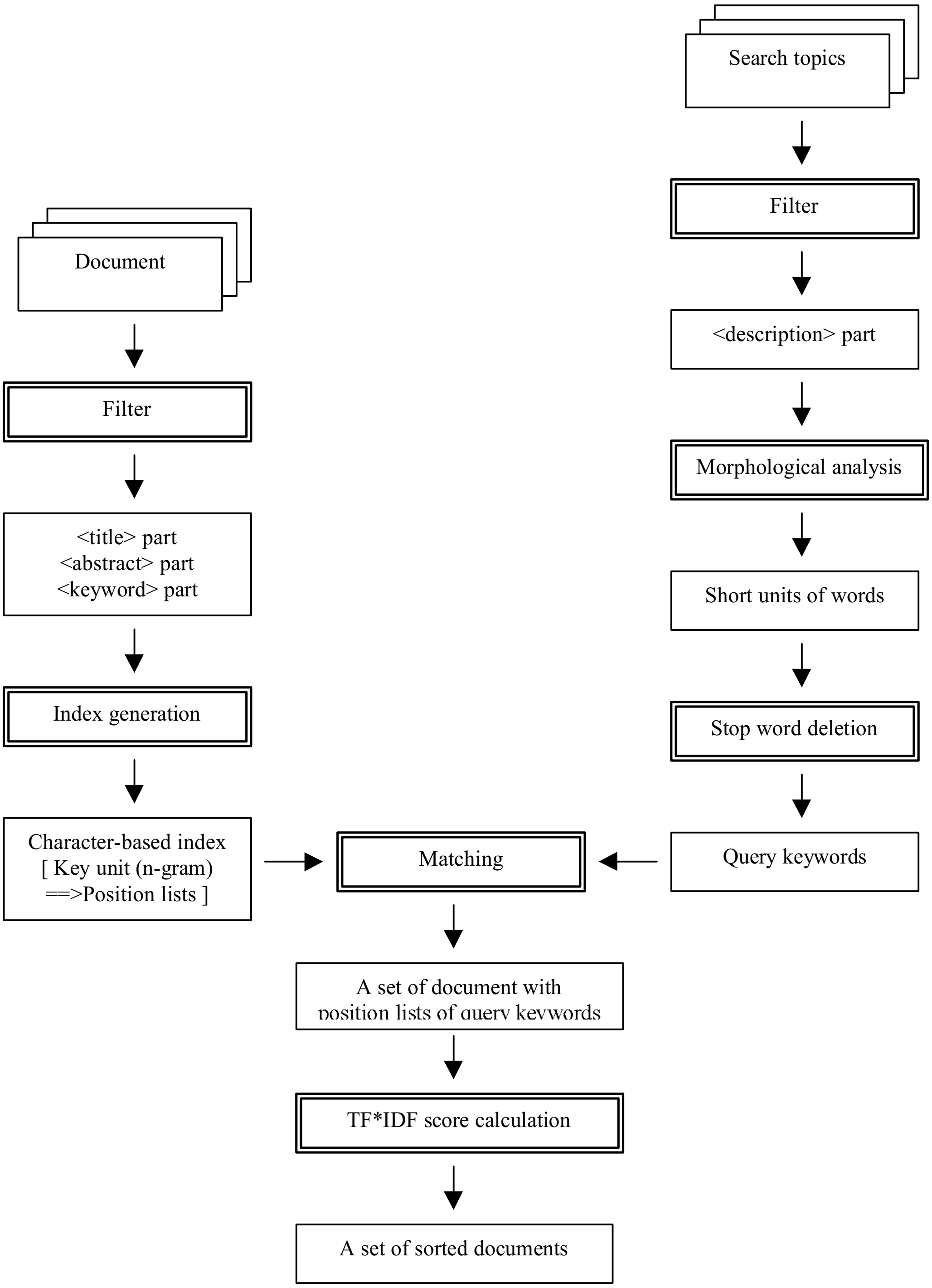

Figure 1. Information Retrieval Process

3.2 Vector Space Based on Short Units of Words

The morphological analyzer separates Japanese text into words. There is the alternative word definition of short units or long units in the analysis. This definition also decides whether the short units or long units are used as the basis of the vector space model. We adopted the short unit definition, because previous experiments [5][10] showed that better IR performance could be obtained in the short unit approach than in the long unit approach.

3.3 Stop Word Deletion

Stop words are deleted from the morphological analysis results of the query text. A part-of-speech filter and a stop word list matching are used to delete them. Our parts-of-speech filter deletes all words except noun classes as stop words. Our stop word list includes 25 words, such as '研究(research)', '論文(paper)', '文献(article)', '方法(way)', '方式(method)', '実現(realization)', '解決(solution)', '提案(proposal)', '記述(description)', '場合(case)', and so on. We consider these words relatively meaningless in the conference paper database.

4. SYSTEM OUTLINE

The information retrieval process shown in Figure 1 is implemented on the system, which we applied to the Japanese ad-hoc IR task in the NTCIR Workshop.

4.1 Indexing Step

The indexing step in our system is as follows.

First, a filter selects <title>, <abstract> and <keyword> parts from the documents. Then, the character-based index is generated for the text of these parts. It stores the position lists for each key unit, which is a character or characters.

4.2 Query Parsing Step

Our system parses only <description> parts in NTCIR search topics as queries, because query length in real life information retrieval is very short. Many people have recently started to use WWW search engines, such as Yahoo!, goo, InfoSeek, AltaVista, and NETPLAZA, to find useful or interesting information on the Internet. We have been engaged in research and development of WWW search engine technologies [8]. It has been reported that the length of queries which search engine users input is very short [9].

The query parsing step in our system is as follows.

First, a filter selects <description> parts from the search topics. Next, a morphological analyzer divides the text of <description> parts into short units of Japanese words, and determines a part-of-speech for each word. Then, stop words are deleted by using the part-of-speech filter and the stop word list as described in 3.3.

This step gives the query keywords. For example, when the NTCIR search topic #0017 "バスへの同時送信に対する排他制御について論じている論文" is given, query keywords 'バス', '同時', '送信', '排他', and '制御' are obtained.

4.3 Matching Step

The matching step in our system is as follows.

Proposed method - using short units of words - using the stop word list	0.36
Case of using long units of words	0.19
Case of not using the stop word list	0.35

Figure 2. Comparison of average precision scores for the training set of search topics.

The position lists where the query keywords appear are obtained by matching between the character-based index and the query keywords. Then, the TF*IDF score is calculated for each document where the keywords appear, as described in 3.1. Finally, the system sorts the documents by the score.

5. EVALUATION

Experimental results for the training set of the search topics are shown in Figure 2. In these experiments, A- or B-judgements are rated as "Relevant". The average precision score is 0.36. The score in the case of using the long units of words, on the other hand, is 0.19. The score in the case of not using the stop word list is 0.35.

These results show the effectiveness of using the short units of words and deleting stop words in a query.

6. CONCLUSION

We have described a Japanese text retrieval system that we applied to the Japanese ad-hoc IR task in the NTCIR Workshop. It utilizes a character-based indexing and word-based ranking method. It uses the vector space based on short units of Japanese words. It deletes stop words in a query, and calculates the TF*IDF score for each document.

Experimental results show the effectiveness of using the short units of words and deleting stop words in a query. In the experiments, an index was generated for <title>, <abstract>, and <keyword> parts in the documents, and only <description> parts in search topics were parsed as queries. The average precision score for the training set of search topics is 0.36. The score in the case of using the long units of words is 0.19, however. The score in the case of not using the stop word list is 0.35.

Such well-known techniques as query expansion and data fusion are not yet implemented on our system. Combining and evaluating them is part of our future works. We are also interested in evaluating WWW search engines as well as traditional IR systems.

7. REFERENCES

[1] Fujii, H., and Croft, W. B., A comparison of indexing techniques for Japanese text retrieval, Proc. of ACM SIGIR'93, pp. 237-246, 1993.

[2] Tokunaga, T., and Iwayama, M., Word-based vs. character-based indexing: an experimental study on Japanese text representation for text categorization, Proc. of IROL'96, pp. 73-78, 1996.

[3] Akamine, S., and Fukushima, T., Flexible string inversion method for high-speed full-text search (in Japanese), Proc. of IPSJ ADBS'96, 1996.

[4] Fukushima, T., and Akamine, S., Development and evaluation of a Japanese full-text retrieval system (in Japanese), Proc. of 3rd Annual Meeting of the Association for Natural Language Processing, pp. 361-364, 1997.

[5] Akamine, S., Fukushima, T., and Seiko, Y., An evaluation of n-gram based ranking method for Japanese text retrieval (in Japanese), Proc. of 56th National Convention of IPSJ, 1998.

[6] Kikuchi, C., A fast full-text search method for Japanese text database (in Japanese), Trans. of Institute of EICE, Vol. J75-D-I, No. 9, pp. 836-846, 1992.

[7] Ogawa, Y., and Iwasaki, M., A new character-based indexing method using frequency data for Japanese documents, Proc. of ACM SIGIR'95, pp. 121-129, 1995.

[8] Fukushima, T., Matsuda, K., and Takano, H., A study of page ranking factors for WWW search engines (in Japanese), Proc. of the 7th Annual Conference of Japan Society of Information and Knowledge, pp. 77-80, 1999.

[9] Jansen, M. B. J., Spink, A., Bateman, J., and Saracevic, T., Real life information retrieval: a study of user queries on the web, SIGIR Forum, Vol. 32, No. 1, pp. 5-17, 1998.

[10] Ogawa, Y., Bessho, A., Iwasaki, M., Nishimura, M., and Hirose, M., A text database system based on simple words (in Japanese), IPSJ SIG Report, DBS-90-6, 1992.

[11] Salton, G., and McGill, M. J., Introduction to Modern Information Retrieval, McGraw-Hill, 1983.