Networked Information Retrieval Using Unconstrained Mandarin Speech Queries

Lee-Feng Chien, Hsiao-Tieh Pu, Ming-Jer Lee,
Ming-Chan Chen, Hung-Ming Chen and Tun-I Huang

Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C.
Tel: 886-2-788-3799 ext. 1801
E-mail: lfchien@iis.sinica.edu.tw
Fax: 886-2-782-481

Abstract

This paper presents new research on Chinese networked information retrieval with the capabilities of processing unconstrained speech queries and providing efficient natural language searching. The present system is designed to be able to resolve the conventional difficulties of both the input of Chinese characters and the retrieval of large Chinese texts in terms of network information access. In dealing with these difficulties, a signature-based natural language search method has been developed. Meanwhile, for handling of unconstrained speech queries, the present system is integrated with continuous Mandarin speech recognition technology and other robust techniques. With the distinguished features of the present system, the preliminary experimental results show that it achieves good performance in many ways, especially in terms of the accuracy of retrieval, the flexibility of query formation and the possibility of developing a spoken dialogue interface.

I. Introduction

The Internet is full of unstructured information resource which has no overall control, no standard format and no comprehensive index compared to printed and electronic sources. 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 on information explosion. There have been a lot of Internet searching tools available by now [Kimmel'96, Alta Vista, Lycos]. Unfortunately, most of them are not designed for Chinese resources on the Internet [Wu'95]. Considering the rapid growth of Chinese resources on the Internet, high-performance Internet searching tools for Chinese resources are highly in demand.

This paper presents new research on Chinese information retrieval with the capabilities of processing unconstrained speech queries and efficient natural language searching in terms of network information access. For this purpose, a number of successful technologies have been developed in our research group in recent years. For example, these include fast search for multi-giga-byte Chinese texts [Chien'95a], fuzzy search for quasi-natural language queries [Chien'95b], continuous Mandarin dictation [Lee'95], information retrieval using unconstrained Mandarin speech queries[Lin'96, Lin'95a, Lin'95b], etc. Based on these efforts, we are currently developing a system which allows users to retrieve Chinese networked information such as Web pages or real-time news fast and efficiently by using unconstrained Mandarin speech queries. This system is designed to be capable of resolving the conventional difficulties of both the input of Chinese characters and the retrieval of large Chinese texts.

The integration of speech recognition and natural language information retrieval technologies can provide users with a more convenient interface for information retrieval systems, yet few works have truly focused on this task so far [Glavitsch'92, Yamada'94]. This is because current speech recognition technology is not reliable enough, and the mechanism of commonly-used natural language search is not robust either. For the Chinese language, since it is not alphabetic, to input Chinese characters into computers is still a difficult and challenging problem. Therefore, speech retrieval of Chinese databases has become more important and needs to be pursued. Fortunately, the Chinese language has many good features, such as the mono-syllabic structure of Mandarin speech, the rigid semantic meaning and coverage of Chinese characters and character bigrams, and the discrimination ability of Mandarin syllable bigrams in similarity estimation, etc. These features make it possible for the present system to process speech queries and perform information retrieval with high efficiency.

There are many inherent difficulties in Chinese word segmentation and proper noun identification

[Chen'92, Lee'91]. To deal with these difficulties, a signature-based fuzzy search method which carefully considers the features of Chinese information retrieval has been successfully developed [Chien'95a, Chien'95b] and effectively improved in the present system. In the mean time, for development of a high-performance system which can handle unconstrained speech queries, the present system is also integrated with the continuous speech recognition technology of the Mandarin dictation machine, Golden Mandarin III [Lee'95]. Golden Mandarin III is the first system in the world which can dictate continuous Mandarin speech with a very large vocabulary and unlimited texts. Such a speech recognition system has been in development for more than ten years. Based on this speech recognition technology and certain enhanced techniques for processing of speech queries, the present system allows users to use either typed or spoken natural language queries to perform retrieval of document databases. Moreover, regarding the robustness of the present system, it includes other efficient techniques which enable it to properly use the knowledge acquired from document databases to provide grammatical constraints for speech recognition, handle the tolerance of speech recognition errors with syllable-based fuzzy search, provide an accurate ranking algorithm for natural language queries, etc. With these special features of the present system, the preliminary experimental results show that it not only accepts unconstrained speech queries, and but also achieves good performance in many ways, especially in terms of the accuracy of retrieval, flexibility of query formation and in the possibility of developing a spoken dialogue interface. This has encouraged us to continue this advanced research and pursue even greater improvements.

In the rest of this paper, Section II will give an overview of the entire system in advance. Then, Section III will focuse on the general fuzzy search for Chinese natural language information retrieval. The techniques used for resource discovery and syllable-based signature extraction will be further described in Section IV and handling speech queries in Section V. Finally, the experimental results and concluding remarks will be given in the Section VI.

II. The Overall Approach

The block diagram of the present system is shown in Fig.1. It can be decomposed into three major blocks. The first subsystem, namely resource discovery subsystem analyzes the HTML documents (or Web pages) obtained from the Internet, extracts important pieces of information, inserts into different databases (e.g., Web database or news database), and generates statistical indices (document signatures) and parameters (language models) for the other two subsystems. This subsystem is operating during the database preparing phase and will be further discussed in Section IV.

When a natural-language speech query (e.g, like "please find the world-wide responses about the President election of Taiwan" in Chinese for the news database) is entered, the speech recognition subsystem and the information retrieval subsystem cooperate to retrieve documents by estimating the similarities between the query and all of the documents in the target database. The responsibility of the speech recognition subsystem is to transcribe the query into the most possible character string with corresponding syllable information. This subsystem includes two modules: the syllable recognition module and the character string search module. The former produces several possible syllable candidates for each syllable in the input query. The later search module then searches for the most possible character string using the language model generated by the first system. The recognized character string with its corresponding syllable information is then formed as the input of the information retrieval subsystem. The technology of the speech recognition subsystem is quite sophisticated and has been described elsewhere [Lee'95, Lin'96, Lin'95a, Lin'95b].

By accepting the recognized character string and its syllable information, the information retrieval subsystem performs signature-based fuzzy search to retrieve the highly relevant documents. This subsystem 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. At the stage of the signature file test, both the character and syllable information of the recognized input query will be used. Then, the fast search module generates the signature of the input query according to its composed single characters, character bigrams and syllable bigrams. It matches the query signature using the document signature file, calculates the similarity values using a signature-based ranking function, and filters out many of the non-qualifying documents. Continuing in the stage of the full-text scanning, both the character and syllable information of the input query are used again to analyze all of the other remaining documents by scanning their document contents and using linguistic heuristics. Finally, the documents which are highly related to the input query will be retrieved and displayed through the network. The general concept of the fuzzy search will be further introduced in Section III. Meanwhile, the extraction of syllable-based signatures will be described in Section IV and the entire fuzzy search for speech queries in Section V.

III. Fuzzy Search for Chinese Natural Language Information Retrieval

In this section, the general concept of the proposed fuzzy search method which forms the basis of the present system will be first described in detail.

Efficient Full-text Search

An efficient indexing and searching method for full-text retrieval is crucial in developing a high-performance IR system [Salton'83]. According to our experience, Chinese IR is not easily handled using conventional word-based indexing and exact match searching methods, whereas character-based indexing and best match searching methods can achieve much better performance if the design truly considers the features of Chinese IR. Since Chinese words are formed by a sequence of characters, and words in a Chinese sentence are not clearly bounded by spaces as are those in English, word-based indexing will cause inappropriate word breaking and proper noun identification [Wu'94]. On account of these difficulties, for many years conventional Chinese IR systems 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 [Liang'94, Chien'95a, Chien'95b]. These methods can also perform efficiently in approximate text searching in finding similar Chinese searching terms. This capability is great in demand in the retrieval of Chinese texts but is difficult to be implemented using conventional methods. Moreover, such signature-based approaches are easier in design for processing natural language queries and achieve accurate ranking results if knowledge from Chinese natural language analysis is properly used.

Natural Language Information Retrieval

For processing of natural language queries, the major processing steps of the conventional approaches include removing stop words and extracting key terms from input queries, determining the weights of the key terms, calculating the relevance values between the extracted terms and each of the documents with a certain ranking function, and then sorting all of the documents into order. Each of these steps relies heavily on the capability of natural language analysis. Since Chinese words are difficult to properly break as mentioned above, such processing steps will not be suitable for dealing with Chinese texts unless there is a breakthrough in Chinese natural language analysis. After careful study for several years, we have found a better method for processing Chinese natural sentences. This method consists of several basic strategies such as extraction of character-level information, character grouping and information sharing, signature-based ranking and reduction of searching space, and text-based ranking and detailed analysis, etc. These strategies have been carefully adopted in the present system to deal with speech queries. Before giving the details of the present system, the concepts behind these strategies will be briefly described below.

Extraction of Character-level Information

First, it has to be pointed out that each single Chinese character and character bigram holds more semantic meaning than 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 key word 中國人 (Chinese) within a document which cannot be correctly segmented, yet its decomposed characters and character bigrams, i.e., 中(centered), 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, there still remain strong relationships between the word and other relevant key words such as 中華民國 (Republic of China), 中華人民共和國(People 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.

Character Grouping and Information Sharing

Based on the above analysis, it can be seen that the features of Chinese natural language queries or documents can be basically formed by their composed single characters and character bigrams rather than by words. The inverse frequencies of these characters and bigrams in complete documents can represent their significance values. In this way, frequently-used single characters such as 的 (of) and 一 (one) can be almost treated as stop characters. Character bigrams appearing in proper nouns or personal names might have very high significance values, which would solve the conventional difficulty of proper noun identification.

Though the composed characters and character bigrams make up the features of a Chinese document, the number of possible character bigrams in a database is often too large (usually exceeding 1M different bigrams in a database of 100MB) and needs to be reduced. Since many of these characters and bigrams carry similar information, similar characters and character bigrams can be grouped into classes and can share same information. In this way, it has been proved that it is possible to effectively reduce the size of the character set without losing too much information[Chien'95a]. As to the similarity among characters and character bigrams, this can be estimated by using the database content (or prepared training corpus) and the highly associated characters and character bigrams. For example, the character bigram 電腦 (computer) seems to be more similar to 資訊 (information) than to 公園 (park) because there are many bigrams highly associated with 電腦 (computer) and 資訊 (information), e.g., 軟體 (software), 程式 (program), etc.

Signature-based Ranking and Reduction of Searching Space

To handle a large document database, the efficiency of the retrieval speed should be carefully considered. Therefore, a specially designed signature extraction method is proposed. The character-level information of the queries or documents are represented using signatures rather than conventional term vectors. The signatures are fixed binary bit strings which indicate the presence of character classes in the queries or documents. Each bit in the signatures records the existence of a certain class (a set of characters or bigrams) and its corresponding semantics. The size of each signature is, thus, determined by the number of grouped character classes. This kind of signatures is easy to access and saves space. The derived document signatures constitute a bit-slice document signature file. Using this signature extraction method, the existence of a queried character string in a test document can be easily checked. For example, the character string does not exist if not every corresponding bit of its composed characters and bigrams is set in the document signature. This signature file, therefore, serves as a filter which can be used to rapidly test for the existence of any given character string. Furthermore, it can be used as a feature vector for rapid similarity estimation.

Text-based Ranking and Detailed Analysis

At this point, the entire method can be formed. The proposed fuzzy search method can be implemented using two primary modules: the fast search and the detailed search modules, as shown in Fig. 1. The purpose of the fast search module is to reduce the number of non-qualifying documents and obtain a higher recall rate. It uses a signture-based ranking function to estimate the similarity values between queries and documents. The documents which have lower similarity values can be filtered out 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 obtained signature-based similarity values of the remaining documents are re-estimated at this stage; the contents of the documents are scanned, the real frequencies of the composed characters and bigrams are calculated, and detailed analysis is performed.

IV. Document Analysis and Syllable-based Signature Extraction

The above fuzzy search method can be a general approach. Since there are various possible errors in speech queries and since the process of speech recognition takes time, the signature extraction method and searching strategy in the present system have been enhanced by considering the characteristics of Mandarin speech.

Mono-syllabic Structure of Mandarin Speech

Although there are 13,053 commonly-used Chinese characters, each character is mono-syllabic, and the total number of phonetically-allowed Mandarin syllables is only 1,345. Generally, the accuracy of syllable recognition is high in Mandarin speech recognition systems[Lee'96]. The number of possible syllable bigrams exceeds 1M (1,345*1,345) and is close to the number of different character bigrams in a common database as mentioned in the last section. This indicates that the number of homonym character bigrams for each syllable bigram is small. A character bigram can use its corresponding syllable bigram to almost retain its semantics in terms of IR. This feature makes it possible to develop a syllable-based information retrieval system for Chinese textual databases [Lin'95a,Lin'95b]. The present system takes advantage of syllable information in all of the subsystems: the resource discovery subsystem, the speech recognition subsystem and the information retrieval subsystem. In the following, the process followed in the document analysis subsystem will be introduced.

Signature Extraction with Both Character and Syllable Information

To effectively reduce the number of speech recognition errors, this subsystem applies syllable knowledge acquired from the database in the language model construction module. It first transcribes all of the Chinese characters of the documents in the database into Mandarin syllables. It then constructs a specially-designed statistical language model, combined with both syllable and character information [Lin'96], to provide more powerful linguistic constraints in speech recognition of the input query. Furthermore, the signature extraction module uses a syllable-based signature extraction method which will be described below.

Different from the way in which typed natural language queries are processed, the proposed signature extraction method considers both character and syllable information in the generation of signatures. With this signature extraction method, each document say a fixed block of text for simplicity, will be assigned a pre-specified two-segment bit string as its signature. Each bit in the first segment of the signature represents the occurrence of a class of single Chinese characters or bigrams, and the second segment represents that of a class of similar Mandarin syllable bigrams. The grouping of similar syllable bigrams is the same as the grouping of similar characters. For a given document, the signature of the document is first obtained by setting the features of the composed Chinese characters and character bigrams in the first segment. Then, the corresponding syllable bigrams of the composed character bigrams are transcribed in order to set the features of the second segment. In this way, the existence of every similar class formed either by characters or by syllables can be clearly indicated.

V. Natural Language Search for Speech Queries

Since there usually exist recognition errors in speech queries, the fuzzy search used to handle speech queries should be more robust. When a natural-language speech query is entered, the speech recognition subsystem transcribes the query into the most possible character string. To increase the reliability of the recognized output, the corresponding syllables of the recognized character string are also considered in the information retrieval phase to obtain high accuracy in Mandarin syllable recognition. After accepting the recognized character string and its corresponding syllable information, the information retrieval subsystem performs a two-stage fuzzy searching process, previously described in Sections II and III, for retrieval of the highly relevant documents.

Signature File Test and Fast Search

In the stage of the signature file test, first both character and syllable information of the recognized input query are used. The fast search module generates the signature of the input query according to its composed characters, character bigrams and syllable bigrams. It then quickly matches the query signature with the document signature file generated by the document analysis subsystem, calculates the similarity values using a signature-based ranking function which will be described below, and finally filters out many of the non-qualifying documents.

In order to quickly filter out non-qualifying documents and to avoid the difficulty of key word extraction, the signature-based ranking function is used and defined below.

Definition of the Signature-based Ranking Function

For each given query q and document d, the function B(q,d) will return a value as their relevance. A higher relevance value means that they have a stronger relation in terms of semantics. The definition of this function is given below.

$$B(q,d) = t_1 \sum_{1 \le i \le n} I_{ci,d} / N_{ci} + t_2 \sum_{1 \le i \le n-1} I_{cici+1,d} / N_{cici+1}$$
$$+ t_3 \sum_{1 \le i \le n-1} I_{sisi+1,d} / N_{sisi+1} \quad \text{,where}$$

q is a given query which is a sequence of n Chinese characters, i.e., $q = c_1, \ldots., c_n$,

d is a given document,

c_i is a composed single character of query q,

$c_i c_{i+1}$ is a composed character bigram of query q,

$s_i s_{i+1}$ is the corresponding syllable bigram of $c_i c_{i+1}$,

$I_{ci,d}$, $I_{cici+1,d}$, $I_{sisi+1,d}$ are binary values to indicate the existence of corresponding classes of c_i, $c_i c_{i+1}$, $s_i s_{i+1}$ in document d, respectively

N_{ci}, N_{cici+1}, N_{sisi+1} are the total numbers of documents in the database which contain c_i, $c_i c_{i+1}$, $s_i s_{i+1}$, respectively,

t_1, t_2 and t_3 are three predefined weight values

Basically, in the above function it is assumed that the relevance value is mainly proportional to the existence and IDF of each extracted class. In fact, this assumption is derived from overall consideration of the difficulty of key word extraction, the reliability of speech recognition, and the significance of each single Chinese character, character bigram and syllable bigram. However, this assumption is only an approximation; text access methods based on this ranking function still work very well and have many advantages in retrieving Chinese documents. For example, the ranking function can be easily implemented with a document signature file and bit-slice access, neither demanding extra space overhead, nor affecting retrieval efficiency. The abstract diagram in Fig. 2 shows this a process. In addition, queries can be expressed with non-controlled vocabulary, and the text access method remains efficient even the search terms of the queries are not expressed or recognized exactly. For instance, an input speech query with one or two mis-recognized characters can be tolerated because the corresponding syllables of these error characters might still be correct or their semantics may mostly remain in the extracted query signature. According to our observation, in terms of recall and precision rate, the above function can work very well in most cases.

Full-Text Scanning and Detailed Search

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 in the stage of full-text scanning. As mentioned in Section III, the obtained signature-based similarity values of the remaining documents are re-estimated in this stage. Both character and syllable information of the input query are used again to analyze all of these remaining documents by scanning their contents and using linguistic heuristics. The frequencies of the queried characters, character bigrams, and syllable bigrams of the recognized query 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. In this way, the documents which are highly relevant to the query can be found and displayed.

VI. Experimental Results and Concluding Remarks

Resource Discovery

At present there are two major databases: news database and Web database which are under construction in the resource discovery subsystem for Internet information service.

The news database is a dynamic and instant collection of real-time Chinese news which are automatically obatined from the news groups. In addition to providing real-time news retrieval service for the Internet users, the construction of the database is also designed for developing many natural language processing applications which demand a large size of training corpus such as speech recognition, natural langauge analysis and automatic text classification, etc. The features conceived in the construction of the database are its high accurate information retrieval functions and powerful compilation capcabilities from different sources of news. Since there still have few representitive Chinese news providers on the Interent, except that of CNA and China Time, currently our collected news are most obtained from these two channels. Every day about 800 pieces of news (mostly news abstracts) on average can be obtained, and every hour the system takes several seconds to add the newly acquired news in the database and update the corresponding index files. So far, the total amount of acquired news abtracts in the database exceed 80,000 pieces. At the same time, to provide quasi-natural language queries for the access of the database, a fast signature-based indexing scheme as mentioned before has been adopted for recording the full-text information of the collected news. Meanwhile, for allowing spoken queries and increasing the accuracy of speech recognition, a Markov language model for the news database is also created and instantly updated to construct domain knowledge for the speech recognition subsystem.

On the other hand, the Web database is a collection of abstracted Chinese Web pages on the Internet. The purpose of constructing such a database is similar to Lycos that is planned to act as the cataloging of the Internet Chinese resources. The searching for the preferred web pages is performed by an information spider with automatic traverse started from important Web sites to everywhere. For each preferred web page, a list of abstracted data as that in Lycos will be extracted. Besides, the corresponding document signatures for information retrieval and statistical parameters for language modeling wll be also created. Since the amount of Chinese Web pages is not large and the construction of the Web database is undergoing, so far there are about 5,000 URLs are extracted and completely indexed for the purpose of experiments. According to our schedule, the collection of the database will be obviously enlarged after the retrieval of the database has been thoroughly tested and evaluated recently.

Information Retrieval

The performance of the signature file test has been evaluated. Its achieved retrieval speed is very fast even if more than one gigabyte of texts 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, as our experiments about 99.2% of irrelated documents can be eliminated after the signature file test if common type-in natural language queries (which are not short and consist of meaningful keywords) are given and the news database tested, while the recall rate is 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 number of desired documents need to be further checked. This efficient performance permits us to use more sophisticated linguistic knowledge and text scanning techniques in the continued detailed search module. The performance of the detailed searching is also evaluated. The recall rate for common quasi-natural language queries can remain on the order of 90% and the precision rate can be achieved to 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 a large scale of testing on the news database and web database will be performed recently by more than 10 tester and 1,000 queries, the above testing results are only empirical values obtained from the test of about 100 queries.

Speech Recognition

As described above the speech recognition subsystem is designed based on Golden Mandarin III. Currently, it is a speaker-dependent continuous speech recognition system. Every new user must take about a half hour to train the system and establish his acoustic data model. The training time is planned to be reduced in the next version of the system. Meanwhile, the system has a capability of speaker adaptation and on-line learning. The recogntion accuracy can be increased after several times of use.

In the tests of the preliminary system, a general language model was used in the character string search module and the syllable information was extracted for the test of the signature file. At the same time, the news database collected by the resource discovery subsystem was tested. At that time, the database consisted of a collection of 20,000 news items(about 45MB). A total of 200 unconstrained speech queries with a length of 8.5 characters on average were prepared for testing with 3 speakers. These queries contained from one to three characters irrelevant to the request subject. In order to simplify the examination of the retrieved results, the most highly relevant news item for each test query in the database was restricted to one. The overall performance of the system was evaluated using the hit rates of the top 10 retrieved news. The experimental results are shown in Table 1. In this table, column 2 shows the results of typed queries and columns 3~5 are those of speech queries for three different speakers. The second row shows the character accuracy rates of speech recognition. The third shows the hit rates of the top 10 retrieved news. From this table, it can be observed that the three speakers had similar performance. In addition, the performance obtained using the speech queries was close to the obtained performance using the typed queries. It needs to mention that the test speakers were all surprised by the flexibility of speech queries. They believed it was a good way to perform an interactive process.

Though the obtained results are encouraging, there still exist some difficulties which need to be overcome. For example, the technology of continuous Mandarin speech recognition is not yet robust. For a not-well-trained speaker, the accuracy of the speech recognition and the precision of the retrieved results are not high enough. To obtain more reliable performance, in the present system the domain-specific language model as shown in Fig. 1 is adopted to increase the accuracy of speech recognition. The language model is exactly the set of Markovian parameters extracted from the content of the target database. Combing with the general language model, the recognition accuracy of the system has an improvement with a rate of 3~8%. In special, it can be rather rubust, especailly on the recognition of proper nouns which are often keywords in terms of information retrieval. This improvement encourages us that a Mandarin speech interface for network information retrieval is possible to be established. Since the integration of speech recognition and natural langauge information retrieval technologies can provide users with a convenient initerface for information retrieval. In addition, because Chinese is not alphabetic, speech retrieval of Chinese databases will be a breakthrough if it can be practically used. The present research is important to be continuously engaged in.

References:

1. [Chien'95a] Lee-feng Chien, Fast and Quasi-Natural Language Search for Gigabytes of hineseTexts, *ACM SIGIR ' 95.*

2. [Chien'95b] Lee-feng Chien, Csmart -- A High-Performance Chinese Document Retrieval System, The 1995 International Conference of Computer Processing for Oriental Languages, *ICCPCOL'95.*

3. [Glavitsch'92] Ulrike Glavitsch, Peter Schauble, A System for Retrievaing Speech Documents, *ACM SIGIR'92.*

4. [Kimmel'96] S. Kimmel, "Robot-generated Databases on the World Wide Web", Database, 1996, p 41-49.

5.[Lee'91] Lee, Lin-Shan, Chien, Lee-feng, Chen, K., J., et al., An Efficient Natural Language Processing System Specially Designed for the Chinese Language, *Computational Linguistics*, Dec., 1991.

6. [Lee'95] Lin-shan Lee, Lee-feng Chien, et al., Golden Mandarin III: A Prosodic Segment based Mandarin Speech Recogntion System, Proceedings of the 1995 IEEE International Conference on Acoustics, Speech and Signal Processing, *ICASSP'95.*

7. [Liang'94] Liang, Tyne, Lee, S. Y. and Yang, W. P. On the Design of Effective Chinese textual Retrieval Based On the Signature Method, *Computer Processing of Chinese and Oriental Languages*, Vol. 8, No. 1, June 1994, pp. 87-96.

8.[Lin'96] Sung-chien Lin, 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*.

9.[Lin'95a] Sung-chien Lin, 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*.

10.[Lin'95b] Sung-chien Lin, 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*.

11.[Lycos] Lycos HomePage (http://www.lycos.com/)

12[Salton'83] Salton, G., Introduction to Modern Information Retrieval, NY,
McGraw-Hill, 1983.

13.[Wu'95]S. Wu, Gais Home Page, Http://gais.cs.ccu.edu.tw/

14.[Wu'94]Wu, Zimin and Tseng, Gwyneth, Chinese Text Segmentation for Text Retrieval; Achievements and Problems. *JASIS*, 44(9), 1994, 532-542.

15.[Yamada'94] M. Yamada, F. Itoh, K. Sakai et al., A Spoken Dialogue System with Active/Non-active Word Control for CD-ROM Information Retrieval, *Speech Communication*,1994, 355-365.

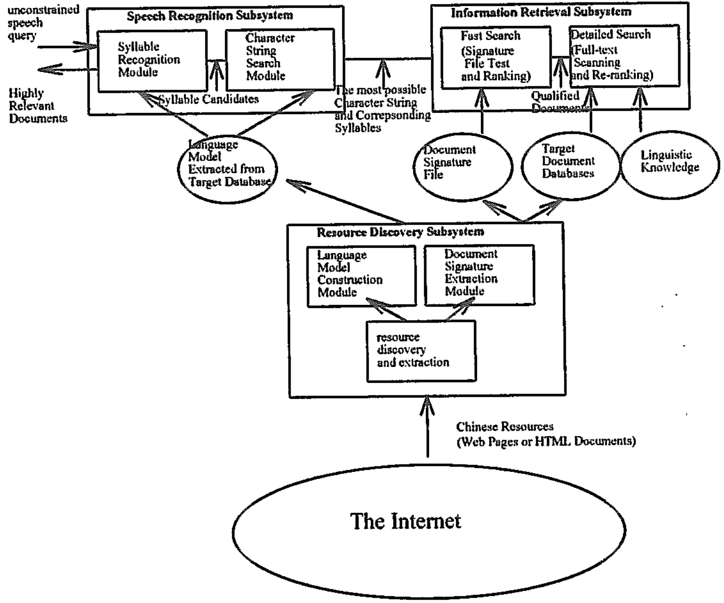
unconstrained speech query
Highly Relevant Documents
Speech Recognition Subsystem
Syllable Recognition Module
Character String Search Module
Syllable Candidates
The most possible Character String and Correpsonding Syllables
Information Retrieval Subsystem
Fast Search (Signature File Test and Ranking)
Detailed Search (Full-text Scanning and Re-ranking)
Qualified Documents
Language Model Extracted from Target Database
Document Signature File
Target Document Databases
Linguistic Knowledge
Resource Discovery Subsystem
Language Model Construction Module
Document Signature Extraction Module
resource discovery and extraction
Chinese Resources (Web Pages or HTML Documents)
The Internet

Fig. 1 The block diagram of the overall system.

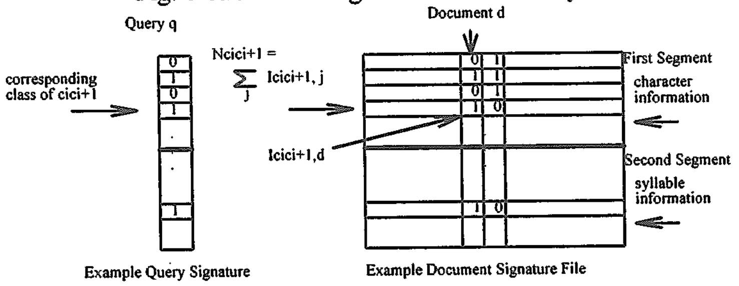

Fig. 2 An abstract diagram which showing the process of signature-based ranking with a bit-slice document signature file, in which the required parameters, e.g, N_{cici+1}, can be calculated without additional space overhead.

	Type-in Query	Speaker A	Speaker B	Speaker C
Character Accuracy in Speech Recognition	~	88.4%	91,3%	82.1%
Hit Rates of Top 10 Retrieved News	90.5	82.3	82.6	80.4

Table 1. The experimental results for the present system.