A New Full-Text Indexing Model with Low Space Overhead for Chinese Text Retrieval

Shuigeng Zhou¹, Jihong Guan²
¹ Department of Computer Science and Engineering, Fudan University, Shanghai, 200433, China
² School of Computer, Wuhan University, Wuhan, 430072, China

Abstract. Text retrieval systems require an index to allow efficient retrieval of documents at the cost of some storage overhead. This paper proposes a novel full-text indexing model for Chinese text retrieval based on the concept of adjacency matrix of directed graph. Using this indexing model, on one hand, retrieval systems need to keep only the indexing data, instead of the indexing data and the original text data as the traditional retrieval systems always do. On the other hand, occurrences of index term are identified by labels of the so-called s-strings where the index term appears, rather than by its positions as in traditional indexing models. Consequently, system space cost as a whole can be reduced drastically while retrieval efficiency is maintained satisfactory. Experiments over several real-world Chinese text collections are carried out to demonstrate the effectiveness and efficiency of this model. In addition to Chinese, the proposed indexing model also works as an efficient retrieval of other Oriental languages, such as Japanese and Korean. It is especially useful for digital library applications where storage resource is very limited (e.g., e-books and CD-based text retrieval systems).

Key words: Digital Library - Information retrieval - full-text indexing - adjacency matrix - Chinese

1 Introduction

With the rapid growth of electronic Chinese documents published in Mainland China, Hong Kong SAR of China, Taiwan area and Singapore etc., there is an increasing need for Chinese text retrieval systems that support fast access to large amount of text documents. Full text retrieval systems are a popular way of providing support for on-line text access, in which word or string search is a very important operation [1,2,4]. These systems' main advantage is that they avoid the complicated and expensive process of semantic indexing. From the end-user point of view, full text searching of on-line documents is appealing because a valid query is just any word or sentence of the document. Generally, full-text retrieval systems have an index to allow fast access to documents, especially for those that manage massive text information. Many text retrieval index models have been developed and used, including inverted lists [2-4], signature files [5], PAT trees [6] and PAT arrays [7,8] and so on.

Based on the granularity of indexed terms, the indexing models can be roughly classified into character-based model, word-based model and phrase(or concept)-based model. Concept-based indexing can provide high-level, semantic description of text documents, but it need the help of natural language processing to extract meaningful concepts from text, which is not an easy job to do. So more often it is used as a kind of complementary to the character/word-based indexing methods.

Although word-based indexing is widely used in English text retrieval [2-4,19,21,22], it can not be easily applied to Chinese. This is because written Chinese text — a string of Chinese characters and punctuation marks — has no delimiter to mark word boundaries. The first step toward word-based indexing of Chinese text is to break a sequence of characters into meaningful words, which is called word segmentation. Word segmentation is known to be a difficult task because accurate segmentation of written Chinese text may require deep analysis of the sentences [9,20]. On the other hand, character-based indexing methods don't depend on word segmentation, thus it is suitable and easy to be implemented for Chinese text retrieval.
In the last decade, there had been an increasing research on character-based indexing for text retrieval of Chinese and other Oriental languages [10–17]. However, traditional character-based full-text indexing methods cost too much storage space because each character in text database is indexed and its positional information is also stored to support exact searching. So it is unfavorable for some application areas where storage resource is very limited, e.g., e-books, CD-based mini-digital libraries and text retrieval systems.

Recently, we have proposed an approach to reconstruct the traditional full-text indexing models (including inverted lists and PAT array) for improving text retrieval efficiency by using adjacency matrix structure [18]. In this paper, we present a new adjacency matrix based full-text indexing model to reduce indexing space overhead of Chinese text database. By treating a text database as a directed graph and extending the concept of adjacency matrix of directed graph, we develop the adjacency-matrix based full-text indexing model. It is a bigram-based indexing technique because each element in the adjacency matrix corresponds to a bigram, or an adjacent-character pair (in this paper, we use these two terms equivalently). With this new model, retrieval system consumes considerably less storage space than using other indexing models because

- the system stores only indexing data after it is built.
- Original text documents can be reconstructed efficiently by using the saved indexing data;
- each occurrence of a bigram is uniquely identified by the label of a certain s-string (i.e., a string with no character occurring twice or a string consisting of only two similar characters) where the bigram occurs, rather than by its position as traditional indexing techniques do. Considering that the total number of s-strings included in a text database is much smaller than the text database’s length, indexing space overhead can be cut down considerably.

The only precondition for the proposed model is that sufficient main memory is available to support an in-memory adjacency matrix. Given this precondition, the model we describe can support efficiently searching of string in large text databases.

Because of its advantage of low space overhead, the proposed model is especially applicable for digital library application areas where storage resource is very limited, such as e-books, CD-based mini-digital libraries and PAD-based text retrieval systems. Although the cost of storage has been reducing, the users' requirement for larger amount of storage has never abated, due to the explosive growth of information and the rapidly increasing of storage amount consumed by system and application softwares. For information retrieval systems, storage requirement is still an obstacle to their wide employment.

Currently, with the popularity of mobile phones, more and more computing applications previously installed on desk PCs are now moving onto people's palms. It is not unrealistic to expect that in the near future people can enjoy e-books in their mobile phones. However, considering the smaller and smaller size of current mobile phones, their limited capability of storage unavoidably restricts the number of ebooks that the users can enjoy on them. For such a potential application scenario, our proposed new model is obviously a promising solution, as a compact indexing engine of mobile phone based ebooks or digital libraries.

Although the initial motivation of the new indexing model is Chinese text retrieval, it is also effective and efficient for text retrieval of other Oriental languages, such as Japanese and Korean. Furthermore, it can be applied into retrieval of any string-like data, such as gene sequences and time series.

In this paper, our focus is on the new model and how to use the model to create indexing and conduct query-processing given an arbitrary text database. We leave some deep topics such as dynamic updating of indices, query results ranking for future work. Major contributions of this paper include the following aspects:

1. A new full-text indexing model with low space overhead based on adjacency matrix of directed graph is proposed for Chinese text retrieval;
2. Implementation techniques of the new model is presented;
3. Space overhead of the new model is analyzed formally and compared with three traditional indexing methods;
4. Experiments over five real-world Chinese text databases are conducted to demonstrate the effectiveness and efficiency of the new model.

The rest of this paper is organized as follows. In Section 2, we describe the novel full-text indexing model. In Section 3 we introduce the implementation techniques in details. We analyze storage overhead of the novel model and compare it with other indexing techniques in Section 4. In Section 5, we present the experimental results. And in Section 6, we give a brief overview of related work. We conclude the paper in Section 7.

2 New Full Text Indexing Model with Low Space Overhead

We begin with the notation of a Chinese character set $\Sigma$: a finite set of Chinese characters, letters, digits, punctuation marks and other symbols that may occur in Chinese text documents. A text string or simply string over $\Sigma$ is a finite sequence of characters from $\Sigma$. The length of a string is its length as a sequence. We denote the length of a string $w$ by $|w|$. Alternatively a string $w$ can be considered as a function $w : \{1, \ldots, |w|\} \rightarrow \Sigma$: the value of $w(j)$, where $1 \leq j \leq |w|$, is the character in the $j$th position of $w$. To distinguish identical characters at different