Design of a full text image Chinese CD-ROM database retrieval system in a campus digital library

Ruey-Shun Chen and C.C. Chen

The authors

Ruey-Shun Chen is an Associate Professor at the Institute of Information Management, National Chiao-Tung University, Taiwan and **C.C. Chen** is a graduate student, NCNU, Taiwan.

Keywords

Digital libraries, CD-ROM, Databases, China

Abstract

The library CD-ROM database with its enormous storage and retrieving from network capabilities, has been gradually replacing some of its printed counterparts. But one of the disadvantages is that it can only access the full-text image CD-ROM database via local area network environment. This paper proposes a new method to solve this problem. The method uses hash function as data structure and image document process technique to perform a practical library full-text image CD-ROM database retrieving system. The results of its function can reduce users storage space, allow multi-users to retrieve the same CD-ROM database, and allow users to retrieve and print out full text image CD-ROM database via campus network.

Electronic access

The research register for this journal is available at
http://www.emeraldinsight.com/researchregisters

The current issue and full text archive of this journal is available at
http://www.emeraldinsight.com/0264-0473.htm

The Electronic Library
Volume 20 · Number 6 · 2002 · pp. 460–465
 · ISSN 0264-0473
DOI 10.1108/02640470210453912

Introduction

In recent years, traditional archives management systems have been replaced by computer media since the mass of data can be restored in the form of electronic media, eg CD-ROMs (Yu and Chen, 2001). Because of its large capacity CD-ROMs can reduce the storage space required and extend the useful life of data. The main advantages of CD-ROMs include:

- *small size*: it is easy to use and to distribute;
- *high storage volume*: the capacity of each CD-ROM is 600Mbytes;
- *easy to store*: the longevity of CD-ROM is at least 20 years;
- *easy retrieval*: the average seek time of current CD-ROM systems is 600ms – slower than hard disks, but faster than magnetic tape or microfilm; and
- *low cost*: CD-ROM drivers are now standard on personal computers.

The data on CD-ROMs can also be retrieved from local area networks (LANs) and can be easily modified. Despite the fact that UMI began to publish CD-ROM products with the abstracts of doctoral dissertations as long ago as 1987 – products which are excellent research tools for both professors and graduate students – there has hitherto been no possibility to access the full-text images of Chinese dissertations on CD-ROM databases via the university campus network.

Current CD-ROM hardware technology is very mature, and it is the time for CD-ROM software to be popularized, practiced and produced. In 1991, the National Library of Taiwan had developed a CD-ROM with 157,360 records of data for indexing the 1,161 kinds of journals and the dissertations in Taiwan. But when the system is used, users can only get the titles and not the full-text Chinese CD-ROM database – and it is difficult to determine the content of an article from just the title. A full-text retrieval system was developed in 1992 and the first version of this system contains the dissertations of doctoral and master students in the National Chiao-Tung University of Taiwan. The users can access and print out the full-text of the dissertations via the campus network

Now, the design and implementation of this system is fully completed. All of the functions for data retrieving and full text reading are ready. Regarding data retrieval in Chinese,

The Electronic Library
Volume 20 · Number 6 · 2002 · 460–465

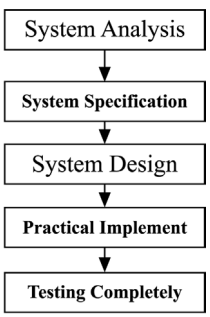
the key value can be a Chinese character or a composition of a Chinese character string. Because there are unavoidable English technical terms in the dissertations, we had to consider this problem with English data. The solution was to allow the key value to be combined using any English words or Chinese character strings in the system. The users can access any information about the key first, and then the full text can be displayed on the screen clearly to permit the users to read the full text directly and get what they want. Furthermore, the image display on the screen can also be printed out (and charged) from the client's laser printer. In this way, the users can avoid the time taken to go to the library and print out the paper there. Users can access this database system from the campus network at any time.

Due to the maturity of the notebook computer, the speed of workstations and the successful techniques of image processing, scanning images and storing them on computer is no longer difficult. The algorithm for data compression can reach a compression ratio of 50 percent at least – thus permitting more image data to be stored. In this paper, we discuss the algorithms used in this system, the data format, the hardware configuration, and the features and the functions of our full-text image database system.

System design and programming methods

The software design steps are shown in Figure 1. An overview of the whole system, i.e. a complete analysis of the five steps of this Chinese-English retrieval system, is presented together with efficient and useful solutions for each particular problem encountered.

Figure 1 System development flowchart



Requirement analysis

In Figure 2, we present the system structure from the viewpoint of the user and the data format of the programming method (Feldman, 1999). The main functions are data retrieval and full text reading which are discussed below.

Data retrieval

The aim of the data retrieval are the actual abstracts of the dissertations. The possible indexes used follow Juha (1996):

- date;
- department;
- author;
- teacher;
- the keywords of title; and
- the keywords of abstract.

In the above six fields, the first two are indexes with a numeric key. The others are indexed with both Chinese and English keys, for example, Chinese name and English name, Chinese keywords and English keywords.

Full-text reading

For proper full-text reading, we must display the content of the dissertation on the computer screen with suitable resolution. Users may page down or page up to see the other parts of the dissertation. The content of the dissertation must be printed out by the laser printer via campus network.

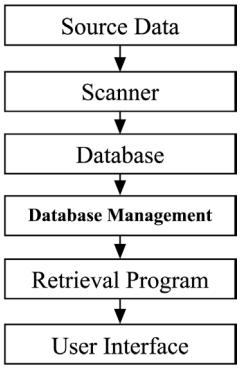
User interface

In order to be useful a system must have a user-friendly interface. This implies that when it is operating, there must be sufficient information displayed on the screen to let users easily find the data they want.

System specification

According to the system requirement analysis in the above section, we can design the

Figure 2 System block diagram



The Electronic Library
Volume 20 · Number 6 · 2002 · 460–465

corresponding system specification to achieve the system desired.

Data retrieval
There are three kinds of keys:
(1) Chinese words;
(2) English words; and
(3) numeric data.

The minimum unit of the key for Chinese words is a Chinese character. The minimum unit of the key for English words is an actual English word, for example, "liquid" and "structure". Not only can a single word be employed as the key to find out the matching data, but the system also permits the utilization of the Boolean "and" and "or" functions to combine the results. The Chinese character can also be used as a master key to find out the matching data.

Full-text reading
Before being stored, the image data must be compressed to 50 percent. Experience shows that the compression ratio should not be too low as this will waste storage space. Furthermore, the decompression ratio may not too slow as this will inhibit full-text reading. So, in order to get the best performance, we must have a trade-off between the above two factors. Regarding printing out speed, considering the output quality and the popularization, we selected the high performance HP Laser Jet as the output device.

User interface
We found that having menu-driven actions was the best way of user interfacing. It is easy to understand and learn from a menu. Since the Internet is popular, we upgraded the latest version of the software with a Web-based version. Users can now access and print out papers with an authorized person via the Internet.

Data format

In the system, the data formats used are as follows:

- *hash table* – for English words;
- *index table* – for all types of data to be indexed;
- *database* – stores the information of papers; and
- *image data* – the real content of papers.

These are discussed in more detail below.

Hash table
The two issues about the hash table are:

(1) *Hash function*. In order to calculate the hash key for an English word, we multiply each ASCII code of the English string. It is then easy to calculate.
(2) *The policy for collision*. Here, we use the dynamic linking technique to link all data of the same hash keys. Therefore, the format for each element in the hash table is A1, A2, A3, where A1 is the next entry pointer → for dynamic linking; A2 is the index pointer value; and A3 is an English word.

Although the hash table can be constructed easily by a collection of the elements for the table, we cannot actually do it this way. There will be lots of hash tables because there are many fields that need such tables: for example, the name of the author, name of professor, the title, the abstract, and then there will be four files to open when we search for data, because each hash table occupies one file. Therefore, it is possible to run out of the file pointers. So, we have to connect different hash tables into one large hash table, so that there will then be only one file for that table to open. Opening one file is enough for searching the contents of four hash tables.

The file format of the large hash table is:

A1 A2 A3 ... An An+1
Ai: basic elements for each hash table $1 \leq i \leq n$
An+1: for dynamic linking used by every hash table.

Index table
There are two tables for indexing: one storing the serial numbers of all the articles and called the index table; and another storing the pointer to the index table and called the index pointer table. The format in each table is index pointer table A1 for only one element in each record:

A1: pointer to index table
Index table A1, A2, ..., An, An+1
Ai: the serial numbers for $1 \leq i \leq n$
An+1: dynamic linking to next record.

From the above analysis, we know that the size of index pointer table is fixed, but the size of the index table grows as more data is added.

The database
There are many fields of information about the documents in the database, and we use a

The Electronic Library
Volume 20 · Number 6 · 2002 · 460–465

fixed length format in order to speed up the search speed. The format is:

- serial number;
- pointer to image data position for this paper;
- department;
- title;
- author;
- professor;
- abstract; and
- date.

The image data

The image data is the real content of the papers. First one page of paper is stored as a PCX file, and the procedures to construct the image data are twofold:

(1) Connect all pages of one paper into one file, and build the header storing the starting address of each PCX file. This has the advantage of being able to search more quickly for the specific files required.

(2) Connect the total files in step 1. The format is:

Paper i $\geq$ header pcx1 pcx2 pcx3, ..., etc.

System design

According to the system specification, we can design the whole system structure to meet the requirement analysis.

Data retrieval method

In order to complete the functions in the system specification, we must establish a database. Because the data before being added must be pre-defined, we only provide the functions of adding and retrieving data in this database. The whole procedure to be carried out is as follows:

(1) *Assign number.* Each data entry must have a unique serial number.

(2) *Packed image data.* After the image data has been scanned, each page is stored as a single file, so we must pack those files to form the content of the dissertation.

(3) *Reading text data.* The text data are title, abstract, author, professor, ... , etc.

(4) *Generate index.* In this step, we had to deal with three kinds of key value:

- *Numerical data.* If the key is numerical data, the index key must be finite, so we can use array to build the index table, for example:

Serial number	Year	Index table
1	80	80 – 1 3 4
2	81	81 – 2
3	80	
4	80	

And then, we can get the dissertation serial number fast by the key input by the user.

- *English word.* Although the best way to build the index table is a B-tree, we used a hash table to build the index table, because it has the advantage of being easy to implement, debug and maintain. It is also easier to build a hash table than a B-tree. In addition, although the performance of a B-tree is the best, the hash table is more flexible and still gives good performance. Furthermore, the size of the hash table can be adjusted freely. Because of the above considerations and a well-selected hash function, we can avoid collisions and speed up the indexing. In this way, the hash table will cover all of the English words in the database.
- *Chinese word.* We use a similar way to build the index table, but the first problem is how to define a Chinese phrase, because Chinese phrases are unlike English words and separated by blanks. After consideration, we proposed an economical way to solve this problem. We use the Chinese character as the key, and store the position of the word in the sentence. In this way, no matter what the characters are that compound the phrase, we can search it smoothly.

(5) *Write data and build database.* In this step, we must write all of the data and index table into the database.

Full-text reading

The main technique to accomplish full-text reading is image processing, so the first problem is the data format of the images. After consideration of the compression ratio and decompression speed, we selected the standard PCX format as our image format. Although this format cannot provide the best compression ratio, it functions well in the decompression speed and memory usage. The output to printer is similar to output to the screen. The only thing needed to do is to

The Electronic Library
Volume 20 · Number 6 · 2002 · 460–465

add suitable control codes before transmitting the image data to the printer.

User interface

From the user's point of view, as already noted the user interface is very important. The menu-driven method of the user interface makes the system very easy to use and very efficient to retrieve the data being searched for. It can act as a filter for user input – thus avoiding causing the program to crash. The user interface for the system is shown in Figure 3.

System implementation

The development platform was an IBM PC, and the programming language used was the portable C. The first Chinese-English full-text retrieving system is completed and all the National Chiao-Tung University dissertations are now written onto a CD-ROM. Users can access the Internet Explorer browser to retrieve the full text CD-ROM database over the university campus network The system structure is presented in Figure 4, including the main component: scanner, laser printer, personal computer and CD-ROM driver.

System and performance evaluation

The database system is proving to be a useful tool and the methods of construction and access have been evaluated. As regards data retrieval, the three kinds of key – numerical data, Chinese words and English words – have been implemented in the system. They include all possible indexing methods. All of the data is within the composition of these three basic indexing methods. If we want to utilize this system in a different field, only a small modification is needed.

All possible indexes about the dissertation are implemented, including department, date, author, teacher, keywords of title and keywords of abstract. This gives a rather flexible index function. Because of the image processing technique, even the text, figure, table, symbols or mathematical formula in the content of the dissertations can be shown in the screen or be printed out.

Users use the IE browser to retrieve full-text images from CD-ROM database system over the university campus network. When a 10M/bps network card is used it can cache 25 frames with good performance with 12 concurrent users. But if a 100M/bps network card is used, it can cache 25 frames with good performance with 25 concurrent users (Figure 5). Although this system is implemented, not enough testing has been carried out. The resolution and the compression ratio of the image must be improved in order to serve the best quality for the user.

Figure 3 Main menu of CD-ROM retrieval system

Chinese-English full text CDROM Retrieval System
National Chiao-Tung University of Taiwan

[A] Introduction of full text CDROM Retrieval System
[B] System User's Manual
[C] Data Query
[D] Quit

Conclusion

The current tendency of libraries is towards Web-based library automation. How to control and utilize information is a necessary condition for successful management. At

Figure 4 Architecture for Chinese-English full-text image CD-ROM retrieval system in the National Chiao-Tung University of Taiwan

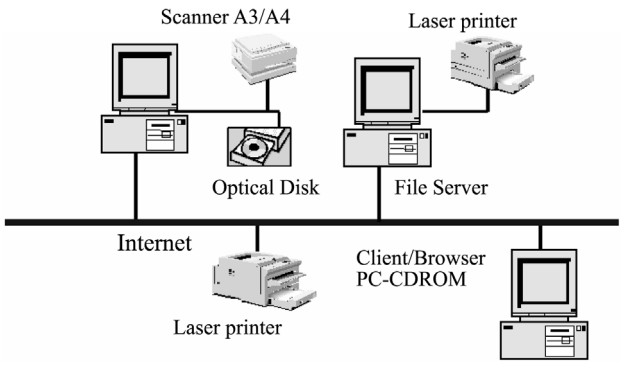
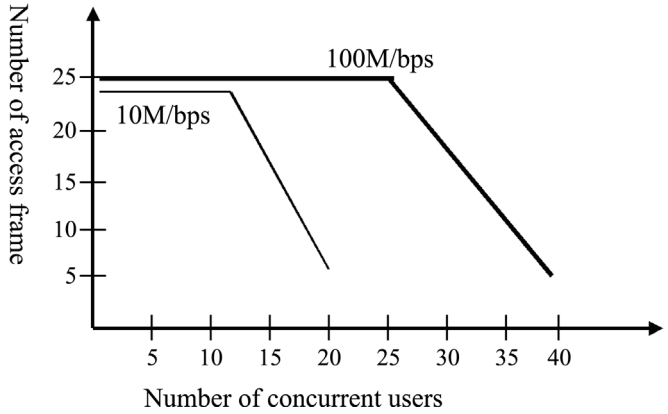


Figure 5 Performance evaluation of full-text CD-ROM retrieval system



The Electronic Library
Volume 20 · Number 6 · 2002 · 460–465

present, both CD-ROM hardware and software are very mature now and the time is opportune to further develop Chinese CD-ROM databases.

In this paper, we have presented the system design of a popularized Chinese-English full-text CD-ROM database retrieval system which has been implemented to permit professors and researchers to retrieve dissertations under the Chinese-English environment. The detailed system analysis and design for this system have also been discussed. The advantages of this retrieval system are:

- its high capacity occasioned by the employment of CD-ROM storage;
- the longevity, stability and security of CD-ROMs;
- the ease of information access and retrieval from a university campus network;
- the low cost for widespread application fields;
- inclusion of all possible keys, and easy to use menu-based interface; and
- full-text retrieval under a Chinese environment – a new departure for such a CD-ROM database system.

From the above statements, we know this system is really suitable for the enormous storage required for Chinese databases and subsequent data retrieval. In the future, we will improve the resolution of the image data, use better compress techniques, speed up the retrieval time and build a better environment for more Chinese full-text database systems.

References

Feldman, S. (1999), "The key to online catalogs that works? Testing one, two, three", *Computers in Libraries*, Vol. 19 No. 5, p. 16.

Juha, H. (1996), "Z39.50-1995 information retrieval protocol: an introduction to the standard and its usage", available at: http://renki.helsinki.fi/z3950/z3950pr.html

Yu, S.-C. and Chen, R.-S. (2001), "Developing an XML framework for an electronic document delivery system", *The Electronic Library*, Vol. 19 No. 2, pp. 102-10.