IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITP4545

Incremental Indexing and Its Evaluation for Full Text Search

Hiroshi Yamamoto, Seishiro Ohmi, Hiroshi Tsuji
Hitachi, Ltd.
3-8,Kitakyuhozi-cho 3-chome, Chuo-ku, Osaka, 541-0057 Japan
Tel: 81-6-6281-8331(Direct), Fax: 81-6-6281-8390
yamamohi@itg.hitachi.co.jp

1. INTRODUCTION

N-gram indexing method is the most popular algorithm for the Japanese full text search system where each index consists of serial N characters [1][2]. N-gram based indices can be made in the system. For the English full text search system, indices are based on a word that consists of N-gram (N characters). For the Japanese full text search system, indices are not based on a word but a gram (a character) [3][4][6][7]. In general, the system has 2-gram index in order to save the volumes of index file while there are many words that consists of more than three serial characters and some serial two characters are meaningless from the view of search terms[3][8]. In short, 2-gram can be uniformly used on indices are extracted from the target document for full text search. The advantage of N-gram indexing method is to avoid false drops in the full text search system because indices are uniformly based on 2-grams that are extracted from target documents. On the other hand, the disadvantage is less efficient of searching because the index that can be often used in searching is created with the same method as the index that cannot be often used. In short, the index that can be often used in searching equally based on 2-grams the same as the index that cannot be often used in searching.

In order to improve the performance of 2-gram based test search system, this paper presents supplemental indexing algorithm, called **incremental word indexing method**. Basic idea under this research is that words used frequently in search terms should be indexed.

With incremental word indexing method, indices that are based on words used frequently in search terms should be added to uniformly 2-gram based indices. So this method can maintain the advantage to avoid false drops. This method can improve the performance searching with using supplemental indices that consist of **words**, without using uniformly 2-gram based indices. Consequently if we can specify the word used in search terms, the performance of searching can be improved efficiently.

The summery of the **incremental word indexing method** is following. About the word that is frequently used in search terms, indices should be based on a word or n-gram (n is more than two). The word that is frequently used in search terms means 'the word that is frequently used in search condition'. With using the above-mentioned supplemental indices those length are shorter than 2-gram based indices, we can improve the performance for searching.

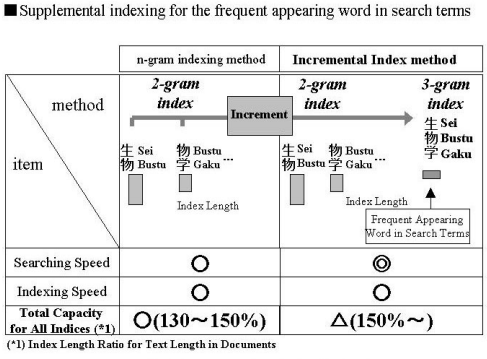

Fig.1 Incremental Indexing Method

Fig.1 shows the outline **incremental word indexing method**. When we search with the Japanese word "Sei-Butsu-Gaku" in case of N-gram indexing method, the both of the 2-gram "Sei-Butsu" based index and the 2-gram "Butsu-Gaku" based index must be referred. On the other hand, in case of **incremental word indexing method**, the only word "Sei-Butsu-Gaku" based index must be used, if the word "Sei-Butsu-Gaku" is defined to be frequently used in search terms. The length of the word "Sei-Butsu-Gaku" based index is much shorter than the amount both of the length of the 2-gram "Sei-Butsu" based index and the 2-gram "Butsu-Gaku" based index. So the searching performance with incremental word indexing method is better than N-gram indexing method based on 2-gram. But generally the total capacity of the indices with incremental word indexing method is larger than with N-gram indexing method based on 2-gram, because the indices based on the word that is frequently used in the search terms should be added to 2-gram based indices with incremental word indexing method.

Then there is a problem: what kinds of words should be indexed and how does it improve the performance? This paper shows the experimental simulation for the variety of retrieval patterns and the guideline for system optimization.

2. TARGET OF INCREMENTAL INDEXING EVALUATION

The searching method is one of the most important factors of the document management system and the knowledge management system. Moreover the searching methods for the above-mentioned system need to be efficient. Under these circumstances, it is important to obtain how to use the incremental indexing system appropriately.

First of all we know there is tradeoff: if the many words are incrementally indexed, the performance becomes better but the index file becomes larger. Second we know Zipf"s law: if the distinct words in some sample texts are arranged in decreasing frequency order and rank orders are assigned, then the frequency of occurrence of the r-th words in frequency order multiplied by rank r is approximately constant.

These imply that there are an appropriate number of words for the incremental word indexing. Our target is to clarify the guideline to optimize followings for the Japanese full text search with incremental word indices.

(a) The relationship between the pattern for the frequency of occurrence in search terms and the pattern for the incremental word index
(b) The performance (searching time) and system resources (memory and capacity of indices) on each condition of (a)

Let us estimate the time (T) for searching and the capacity (C) of indices with the following two values as parameters for our evaluation:

(i) The number (M) of words for incremental index those are added to the basic 2-gram index

(ii) The appearing ratio of the word in the search terms in the search transaction with incremental word indices

Our experimental simulation uses the newspaper articles for one year. It is expected that the analysis for the simulation result shows the number and the kind of the word used with incremental indexing in a suitable direction. It is also expected that the sensitive analysis for the parameters suggest the suitable guideline to estimate the search execution time and the index capacity.

For example, with using results of this analysis, it can be expected to take the following approach.

- In case that the appearing ratio of the word in the search terms is XX in the searching system, and if the average time for searching is YY as the needed condition, it is clarified that the total capacity of indices need to be AA and the pattern for the incremental word index need to be BB.
- In case that the appearing ratio of the word in the search terms is XX in the searching system, and if the pattern for the incremental word index is WW and the total capacity of indices is ZZ, it is clarified that the average time for searching must be CC.

3. EVALUATION FOR INCREMENTAL WORDS INDEXING METHOD

Our simulation for above-mentioned our target are followings.

(1) The environmental conditions for our simulation are followings.
  (a) Search Engine: Windows2000 platform Bibliotheca21 (made by Hitachi),
  (b) Database: One hundred thousands articles (About 200MB),
  (c) Machine Environment:
  OS: Windows2000 Professional /CPU: 250MHz/ Real Memory: 256MB,
  (d) Search Execution: ten thousands executions per index pattern per search pattern as follows with searching trace log for the analysis,
  (e) Indices: Not incremental word index; incremental 3-gram based index; about the word used frequently in search terms, incremental index is based on 3-gram; for example, when we search with the Japanese word “Kei-Zai-Haku-Syo”, incremental indices for searching are based on 3-gram “Kei-Zai-Haku” and “Zai-Haku-Syo”,

(2) Words for incremental indexing are extracted as follows:
  (a) Extract all words that consist of more than three serial characters from database,
  (b) Select M words from all N words extracted. (M=0, 100, 500, 1000).
  (c) On this simulation, we select M (M=1000) words about an “economy” as frequent appeared in search terms. Incremental 3-gram based indices correspond to M (M=0,100,500,1000) words can be added to basic 2-gram indices. Table 1 shows all the pattern of indices on this simulation. (Each index consists of incremental 3-gram based indices and basic 2-gram indices.

Table.1 All the pattern of indices

Index Pattern	Structure for Full-text Search Indices	
	Basic Index (2gram)	Incremental Index (3gram)
Index Pattern 0	2gram Index for 100,000 paper articles	None
Index Pattern 1	2gram Index for 100,000 paper articles	3gram index for frequency of top 100
Index Pattern 2	2gram Index for 100,000 paper articles	3gram index for frequency of top 500
Index Pattern 3	2gram Index for 100,000 paper articles	3gram index for frequency of top 1000

Table.2 Search patterns

Case (Search Patterns)	Frequency of top 1000 words (%)			
	Top 100	101～500	501～1000	Others
Case 1	100	0	0	0
Case 2	50	50	0	0
Case 3	34	33	33	0
Case 4	70	20	10	0

(3) Simulation cases are shown in Table 2.

On each case in Table 2, with the index pattern shown in Table 1, the total capacity of all indices and the average, maximum and dispersion of execution time with search execution for 100,000 paper articles can be obtained.

(4) We will analyze the tendency for the search execution time on following conditions.

This simulation shows the guideline and condition for the high performance searching depends on the value for indices capacity and the searching execution time for each search pattern.

(i) On condition that each pattern of indices in Table.1 are used for searching;
  If the number (M) of words that is defined to be frequently appeared in search terms can be 0,100,500 or 1000 (M=0,100,500or1000), incremental 3-gram based indices are added to be created depend upon the number (M) of words appeared frequently in search terms.
(ii) On condition that each search pattern in Table 2,
  On condition that the appearing ratio of the word in the search terms can be changed, the average searching execution time is depend on each condition.

The above-mentioned simulation is on going, the result of analysis can be obtained until the middle of February 2003.

4. EXTENDED ANALYSIS FOR INCREMENTAL INDEXING

The followings are complemented experimentation plan:

(1) Extension for search patterns and indices patterns,
(2) Depend upon the tendency of the result of this simulation, We increase the number of the frequent appearing word, and refine the kind of the word in search terms,
(3) Document database in other domain such as patents and technical documents,
(4) Relationship between term frequency and document frequency,
(5) The number of the target document for searching.

5. CONCLUSION

On search performance, the incremental word indexing supplements simple n-gram indexing for Japanese full text search system. As side effect, the supplemental algorithm adds the volume of index file. Trend analysis in our research will show the guideline for full text search system design.

REFERENCES

[1] Shannon, Claude E., “Prediction and Entropy of Printed English”, Bell Systems Technical Journal, 30, 50-64. (1950)

[2] Marc Damashek. Gauging Similarity with N-Grams: Language-

Independent Categorization of Text. Science, Vol.267, pp.843-848, 10 February 1995.

[3] Sugaya, N et al., A full-text search system for large Japanese text basis using n-gram indexing method, *proc. 53rd Annual Convention IPS Japan*, 5T-2, 3(1996)

[4] Hosono, K., Current State of Research and Development on Digital Libraries in Japan, 2nd IFLA General Conference - Conference Proceedings - August 25-31, 1996

[5] Sato, T., *et al*., NTCIR-2 Experiments Using Long Gram Based Indices, Osaka Kyoiku University

[6] Matsui, K., Namba, I., Igata, N., Hi-speed Fulltext Search Engine, IPSJ SIGNotes Contents Digital Document No.007,1997

[7] Ogawa, Y. and Iwasaki, M., A new character-based indexing method using frequency data for Japanese documents, In Proc. 18th ACM SIGIR Conf., pp. 121—129 (1995).

[8] Kawashimo, S., et al. Development of full text search system Bibliotheca/TS (in japanese). In Proc. of 45th JIPS Conf. (3), pp. 241-242, 1992.