Computational Linguistics and Chinese Language Processing
vol.1, no.1, August 1996, pp. 205-221

Important Issues on Chinese Information Retrieval

Lee-Feng Chien*, Hsiao-Tieh Pu+

Abstract

In this paper, we will emphasize the significance of Chinese information retrieval in this age of the Internet, and raise several important research issues which are fundamental and require further investigation. At the same time, we will point out some problems and requirements which have often been neglected in designing general Chinese IR systems. Furthermore, experiences obtained from the design of the Csmart system will be described also.

keywords: information retrieval, full-text searching, Chinese information processing

1. Introduction

Information Retrieval (IR) is a research area with a long-term research goal of exploration of information storage, classification, extraction, indexing and browsing techniques for the retrieval of non-structural databases such as textual documents[Salton 83]. With the rapid growth in the number of electronic documents published and distributed through the Internet, the significance of IR techniques is clear, and new research directions have been developed, such as Networked Information Retrieval, Intelligent Information Retrieval and Multi-media Information Retrieval [Kahle 92, O'Kane 96, Chakravarthy 95, Lewis 96]. In fact, progress in conventional IR techniques has long been emphasized in fundamental researches, such as keyword extraction and indexing, full-text searching, term weighting, document ranking, relevance feedback, etc. [Salton 83, Salton 86]. Having been thoroughly studied, these conventional IR techniques are well established and have been successfully applied in retrieving English or Romanized documents. Recently, researchers have begun to move in new directions in exploring more advanced techniques, for instance, information filtering and adaptation techniques for Intelligent IR [Belkin 92], Web spiders and Internet searching tools for Networked IR [Koster 95,Yahoo, Lycos, Alta Vista], and speech and audio retrieval techniques for Multi-media IR [Glavitsch 92]. It can be expected that more advanced IR

* Institute of Information Science, Academia Sinica, Taipei, Taiwan. E-mail: lfchien@iis.sinica.edu.tw
+ The World College of Jounalism & Communications,Taipei, Taiwan. E-mail: htpu@cc.wcjc.edu.tw

techniques will be introduced in the near future and will contribute to areas of English or Romanized information retrieval.

Under the trend of global networking through the Internet, the growth in the number of electronic documents in Chinese and oriental languages has also been rapid, and these have mostly been published in Japan, Korea, Singapore, Hong Kong, Mainland China, Taiwan, etc. These documents are mostly non-structured and usually demand efficient IR techniques for retrieval. There is a increasing need to retrieve large numbers of such documents quickly and intelligently through world-wide information networks. Unfortunately, it is generally believed that due to the inherent differences in languages such as the lack of explicit separators, i.e. blanks or delimiters, in written oriental sentences to indicate word boundaries, the techniques developed for retrieving Romainized documents can not be directly applied to retrieval of oriental language documents. Therefore, many researchers in oriental countries are exploring new techniques adaptable to their native languages [Tseng 89, Zeng 90, 施 92, Chang 92, Wu 94, Chien 95, Ogawa 95, Wu 95, Nie 95, Liang 96, Lee 96].

It is the same motivation which urges researchers in Chinese regions to develop more efficient Chinese IR techniques. Over the last two decades, researchers have focused much effort on developing efficient indexing schemes for full-text searching, yet without many significant achievements. This might have been the result of inherent linguistic difficulties and less computer power in full-text searching. Meanwhile, if too much effort is focused on the same research topic, this will not help but even prevent extension of the scope of Chinese IR. Considering the urgent need to promote of Chinese IR, in this paper, we will emphasize the significance of Chinese IR and raise several important research issues which are fundamental and need to be further investigated. These issues include standard text collection, keyword indexing, document classification, information filtering and Internet searching tools. In addition, we will point out some problems and requirements which have often been neglected in designing general Chinese IR systems. These include the need for approximate queries and quasi-natural language queries, and technical design issues, such as two-stage searching, character-level indexing and best match searching. Much of the discussion and many of the suggestions in this paper were derived from experiences in designing the Csmart system [Chien 95a, Chien 95b], which is a high-performance Chinese IR system for the Internet resource discovery and is undergoing continuous development of Academia Sinica, Taiwan.

In the following, the important research issues for Chinese IR will be discussed in Section 2. The design issues of Chinese IR systems including both functional and

technical requirements will be pointed out in Sections 3 and 4. The relevant experiences obtained from the Csmart system will be described briefly in Section 5. Finally, concluding remarks will be given at the end of the paper.

2. Important Issues for Chinese IR

2.1 Standard Test Collection

For research and development of computer systems, standards for measurements are essential in system evaluation. IR systems are no exception. It is important to use standard test collections and measures to evaluate working systems proposed by different agencies. In Europe and America, the major reported test collections are provided by CACM, MEDLARS, TIPSTER etc. The TREC conference (Text REtrieval Conference) [Harman 95], co-sponsored by ARPA and NIST, is an ongoing conference dedicated to encouraging research in retrieval of large-scale test collections and to increasing cooperation among research groups from industry and academia. The conference brings together IR researchers to discuss system performance on standard test collections and is very effective in promoting the development of IR techniques. On the other hand, in Japan and Korea, standard test collections from the viewpoint of evaluating system functions are being constructed by unified working groups [Funio 96].

Although there are many research teams for Chinese IR development, they are all located in different regions and have little interaction. The first and immediate step is to bring together all of these researchers to share their experiences, and to verify the system performance of their systems by setting up standard test collections and evaluation measures.

2.2 Internet Searching Tools

The Internet is full of unstructured information resources which have no overall control, no standard format and no comprehensive index compared to printed and electronic sources which are quite systematic and familiar to IR researchers. Since the Internet is by far the biggest channel around the world for information exchange, distribution and retrieval, it is urgent and challenging to develop efficient Internet searching tools suitable for resolving the serious problems that have come with the information explosion. Many Internet searching tools are now available [Kimmel 96]. Unfortunately, most of them are not designed for Chinese resources on the Internet. Considering the rapid growth of Chinese resources on the Internet, high-performance Internet searching tools for Chinese

resources are in great demand.

2.3 Keyword Indexing

In addition to full-text searching, conventional IR systems also provide keyword searching in order to achieve a higher precision rate in the retrieved results. Systems with keyword searching capabilities rely on automatic keyword extraction from non-structured documents. It is noted that keyword extraction is very helpful for network IR systems, such as Excite and Infoseek, to have higher precision in retrieval, and that this capability is in demand for all of the Internet searching tools. For Chinese IR systems, automatic word extraction from text is quite difficult especially for unknown words, such as names, locations, translated terms, technical terms, abbreviations, etc. So far, there have been few related works on Chinese keyword extraction [Chang 92, Chen 94]. It is important to note that without efficient keyword extraction, many IR applications, for instance, book indexing, document classification, information filtering and text summary, cannot obtain satisfactory results [Mckeown 95].

2.4 Document Classification

With the rapid growth of Internet resources, large numbers of documents need to be managed with a good organization for efficient retrieval. Document classification [Lewis 95] is considered an important technology to reduce manual effort and to help create a more structured organization for information browsing. However, efficient document classification relies on the extraction of meaningful keywords from text and still challenges researchers working on Chinese IR [楊 93] .

2.5 Information Filtering

Information filtering is a field of study designed for creating a systematic approach to extracting information which a particular person finds useful from a large stream of information. It shares a lot of similarities with information retrieval, which actively searches out information from an existing database. Since this is an age of information anxiety, the development of information filtering technologies is now quite active [Belkin 92, Terry 92, Foltz 92]. For example, IBM provides a service which can filter articles from various newswires. Furthermore, there are commercial news archiving systems which have the capability of selective dissemination. Since research in information filtering is becoming an active area and will link the activities of IR researches in the future, it is important to bring it into the world of Chinese IR and study it extensively.

3. Basic Searching Functions

Concerning issues in designing a Chinese IR system, they depend on many different considerations, such as the characteristics of application domains, the scale and cost-effectiveness of systems, the population of probable users, the number of documents stored, etc. In this and the next sections, we will only focus on several important issues in terms of functional and technical requirements which have often been neglected in designing such a system.

3.1 Demand of Processing Approximate Queries

First of all, it has been pointed out that although there is little in the literature concerning the need for approximate searching, experiences obtained from practical system development and information service reveal the significance of approximate queries for Chinese IR. For example, Table 1 shows that in many cases, Chinese searching terms are difficult to express exactly, e.g. variations of words, uncertain terms or typos, etc. These types of searching requests commonly appear in user's queries but are difficult to properly transform into conventional Boolean queries. Fortunately, since the variations of most of the approximate terms in these searching requests have similar patterns, they can be efficiently retrieved without the aid of a thesaurus if powerful approximate text searching is developed.

Table 1. *Examples of searching terms which are difficult to express exactly*

Type	English Description	Examples
Title	President Deng-Hui Lee, Chair Deng-Hui Lee, Mr. Deng-Hui Lee	李登輝、李總統登輝、李主席登輝
Abbreviation	NTU, National Taiwan University	台大、台灣大學
Name	Names with uncertain characters	郭李建夫、郭李健夫
Translated Term	Terms with uncertain characters	巴塞隆納、巴塞隆那
Typo	Terms with typing errors	電腦概論、電腦概論
Similar Patterns	Terms with similar patterns	中華文化、中國文化
Similar Patterns	Terms with similar patterns	超導技術、超導體技術、超導材料技術

3.2 Elimination of Searching Errors in Terms of Word Semantics

In the mean time, there is a special requirement resulting from the difficulty of Chinese word segmentation. For instance, there are cases of searching terms which cause

ambiguities in terms of word semantics as shown in Table 2, where each of the illegal patterns contains the same character string of the searching term but leads to incorrect meaning in terms of semantics. Although these types of ambiguities often occur in short searching terms composed of common characters, an advanced Chinese IR system is expected to be able to eliminate these searching errors.

Table 2. *Examples of searching terms which might cause ambiguity in retrieval in terms of word semantics*

Searching Terms	Illegal Patterns
語言學	組合語言學、程式語言學
腦科	電腦科學
陳健康	指陳健康的重要
中共	其中共有、美中共同參與
化學	國際化學術會議、電腦化學理、動物演化學

3.3 Quasi-natural Language Queries

Furthermore, an advance IR system is also expected to provide more efficient searching functions which are robust in error tolerance and can easily formulate sophisticated searching requests. The design of quasi-natural language queries in advanced English IR systems [Kahle 92] is being pursued to meet the above requirements. Using quasi-natural language queries, users are allowed to formulate a request subject to a non-constrained vocabulary and searching terms. The system, provided with such a capability, can retrieve documents with the rank of statistical relevance related to the request (or query). In fact, it is obvious that formulating a quasi-natural language query is much easier and more understandable than formulating a Boolean expression query. For instance, two illustrative examples with queries formulated by Boolean expressions and quasi-natural sentences are shown below in Fig. 1.

If such quasi-natural language queries can be implemented, the extent of acceptable searching requests can be broader and more easily formulated, and the capability of an IR system can be effectively improved. This can be further seen by referring to the illustrations in Table 3 where some of the queries which can be handled with the Csmart system are shown.

EX-1:

a. an example query formulated by a Boolean expression
宏碁 and (電腦主機 or 主機板) and (產量 or 產出量)
Acer Computer main processor main board productivity total productivity

b. a similar query formulated by a quasi-natural sentence
宏碁公司電腦主機板產量
(the total number of computer main boards produced by Acer Corp.)

EX-2:

a. an example query formulated by a Boolean expression
李登輝 or 李主席登輝 or 李總統登輝
Lee, Teng-Hui Chairman Lee, Teng-Hui President Lee, Teng-Hui

b. a similar query formulated by a quasi-natural sentence
李登輝總統
(President Lee, Teng-Hui)

Figure 1 *Two examples with queries formulated by both Boolean expressions and quasi-natural sentences*

Table 3. *Cases of quasi-natural language queries with request databases*

Chinese Queries	English Description	Data Base
資訊所圖書館電話	phone no. of IIS Library	Web Page Database
尋易 Csmart 系統地址	URL of Csmart System	Web Page Database
張德培網球排名	Michael Chang's tennis ranking	Real-time News
芝加哥公牛對西雅圖超音速	Chicago Bulls vs. Seattle Super Sonics	Real-time News
奧斯卡最佳影片	Oscar best film award	Real-time News
形容美麗漂亮的女人	words describing a beautiful woman	Encyclopedic Dictionary
諾貝爾獎得主	Nobel prize winner	Encyclopedic Dictionary

4. Indexing and Searching

An efficient indexing and searching scheme for full-text retrieval is crucial in developing a high-performance IR system. According to our experience, conventional word indexing and exact match searching methods cannot be efficiently applied to Chinese IR while character-level indexing and two-stage best match searching methods can achieve better performance if the design truly considers the features of the Chinese language. In this and the next sections, the above concepts will be described and discussed in detail.

4.1 Inverted File vs. Signature File

The most important design issue for a conventional IR system is the choice of the adopted text access method. According to investigations of text access methods available, inverted file and signature file [Faloutsos 85] are two of the most important classes of methods. Basically, both the inverted file and signature file methods rely heavily on certain indexing systems to speed up the searching time. Based on experiments carried out on English IR, the inverted file method, compared with the signature file method, is much preferred for designing commercial systems because of its fast retrieval speed and flexibility in implementing diverse searching functions. On the other hand, according to the analysis of some recent researches[Chien 95b] [Liang 96], the signature file method is believed to be the most promising approach for retrieving large Chinese document databases considering the difficulties of Chinese word segmentation. Similar results were also found in Ogawa's work in retrieving Japanese documents [Ogawa 95]. However, the above findings were all based on the need to improve retrieval speed and to reduce indexing space for conventional full-text searching. With the advances in modern IR, many other retrieval functions or techniques, such as quasi-natural language queries, document ranking and relevance feedback, all need to be included in an advanced IR system. To meet these requirements, it is found that neither the inverted file nor signature file method is sufficient. Therefore, whether the signature file or inverted file method is better is not very critical. According to our observations, except for providing the functions of approximate and quasi-natural language queries, a text access method can be considered as the most efficient and promising approach only when it can reduce the difficulty of Chinese word segmentation, create character-level indexing, carry out efficient best match searching and perform two-stage searching.

4.2 Best Match Searching

Conventional Chinese IR approaches are primarily designed for exact match searching and to support Boolean queries. As indicated in the last section many search terms in Chinese are difficult to express exactly; thus, best match searching (or fuzzy search) [Croft 88] [Stanfill 86] is suggested to be more useful in retrieving Chinese documents [Chien 95a]. Basically, best match searching can remedy some difficulties which exist in conventional Boolean searching by implementing document ranking and relevance feedback capabilities [Salton 83]. More importantly, it allows inexact query expressions and can meet the special requirements addressed in the last section.

4.3 Reduction of Word Segmentation Difficulties

Since words are easily segmented in English text, the retrieval functions in English IR

systems are mostly based on word information, such as word-inverted file methods, term weighting schemes, term vector approaches, etc. In written Chinese, no explicit separators are inserted between written words to indicate boundaries. A Chinese sentence usually can be segmented into many different possible word combinations, and it is difficult to decide on the correct combination. Appropriate word segmentation of Chinese text should rely on sophisticated syntactic and semantic analysis on the text [梁 89, Sproat 90, Lee 91, Chen 92, Chang 92]. Segmentation of correct word combinations and identification of proper nouns is a very difficult task [Wu 94, Wu 95, Nie 95]. This results in two serious problems for Chinese IR. Firstly, there will be no consistency between word segmentation of documents and queries if words are to be used as the basic processing unit in creating indices (word-based indexing). Secondly, it is difficult to identify proper nouns, such as names and locations, which are usually the keywords in queries. Furthermore, it creates many implementation difficulties, for instance the construction of a lexicon, the formation of lexical rules, the design of ambiguity resolution methods, etc. Therefore, many highly-regarded word-based document retrieval approaches in English IR, such as the word inverted file and the vector space model approach, cannot be directly applied in retrieving Chinese documents [Tseng 89, Larsson 90]. These problems or difficulties will not exist if characters instead of words are used as the basic processing unit in creating an index (character-based indexing).

4.4 Character-level Indexing

The advantages of character-based indexing compared with word-based indexing consist of efficient processing in indexing, no requirement for word segmentation analysis, robustness, and high recall rate in information access. However, character-based indexing has many weaknesses, such as the demand for large space overhead, slower retrieval speed, the lack of high-level semantic information and poor retrieval precision, etc.

In fact, it has to be pointed out that each Chinese character and character bigram holds more semantic meaning than does a letter in English. This is because there are about 13,053 commonly-used Chinese characters, and each Chinese word is usually composed of one or two of these characters. In a Chinese document, though the composed words are difficult to correctly segment using a computer, it is easy to extract all of the composed characters and character bigrams (each pair of adjacent characters) from the text. These characters and character bigrams hold certain semantics and form the features of the document. For example, there may be a three-character keyword 中國人 (Chinese) within a document which cannot be correctly segmented, yet its decomposed characters and character bigrams, i.e., 中 (center), 國 (country), 人

(people), 中國 (China) and 國人 (citizen), remain relevant semantics of the word. That is, if we use these characters and character bigrams to form the features of the word, then there still remain strong relationships between the word and other relevant keywords such as 中華民國 (Republic of China), 中華人民共和國 (People's Republic of China) and 美國人 (American people). In addition, the word can even be distinguished from irrelevant words, such as 電腦 (computer), 資訊 (information), etc.

On account of the difficulties of word indexing, for many years, conventional Chinese IR systems have been compromised by adopting character-based indexing methods, which were often implemented with inverted files and suffered from the demand for large space overhead and from low retrieval speed. Fortunately, it has recently been proved that the above difficulties can be effectively reduced by using specially designed character-based signature file methods[Chien'95]. These methods can also perform efficiently in approximate text searching in finding similar Chinese searching terms. This capability is in great demand in the retrieval of Chinese texts, but is difficult to implement using conventional inverted file methods. Moreover, such signature-based approaches are easier in design for processing of quasi-natural language queries and can achieve accurate ranking results if knowledge from Chinese natural language analysis is properly used and if a two-stage searching approach is adopted.

4.5 Two-Stage Searching

It can be found from the above discussion that neither word indexing nor character indexing is sufficient to record extensive information in the indexing system or to handle all of the possible searching functions. In this way, a scheme of two-stage searching is necessary in modern IR systems and is specially useful for retrieving oriental language documents. A general two-stage searching process is implemented using two primary modules: the fast search and the detailed search modules. The purpose of the fast search module is to reduce the number of non- qualifying documents and to obtain a higher recall rate. It uses a simple but easily constructed indexing system to perform fast search. The documents which are irrelevant can be filtered out as much as possible after the fast search process is carried out. At the same time, the purpose of the detailed search module is to obtain a higher precision rate. The remaining documents are re-examined at this stage; the contents of the documents are scanned, detailed analysis is performed, and the relevance values are obtained. The detailed analysis may include sophisticated functions such as word segmenation, approximate matching, document ranking, keyword extraction, etc. By scanning the document contents, these functions can be easier to implement.

5. Experiences with Csmart

In order to pursue high performance Chinese information access on the Internet, Csmart is being developed at Academia Sininca to obtain natural language information retrieval techniques in Chinese networked information discovery and retrieval [Chien 95a, Chien 96]. The design of the IR system is completely based on the above two-stage searching concept and character-level signature file approach for fast and intelligent retrieval of large Chinese full-text document databases. Using this approach, the inherent difficulties of Chinese word segmentation and proper noun identification have been effectively reduced, and efficient approximate and quasi-natural language queries implemented. In the following, the experiences obtained in developing Csmart will be briefly described .

To enable the proposed signature file approach to process approximate and quasi-natural language queries and perform both best and exact match searching in a large database efficiently, Csmart uses a statistical character-level signature extraction method for generation of document signatures, and a signature-based ranking function for estimation of the relevance between queries and documents. Conventional English signature file approaches mostly rely on well-tuned hashing functions in the transformation of word signatures. It is not easy for such hashing functions to consider the characteristics of various document databases in determining appropriate word signatures; hence, the probability of false drops in full-text searching cannot be effectively reduced. Furthermore, the generated document signatures only serve as filters for full-text searching and cannot serve as feature vectors for best match searching at the same time. To remedy the weaknesses of conventional signature file approaches, the presented signature extraction method replaces the use of hashing functions with a statistical text modeling approach and a specially designed character grouping algorithm. Utilizing this method, it is possible to consider the characteristics of various document databases in the generation of signatures. Meanwhile, each bit of the signatures has more semantic meaning, which is often vague using conventional methods.

The IR kernel of the Csmart system, as shown in Fig. 2, is composed of two modules: the fast search module and the detailed search module. These two modules cooperate to perform a two-stage searching process, i.e., a signature file test and full-text scanning individually. For input of a quasi-natural language query, the IR system performs the two-stage best match search process to retrieve the highly relevant documents. At the stage of the signature file test, the character information of the input query is extracted. Then, the fast search module generates the signature of the input query according to its composed single characters and character bigrams (and syllable bigrams, depending on whether or not it is a spoken query [Lin 95a, Lin 95b, Lin 96]). It matches

the query signature using the document signature file created at the indexing stage, calculates the similarity values using a signature-based ranking function [Chien 95a], and filters out many of the non-qualifying documents.

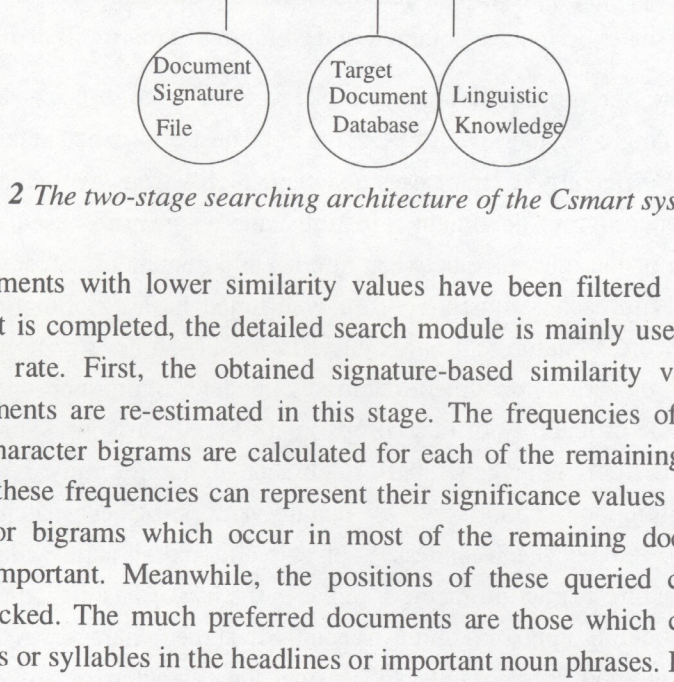

Figure 2 *The two-stage searching architecture of the Csmart system*

Since documents with lower similarity values have been filtered out after the signature file test is completed, the detailed search module is mainly used to obtain a higher precision rate. First, the obtained signature-based similarity values of the remaining documents are re-estimated in this stage. The frequencies of the queried characters and character bigrams are calculated for each of the remaining documents. The inverses of these frequencies can represent their significance values at this stage. The characters or bigrams which occur in most of the remaining documents are, therefore, less important. Meanwhile, the positions of these queried characters or syllables are checked. The much preferred documents are those which contain more queried characters or syllables in the headlines or important noun phrases. Based on the above strategies, the preferred documents can be extracted. The keywords conceived in the query are then extracted to find the relevant sentences in the preferred documents by performing word segmentation on the relevant text fragments. In this way, the documents with important selected text fragments which are highly relevant to the query can be found and displayed.

The performance of Csmart system has been carefully evaluated. First of all, it is found to be able to achieve fast retrieval speed even if more than one gigabyte of text is to be retrieved. Meanwhile, the space overhead of the document signature file is scalable and occupies only about 20~40% of the database size on average. Furthermore, in our experiments, about 99.2% of the irrelevant documents can be eliminated after the signature file test if common quasi-natural language queries (which were not short and consisted of meaningful keywords) were given while the recall rate was on the order of

90%. This means that for a document database with a total of 10,000 documents at most, only about 80 plus the desired documents need to be further checked. This efficient performance permits us to use more sophisticated linguistic knowledge and text scanning techniques in the detailed search module.

An example to illustrate the retrieving process of the IR system is shown in Fig. 3, which shows a Chinese interface with a list of selected databases in the top window, a given quasi-natural language query for retrieving the news database in the left window and the retrieved news titles with the processing time and belief scores shown on the right window. The recall rate for common quasi-natural language queries can remain on the order of 90%, and the precision rate can as high as 95%. Meanwhile, the total retrieval speed without including the processing time of speech recognition and network communication can be achieved within one second. However, since large scale testing is still under way, the above testing results are only values obtained from initial experiments.

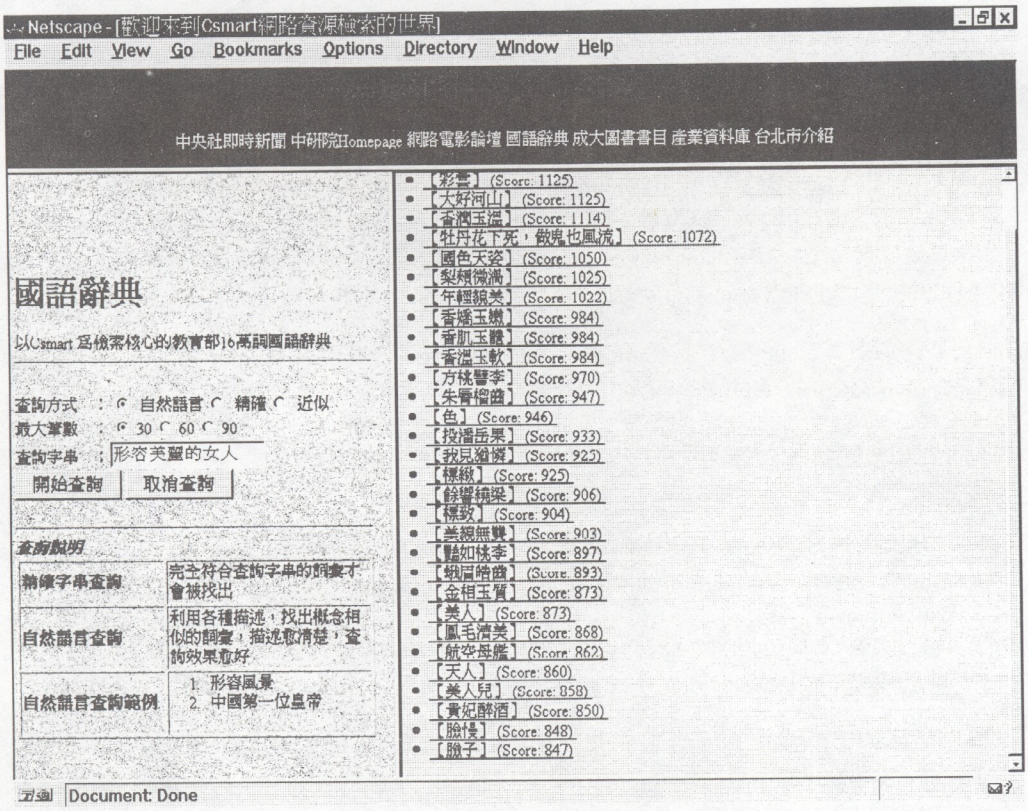

Figure 3 *An example to illustrate the retrieving process of the Csmart system*

6. Concluding Remarks

In the above, we have raised important research issues, including Test Collection, Internet Searching Tools, Keyword Indexing, Document Classification and Information Filtering, which are fundamental and need to be further investigated. On the other hand, we pointed out several issues required but neglected in designing general Chinese IR systems. These consist of the basic searching functions, such as approximate queries and quasi-natural language queries, and technical issues such as two-stage searching, best match searching and character-level indexing which are efficient for Chinese IR. Furthermore, skills and experiences derived from the Csmart system have also been reported.

References

Alta Vista Home Page (http://altavista.digital.com/)

Belkin, Nicholas J., *et al.*, "Information Filtering and Information Retrieval: Two Sides of the Same Coin?" *Communications of the ACM*, Vol. 35, No 12, Dec. 1992.

Chakravarthy, Anil S., K. Hnase, "Netserf: Using Semantic Knowledge to Find Internet Information Archives," *ACM SIGIR'95*.

Chang, Jyun-Sheng, Tsung-Yih Tsengm, Ying Chen, Huey-Chyun Chen, Shun-Der Chen, John S. Liu and Sur-Jin Ker, "A Corpus-based Statistical Approach to Automatic Book Indexing", *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP'92)*, 1992, pp. 147-151, Trento, Italy.

Chang, T.H., T.S. Chang, M.Y. Lin and K.Y. Su, "Statistical Models for Word Segmentation and Unknown word Resolution," *ROCLING'92*, 1992.

Chen, Keh-Jiann, *et al.*, "Word Identification for Mandarin Chinese Sentences," *COLING'92*, 1992.

Chen, K.H.and H.H. Chen, "Extracting Noun Phrases from Large- Scale Texts: A Hybrid Approach and Its Automatic Evaluation," *ACL'94*, 1994, pp. 234- 241.

Chien, Lee-Feng, 'Fast and Quasi-Natural Language Search for Gigabytes of Chinese Texts", *ACM SIGIR'95a*, 1995a.

Chien, Lee-Feng, " 尋易 (Csmart) -- A High-Performance Chinese Document Retrieval System", The 1995 International Conference of Computer Processing for Oriental Languages, *ICCPOL'95b*, 1995b.

Chien, Lee-Feng, "A Model-Based Signature File Approach for Full- text Retrieval of Chinese Document Databases," To appear on Computer Processing of Chinese and Oriental

Languages, 1995c.

Chien, Lee-Feng, "Natural Language Information Retrieval with Speech Recognition Techniques for Network Chinese Resources Discovery", Intenational Workshop on Information Retrieval with Oriental Languages, 1996, Korea.

Croft, W.B. Croft and P. Savino, "Implementing Ranking Strategies Using Text Signatures", ACM Trans. *Office Information System*, Vol. 6, No. 1, Jan. 1988, pp. 42-62.

Faloutsos, C., "Access Methods for Text", *ACM Computing Surveys*, March 1985, pp.49-74.

Funio, M., *et al.*, "Test Collection for Japanese Information retrieval Systems from the Viewpoint of Evaluating System Functions", Proceedings of the 1996 Workshop on Information Retrieval with Oriental Languages, Taejon, Korea, 1996, pp. 42-47.

Foltz, Peter W., *et al.*, "Personalized Information Delivery: An Analysis of Information Filtering Methods", *Communication of the ACMs*, Vol. 35, No. 12, Dec., 1992.

Glavitsch, U. and P. Schauble, "A System for Retrieving Speech Dcouments", ACM SIGIR Conference on R&D in Information Retrieval, 1992, pp. 168-176.

Harman, D., "The 3rd Test Retrieval Conference (TREC-3)," National Institute of Standards and Technology, 1995.

Kahle, Brewster, *et al.*, "Wide Area Information Servers: An Executive Information System for Unstructured Files," Electronic Networking: Research, *Applications and Policy*. 2(Spring 1992): 59-68.

Kimmel, S., "Robot-generated Databases on the World Wide Web", *Database*, 1996, pp. 41-49.

Koster, Martijn, "Robots in the web: Threat or Treat ?" Available on the World Wide Web at http://web.nexor.co.uk/mak/doc/robots, 1995.

Larsson, R. and Sunneback, J., "Chinese in a Text Database System for Fuli Text Searching", In Database Development and Chinese Information needs, edited by Minzu Zeng, London, Aslib, 1990.

Lee, H. J. *et al.*, "Rule-based Word Identification for Mandarin Chinese Sentences - A Unification Approach", *CPCOL*, Vol. 5, No 2, 1991.

Lee, Ahn and Shin, "An Effective Indexing Method for Korean Text Retrieval," International Workshop on Information Retrieval with Oriental Languages, Korea, 1996.

Lewis, David D. and Karen Sparck Jones, "Natural Language Processing for Information Retrieval," *Communication of the ACM*, Vol. 39, No. 1, Jan. 1996, pp. 92-101.

Lewis, D., "Evaluating and Optimizing Autonumous Text Classification Systems," *ACM SIGIR'95*, 1995.

Liang, Tyne, Suh-yin Lee and Wei-Pang Yang, "Optimal Weight Assignment for a Chinese Signature File," *Information Processing and Management*, Vol 32, No. 2, 1996, pp. 227-237.

Lin, Sung-Chien, Lee-Feng Chien and Lin-Shan Lee, "An Efficient Voice Retrieval System for Very-Large-Vocabulary Chinese Textual Databases with a Clustered Language Model," To appear on Proceedings of the 1996 IEEE International Conference on Acoustics, Speech and Signal Processing, *ICASSP'96*, 1996.

Lin, Sung-chien, Lee-Feng Chien and Lin-Shan Lee, "A Syllable-based very-Large-Vocabulary Voice Retrieval System for Chinese Databases with Textual Attributes," The 4th European Conference on Speech Communication and Technology, *EuroSpeech95*, 1995.

Lin, Sung-Chien, Lee-Feng Chien and Lin-Shan Lee, "Unconstrained Speech Retrieval for Chinese Document Databases with Very Large Vocabulary," The 4th European Conference on Speech Communication and Technology, *EuroSpeech95*, 1995.

Lycos HomePage (http://www.lycos.com/)

Mckeown, K., D. Radev, Generating Summaries of Multiple News Articles, *ACM SIGIR'95.*,1995.

Nie, Jian-Yu , "A Unifying Approach to Segmentation of Chinese and Its Application to Text Retrieval," *ROCLING'95*, 1995.

Ogawa, Y., "A New Character-based Indexing Organization Using Frequency Data for Japanese Documents," *ACM SIGIR'95*, 1995.

O'Kane, Kevin C., "World Wide Web-based Information Storage and Retrieval," *Online & CDROM Review*, Vol. 20, No.1, 1996.

Salton, G. and J. Michael McGill, "Intoduction to Modern Information Retrieval," New York, NY: McGraw-Hill, 1983.

Salton, G., "Another Look at Automatic Text Retrieval", *CACM*, Vol. 29, No. 7, 1986.

Sproat R. and C. Shih, "A Statistical Method for Finding Word Boundaries in Chinese Text," *CPCOL*, Vol. 4, March 1990.

Stanfill, C. and B. Kahle, "Parallel Free-Text Search on the Connection Machine System," *Communication of ACM*, Vol. 29, No. 12, Dec. 1986, pp. 1229-1239.

Terry, D. and S. Loeb, "Special Section on Information Filtering," *CACM*, Dec. 1992, pp. 26-81.

Tseng, S.S., C.C. Yang and Ching-Chun Hsieh, "On the design of Chinese Textual Database," *Computer Processing of Chinese and Oriental Languages*, 4, 1989, pp. 240-271.

Wu, Zimin and Gwyneth Tseng, "Chinese Text Segmentation for Text Retrieval:Achievements and Problems," *JASIS*, 44(9), 1994, pp. 532-542.

Wu, S., Gais Home Page, http://gais.cs.ccu.edu.tw/

Wu, Zimin and Gwyneth Tseng, "ACTS: An Automatic Chinese Text Segmentation System for Full-text Retrieval," *Journal of American Society for Information Science*, 46(2), 1995, pp. 83-96.

Yahoo HomePage (http://www.yahoo.com/)

Zeng, Minzu, "Database Development and Chinese Information needs," London, Aslib, 1990.

梁南元，「書面漢語自動化分詞系統 -CDWS ，」In 中文信息處理， No. 1, 1989.

施水才等，「中文自動標引中蘊含概念的分析，」*ICCPOL'92*, pp. 196-201.

楊允言，謝清俊，陳淑美，陳克健，「中文文件自動分類之研究，」*ROCLING*, 1993.