Combination and Boundary Detection Approaches on Chinese Indexing

Christopher C. Yang,* Johnny W.K. Luk, Stanley K. Yung, and Jerome Yen
Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong. E-mail: yang@se.cuhk.edu.hk

Digital libraries store materials in electronic format. Research and development in digital libraries includes content creation, conversion, indexing, organization, and dissemination. The key technological issues are how to search and display desired selections from and across large collections effectively [Schatz & Chen, 1996]. Digital library research projects (DLI-1) sponsored by NSF/DARPA/NASA have a common theme of bringing search to the net, which is the flagship research effort for the National Information Infrastructure (NII) in the United States. A repository is an indexed collection of objects. Indexing is an important task for searching. The better the indexing, the better the searching result. Developing a universal digital library has been the dream of many researchers, however, there are still many problems to be solved before such a vision is fulfilled. The most critical is to support a cross-lingual retrieval or multilingual digital library. Much work has been done on English information retrieval, however, there is relatively less work on Chinese information retrieval. In this article, we focus on Chinese indexing, which is the foundation of Chinese and cross-lingual information retrieval. The smallest indexing units in Chinese digital libraries are words, while the smallest units in a Chinese sentence are characters. However, Chinese text has no delimiter to mark word boundaries as it is in English text. In English or other languages using Roman or Greek-based orthographies, often, spacing reliably indicates word boundaries. In Chinese, a number of characters are placed together without any delimiters indicating the boundaries between consecutive characters. In this article, we investigate the combination and boundary detection approaches based on mutual information for segmentation. The combination approach combines *n*-grams to form words with more number of characters. In the combination approach Algorithm 1 does not allow overlapping of *n*-grams while Algorithm 2 does. The boundary detection approach detects the segmentation points on a sentence based on the values and the change of values of the mutual information. Experiments are conducted to evaluate their performances. An interface of the system is also presented to show how a Chinese web page is downloaded, the text in the page filtered, and segmented into words. The segmented words can be submitted for indexing or new unknown words can be identified and submitted to a dictionary.

1. Introduction

The research in a digital library is focused on developing tools, technologies, and concepts to use the inherent meaning and knowledge in digital collections effectively. For users, an intelligent search and retrieval of the useful information is desired. In order to achieve successful searching, indexing on the electronic documents is important.

Indexing has been one of the most important research issues in information science. Without proper indexes, search and retrieval of information are impossible. A traditional indexer recognizes and selects the essence and then represents it. However, traditional indexing is time consuming and expensive. An automatic indexer is desired and it is critical to the success of the generation of concept spaces [Schatz & Chen, 1996; Schatz, Mischo, Cole, Hardin, Bishop, & Chen, 1996; Schatz, Mischo, Cole, et al., 1999]. In order to index a document, extracted concepts of the document are input pattern to a concept space, which is represented as a Hopfield network. The parallel spreading activation process produces a new set of concepts that are relevant to the concepts of the input document [Chen, Ng, Martinez, & Schatz, 1997]. Much research has been done on English indexing, while it is relatively less so in Chinese indexing. An efficient and effective Chinese segmentation is essential for the success of Chinese indexing.

The population in Mainland China is 1.2 billion and there are Chinese living all over the world. For some major cities in North America, such as Vancouver, Toronto, and San Francisco, the Chinese may not be considered a minority anymore. The number of users and collections providers of Chinese information in the world is unimaginable. As a result, the need for research and development in Chinese information retrieval and a digital library is obvious.

* Corresponding author.

Accepted July 2, 1999.

Cross-lingual information retrieval has been one of the focuses of digital library research recently. The DLI-2 is actively promoting activities and processes that cross the boundaries of language and politics. A new program in International Digital Libraries Collaborative Research has been initiated [Schatz & Chen, 1999]. Based on these reasons, Chinese indexing and information retrieval is crucial for the future development of cross-lingual information retrieval between Chinese languages and other western and Asian languages.

Information retrieval in English has over 30 years of history, however, information retrieval in Chinese is relatively recent [Ikehara, Shirai, & Kawaoka, 1995]. Written Chinese consists of strings of characters (or ideographs) separated by punctuation. A character can perform as a word with meaning(s) or function as an alphabet forming a short word with one or more adjacent characters and having specific meaning [Ikehara, Shirai, & Kawaoka, 1995]. The first step of word-based indexing is segmentation. Segmentation is the process of breaking a string of characters into words. In other words, segmentation is determining the boundaries of single or multicharacter words in a string. Unlike English, written Chinese does not have explicit word boundaries. Word segmentation in Chinese is known to be a difficult task, especially for unknown words, such as names, locations, translated terms, technical terms, abbreviations, etc.

Previous work on Chinese word segmentation can be divided into three categories, statistical approach [Gan, Palmer, & Lua, 1996; Lua, 1990; Schatz & Chen, 1996; Sproat & Shih, 1990], a lexical rule-based approach [Nie, Jin, & Hannaan, 1994; Wang, Pei, Li, & Huang, 1900; Wu & Tseng, 1995], and a hybrid approach that is based on statistical and lexical information [Leung & Kan, 1996; Nie, Jin, & Hannaan, 1994]. In newspaper articles, names, places, organizational and historical events, and specialized terms are very common. These words are considered unknown words that do not appear in dictionaries. Unknown words, such as names, have little linguistic constraints, and hence, a lexical approach may not be appropriate. The statistical approach is particularly applicable to identify unknown words. In this paper, we investigate two approaches based on statistics, the combination approach and the boundary detection approach. Both of these approaches use mutual information between adjacent Chinese characters or *n*-grams. Experiments are conducted to evaluate how these approaches perform in segmentation on newspaper articles on the web. Experimental results show that the proposed boundary detection approach performs well in segmenting the unknown words. A system interface is also presented to show how web-based news articles are downloaded, segmented, and submitted for further processing in the Chinese information retrieval system.

2. Word Segmentation

There are three major approaches to Chinese segmentation, the statistical approach, the lexical rule-based approach, and the hybrid approach based on statistical and lexical information.

2.1 Statistical Approach

Sproat and Shih [1990] uses mutual information as a statistical measurement of any two characters. Two adjacent characters with the highest value of mutual information are extracted first. Bi-grams are recursively extracted until no more pairs can be extracted. A lower limit is set, below which association strength characters are not grouped. Rigorous experiments are conducted and 94% of the segmentation is correct. The failure occurs on common words with insufficient representation in corpus, unpopular technical terms, styles not well represented in corpus such as fortune-telling, and extracting wrong bi-grams out of several possible groupings. Sproat and Shih's technique only extracts bi-grams, however, most of the unknown words such as names and technical terms contain more than two characters and long words are often used for indexing.

Lua [1990] investigates the formation of Chinese words from characters by observing the change in information content. The concept is similar to energy change of a molecule from its constituent atoms. A gain in information content is usually associated with significant deviation from the meaning of its composing characters. A loss in information content often indicates little change in the meaning and strong binding between the characters. The loss in information content for tri-grams and quadra-grams are more significant than the loss in information content for bi-grams. However, it is not due to a change in meaning but simply due to more numbers of characters in the words.

Chien [1997] developed a PAT-tree-based approach for keyword extraction. All of the lexical patterns without limitation of pattern length are first extracted. A mutual-information-based filtering algorithm is then applied to filter out the character strings in the PAT tree. A refined method based on a common-word lexicon, a general domain corpus and a keyword determination strategy are utilized finally. The performance is good but building a PAT tree is time consuming and large space overhead is required. In this paper, we are investigating a more efficient technique to segment a Chinese sentence.

Several Japanese researchers have also applied statistical techniques for segmentation and keyword extractions for Japanese documents [Ikehara, Shirai, & Kawaoka, 1995; Ogawa & Iwasaki, 1995]. Frequency data were used to index *n*-grams and extract collocations from very large Japanese corpora.

2.2 Lexical Rule-Based Approach and Hybrid Approach Using Lexical and Statistical Information

The major concerns of lexical rule-based approach are how to deal with ambiguities in segmentation, and how to extend the lexicon beyond dictionary entries. The lexical rule-based approach that deals with mentation ambiguities

is also known as the dictionary-based approach and the most popular method is the maximum matching method. Starting from the beginning or end of a sentence, the maximum matching method groups the longest initial sequence of characters that matches a dictionary entry as a word, these are called forward maximum matching and backward maximum matching. The idea is to minimize the number of words segmented. A variation of the maximum matching is the minimum matching or shortest matching, which treats a word as the shortest initial string of characters that match a dictionary entry. Leun and Kan [1996] have developed a parallel maximum matching algorithm that increases the execution speed. Some other techniques in the lexical rule-based approach depend on constraint satisfaction based on syntactic or semantic [Yeh & Lee, 1900] and lexical heuristics [Chen & Liu, 1992]. For example, Yeh and Lee use a unification-based approach. Some others depend on lexical heuristics.

Many researchers have also developed techniques based on the hybrid approach using lexical and statistical information. For example, Nie and colleagues [1994, 1995] first segment the text using the maximum-matching approach and then utilize the statistical method to locate and propose candidates for the unknown words contained in the segmented results of the maximum-matching method (see Fan and Tsai's work using relaxation technique [1988]).

In this article, we focus on the statistical approach using mutual information.

3. Mutual Information

Mutual information $I(a, b)$ is the statistical measurement of association between two events, a and b. In Chinese segmentation, mutual information, $I(c_1, c_2)$, measures association between two consecutive characters, c_1 and c_2, in a sentence. Characters that are highly associated are considered to be grouped together to form words.

Equation (1) shows the computation of mutual information of event a and b, where $P(a, b)$ denotes joint probability of events a and b, and $P(a)$ and $P(b)$ denote probabilities of event a and event b.

$$I(a, b) = \log_2\left(\frac{P[a, b]}{P[a]P[b]}\right). \quad (1)$$

For Chinese segmentation, probability of character, c_i, is the frequency of $c_i(f[c_i])$ divided by the total number of characters in the corpus, N. Joint probability of two characters, c_i and c_j, is the frequency of c_i followed by $c_j(f[c_i, c_j])$ divided by N. Therefore, mutual information of c_i and c_j is computed as follows:

$$I(c_i, c_j) = \log_2\left(\frac{\frac{f(c_i, c_j)}{N}}{\frac{f(c_i)}{N}\frac{f(c_j)}{N}}\right) = \log_2\left(\frac{Nf[c_i, c_j]}{f[c_i]f[c_j]}\right). \quad (2)$$

Mutual information of two characters shows how strongly these characters associated with one another. If the characters are independent to one another, $P(c_i, c_j)$ equals to $P(c_i)P(c_j)$. Substituting into Eq. 2, $I(c_i, c_j)$ equals to 0. If c_i and c_j are highly correlated, $I(c_i, c_j)$ increase.

Similarly, computing mutual information between bi-gram and uni-gram ($[c_ic_j][c_k]$ or $[c_i][c_jc_k]$) is as follows:

$$I(c_ic_j, c_k) = \log_2\left(\frac{Nf[c_ic_jc_k]}{f[c_ic_j]f[c_k]}\right) \quad \text{and}$$

$$I(c_i, c_jc_k) = \log_2\left(\frac{Nf[c_ic_jc_k]}{f[c_i]f[c_jc_k]}\right). \quad (3)$$

Computing mutual information between bi-gram and bi-gram ($[c_ic_j][c_kc_l]$) is as follows:

$$I(c_ic_j, c_kc_l) = \log_2\left(\frac{Nf[c_ic_jc_kc_l]}{f[c_ic_j]f[c_kc_l]}\right) \quad (4)$$

Computing mutual information between tri-gram and uni-gram ($[c_ic_jc_k][c_l]$ or $([c_i][c_jc_kc_l]$) is as follows:

$$I(c_ic_jc_k, c_l) = \log_2\left(\frac{Nf[c_ic_jc_kc_l]}{f[c_ic_jc_k]f[c_l]}\right) \quad \text{and}$$

$$I(c_i, c_jc_kc_l) = \log_2\left(\frac{Nf[c_ic_jc_kc_l]}{f[c_i]f[c_jc_kc_l]}\right). \quad (5)$$

4. Segmentation Based on Combination and Boundary Detection

We investigate two approaches of segmentation, combination and boundary detection. In the previous studies [Chen, He, Xu, Gey, & Meggs, 1997; Sproat & Shih, 1990], mutual information is used to extract bi-grams, *n*-grams with more than two characters are not segmented. Although approximately 70% of Chinese words are bi-grams [Sproat & Shih, 1990], most of the names, places, organizational and historical events, and specialized terms in various fields of studies that are used for indexing are not bi-grams. These words usually contain three characters or more. Techniques for segmenting *n*-grams with more than two characters are desired.

The first approach is based on the combination of segmented *n*-grams to form words with numbers of characters more than *n.* In this approach, we develop two algorithms. The second algorithm allows overlapping in the segmented *n*-grams, but the first algorithm does not allow so.

The second approach is based on boundary detection. The boundary of a word usually occurs at the mutual information value lower than a threshold and/or local minimum of mutual information values in a sentence. For edge detection in image analysis, a pixel is considered an edge point (or boundary) if there is an abrupt gray-level change. Gradient operators, such as Roberts operator [Roberts, 1977].

Smoothed operator [Davis, 1975], and Sobel operator [Prewitt, 1970], are used in image edge detection. In our boundary detection approach for Chinese segmentation, we use the mutual information and the analogy of the edge detection technique on images. If the mutual information value is less than a threshold, it means the two characters are independent and therefore it is considered a segmentation point. Similar to gradient operators in image edge detection, change in mutual information value is used to detect the points of valley and bowl shape curve. These abrupt changes in mutual information values are employed to detect the segmentation points.

4.1 Combination

Algorithm 1

1. *Counting occurrence frequencies*
   Obtain occurrence frequencies for all possible *n*-grams, for $n = 1$ to 4.
2. *Extracting bi-grams*
   Compute mutual information for all *n*-grams. Determine the bi-grams with the highest mutual information value and remove it from the sentence. Repeat the removal of bi-grams until no more bi-gram existing in the sentence or the mutual information values are less than a threshold, T_1 [Sproat & Shih, 1990].
3. *Combining the extracted bi-grams and uni-grams to form tri-grams*
   Compute the mutual information for all possible combinations of uni-gram and bi-gram to form tri-grams, i.e., (bi-gram, uni-gram) or (uni-gram, bi-gram). Combine the bi-grams and uni-grams with the highest mutual information value until no such patterns exist in the sentence or the mutual information values are less than a threshold, T_2.
4. *Combining the extracted tri-grams, bi-grams and uni-grams to form quadra-grams*
   Compute the mutual information for all possible combinations of uni-grams, bi-grams and tri-grams to form quadra-grams, i.e., [uni-gram, tri-gram], or [bi-gram, bigram], [tri-gram, uni-gram]. Combine the uni-grams, bi-grams, or tri-grams with the highest mutual information value until no such patterns exist in the sentence or the mutual information values are less than a threshold, T_3.

Algorithm 2

Steps 1 and 2 are the same as those in Algorithm 1. However, in Steps 3 and 4, portions of the extracted bi-grams and tri-grams are allowed to combine with other tri-grams, bi-grams, and uni-grams. For example, given two consecutive bi-grams, (c_1c_2) and (c_3c_4), it allows to combine (c_2) and (c_3c_4) or (c_1c_2) and (c_3) to form tri-grams $(c_1c_2c_3)$ or $(c_2c_3c_4)$. Although the portion of an *n*-gram, *A*, is combined with another *n*-gram, *B*, to form a new *n*-gram, *C*, the original *n*-gram, *A*, is still saved as a segmented *n*-gram in the sentence. For example, $[c_1c_2]$ and $[c_3]$ is combined from two bi-grams (c_1c_2) and (c_3c_4) to form a tri-gram $(c_1c_2c_3)$, the segmented *n*-grams in the sentence are $(c_1c_2c_3)$, (c_1c_2) and (c_3c_4).

In some conditions, a portion of an *n*-gram may combine with another *n*-gram to form such a word that the two original *n*-grams and the combined *n*-gram are all valid words. For example, given two bi-grams, (愛 滋) (Acquired Immune Deficiency Syndrome's [AIDS] Chinese abbreviation) and (病 毒) (virus), (愛 滋) (AIDS's Chinese abbreviation) can combine with the first character, (病) (disease), of the second bi-gram, (病 毒) (virus), to form a tri-gram (愛 滋 病) (AIDS's Chinese translation).

The following example is used to illustrate how Algorithm 1 and Algorithm 2 are processed.

Example

Sentence:	主	禮	嘉	賓	將	包	括	國	家	主	席
$I(c_i,c_j)$	1.87	4.67	9.26	1.49	1.14	10.38	0.51	5.10	1.83	7.60	

Sentence in English: The guests ceremony will include the country chairman.

Algorithm 1

In Step 2, the sentence is segmented to (主禮) (嘉賓) (將) (包括) (國家) (主席) ([ceremony] [guests] [will] [include] [country] [chairman]). In Step 3, there are only two possible tri-grams that can be combined from the bi-grams and uni-grams obtained from Step 2, i.e., (嘉賓) (將) ([guest] [will]) and (將) (包括) ([will] [include]). The mutual information value of (嘉賓) (將) is higher than that of (將) (包括), however, both mutual information values are less than the thresholds of 5.0, (嘉賓) and (將) are not combined. The sentence is not further combined in this step. In Step 4, (主禮) (嘉賓) ([ceremony] [guest]), (包括) (國家) ([include] [country]), (國家) (主席) ([country] [chairman]) are the possible combinations to form quadra-grams. Given 5.0 as the threshold, the sentence is segmented (主禮嘉賓) (將) (包括) (國家) (主席) ([The guest of ceremony] [will] [include] [country] [chairman]).

Algorithm 2

Steps 1 and 2 in this algorithm are the same as those in Algorithm 1. Possible tri-grams that can be combined from the bi-grams and uni-grams in Step 2 are (主禮) (嘉), (禮) (嘉賓), (嘉賓) (將), (將) (包括), (包括) (國), (括) (國家), (國家) (主), (家) (主席). The possible combinations are four times more than those in Algorithm 1. If the threshold is 5.0, the sentence is segmented to (主禮) (主禮嘉) (嘉賓) (將) (包括) (國家) (主席) in Step 3. Possible quadra-grams that can be combined from tri-grams, bi-grams, and uni-grams are (主 禮) (嘉 賓), (包 括) (國 家), (國 家) (主 席). The sentence is segmented to (主禮 嘉) (主禮 嘉賓) (將) (包括) (國 家) (主 席) in Step 4.

4.2 Boundary Detection

1. *Counting occurrence frequencies*
   Obtain occurrence frequencies for all uni-grams and bi-grams.

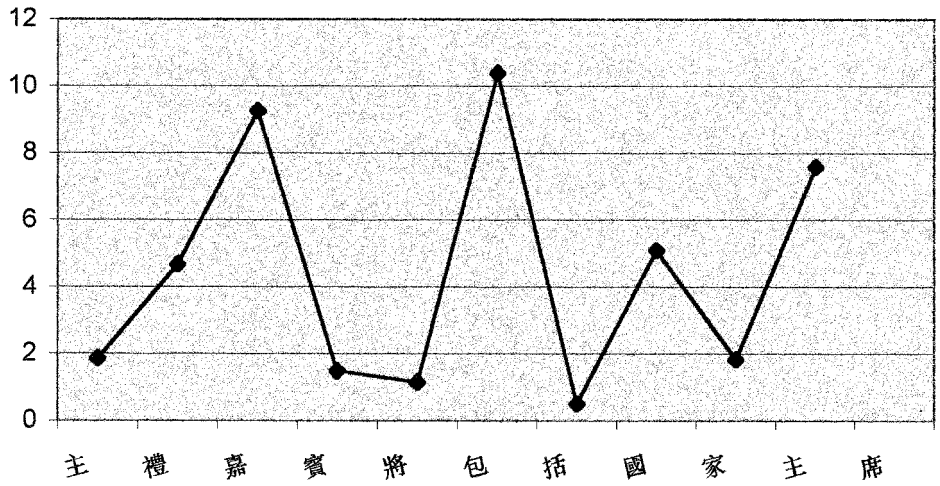

FIG. 1. Mutual information values of the sentence 主 禮 嘉 賓 將 括 國 家 主 席.

2. Compute mutual information for all bi-grams

3. Determine the segmentation points

a. If the mutual information value for a bi-gram is less than a threshold, T_1, the point between the two characters in the bi-gram is treated as the segmentation point. T_1 is greater than or equal to 0.

b. Given a string of characters . . . , c_{j-1}, c_j, c_{j+1}, c_{j+2}, c_{j+3} . . . ,

Determine the valley point:

If $I(c_{j-1}, c_j) > I(c_j, c_{j+1})$ and $I(c_{j+1}, c_{j+2}) > I(c_j, c_{j+1})$

Then the point between c_j and c_{j+1} is a valley point and the point is treated as a segmentation point

Determine the points of bowl shape curve;

If $I(c_j, c_{j+1}) - I(c_{j-1}, c_j) < 0$ and $I(c_{j+2}, c_{j+3}) - I(c_{j+1}, c_{j+2}) > 0$ and

$$\frac{I[c_{j-1}, c_j] - I[c_j, c_{j+1}]}{|I[c_j, c_{j+1}] - I[c_{j+1}, c_{j+2}]|} > T_2 \text{ and}$$

$$\frac{I[c_{j+2}, c_{j+3}] - I[c_{j+1}, c_{j+2}]}{|I[c_j, c_{j+1}] - I[c_{j+1}, c_{j+2}]|} > T_2$$

where T_2 is a threshold

Then the points between c_j and c_{j+1} and between c_{j+1} and c_{j+2} are points of bowl shape curve, these points are treated as a segmentation points

For example, given the following sentence and the computed mutual information values,

Example 1

Sentence:	中	國	大	陸	新	發	現	的	油	田
$I(c_i,c_j)$		4.69	0.18	6.13	-1.46	-0.30	4.07	0.00	-0.30	7.87

Sentence in English: The newly discovered oil field in Mainland China . . .

If $T_1 = 0$, the sentence is segmented to (中國大陸) (新) (發現) (的) (油田) ([Mainland China] [new] [discovered] [*adjective marker*] [oil field]). If $T_1 = 0.5$, the sentence is segmented to (中國) (大陸) (新) (發現) (的) (油田) ([China] [Mainland] [new] [discovered] [*adjective marker*] [oil field]). The smallest the threshold is, the largest the segments' size is after processing.

Example 2

Sentence:	主	禮	嘉	賓	將	包	括	國	家	主	席
$I(c_i,c_j)$		1.87	4.67	9.26	1.49	1.14	10.38	0.51	5.10	1.83	7.60

Sentence in English: The guests of ceremony will include the country chairman.

If $T_1 = 1.0$, one segmentation point is found between 括 and 國. The sentence is segmented to (主禮嘉賓將包括) (國家主席). Three valley points are detected and the sentence is further segmented to (主禮嘉賓將) (包括) (國家) (主席) ([The guest of ceremony will] [include] [the country] [chairman]). In addition, two points of bowl shape curve are found where one of these two points has already been detected as valley points. The sentence is further segmented to (主禮嘉賓) (將) (包括) (國家) (主席) ([The guests of ceremony] [will] [include] [the country] [chairman]).

As discussed in Section 3, if two characters are independent to one another, the mutual information equals to 0. The threshold, T_1, is set as 1 in our experiments (Section 5). The value of 1 is determined experimentally. We also observe that, even if T_1 is set to 0, many of the missing segmentation points will be determined as points of valley or bowl shape curve. Setting the value of T_1 between 0 and 1 does not change the segmentation result significantly.

T_2 is used to determine the segmentation points for the bowl shape curve. If the changes of mutual information between $I(c_{j-1}, c_j)$ and $I(c_j, c_{j+1})$ and between $I(c_{j+2}, c_{j+3})$ and $I(c_{j+1}, c_{j+2})$ are significantly larger than the change of mutual information between $I(c_j, c_{j+1})$ and $I(c_{j+1}, c_{j+2})$, the points between c_j and c_{j+1} and between c_{j+1} and c_{j+2} are considered as segmentation points. Experimentally, we have set T_2 as 2. The bowl shape curve is

Table 1. *N*-gram size in testbed.

n-gram	Number of distinct *n*-grams	Number of *n*-grams
Uni-gram	4727	3,570,523
Bi-gram	499,469	3,213,203
Tri-gram	1,763,629	2,926,959
Quadra-gram	4,456,723	2,710,034

good at retrieving the uni-grams that are adjective makers and always attached before or after other *n*-grams, e.g., 的 is an adjective maker (Figure 1).

5. Experiment

To evaluate the performance of the two approaches, an experiment is conducted. A training corpus is made up of Hong Kong local news articles collected from three local newspaper web sites, which include Apple Daily, Ming Pao, and Sing Tao Daily. The size of the collection is about 15 Mbytes. There are a total of 3514 documents and 3,570,523 characters in the testbed. The occurrence frequencies of all uni-grams, bi-grams, tri-grams, and quadra-grams are obtained. Table 1 shows the number of *n*-grams collected in the testbed. We perform the segmentation using the combination and boundary detection approaches and their results are compared.

To measure the accuracy, all the documents are first segmented manually, the manual segmentation results are then compared with the segmentation results using the combination approach and boundary detection approach. The accuracy measures the number of correctly segmented *n*-grams by the algorithms over the total number of manually segmented *n*-grams in all the documents. The correctness of segmentation is determined syntactically, semantically, and contextually. Table 2 shows the results of the accuracy. The boundary detection approach has the highest accuracy (92%) while the accuracy of the algorithms using the combination approach are only 84% and 83%.

The efficiency of the combination approach and boundary detection approach is also considered in our experiment. We conduct the experiment using a Pentium II 300 MHz machine with 256 MB RAM. The clock time of segmentation for each document using the combination approach, Algorithm 1 and Algorithm 2, and boundary detection approach are recorded. The average clock time taken by each algorithm for all the documents are computed and presented in Table 3. The boundary detection approach is significantly faster than the combination approach. The boundary detection approach takes only 4.2 seconds in average while Algorithm 1 and Algorithm 2 in combination approach take 10.4 seconds and 14.5 seconds in average, respectively. The boundary detection approach only utilizes the frequency tables of unigrams and bi-grams. However, the combination approach utilizes the frequency tables of unigrams, bi-grams, tri-grams, and quadra-grams. The combination approach is more time consuming in retrieving data from the frequency tables. The procedures of combining unigrams, bi-grams and tri-grams are also more tedious than recognizing the valley and bowl shape patterns in the boundary detection approach.

As a conclusion, we find that the boundary detection approach is more efficient and accurate than the combination approach.

Table 2. Accuracy of the combination approach and boundary detection approach on segmentation.

Algorithm	Accuracy (%)
Combination approach—Algorithm 1	84
Combination approach—Algorithm 2	83
Boundary detection approach	92

Table 3. Average speed (clock time) of the combination approach and boundary detection approach on segmenting a document of approximately 1000 characters.

Algorithm	Average speed
Combination approach—Algorithm 1	10.4seconds
Combination approach—Algorithm 2	14.5seconds
Boundary detection approach	4.2seconds

5.1 Discussion

Both Algorithm 1 and Algorithm 2 in the combination approach produce fair segmentation results. However, noise segments are produced in Algorithm 2 as by-products and some segmented words can indeed be decomposed into shorter word phrases in both algorithms. In some cases, wrong segments are also obtained.

Example:

Sentence: 主 禮 嘉 賓 將 包 括 國 家 主 席 江 澤 民 及 副 總 理 錢 其 琛

Sentence in English: The guests of ceremony will include the chairman of country, Jiang Zemin, and vice-president, Qian Qichen.

Using Algorithm 1, the sentence is segmented to (主 禮 嘉 賓) (將) (包 括) (國 家) (主 席) (江 澤 民) (及) (副 總) (理 錢 其 琛) (The guests of ceremony include the country chairman, Jiang Zemin, and vice president, Qian Qichen). Wrong segments (副 總) (理 錢 其 琛) are formed because bi-grams (副 總) (理 錢) (其 琛) are formed in Step 2, where (理 錢) is a wrong segment. 理 and 錢 are extracted in Step 2 because (副 總) and (其 琛) are already extracted and the mutual information of the only possible combination left, 理 and 錢 is higher than the threshold. The correct segments are (副 總 理) (vice president) and (錢 其 琛) (Qian Qichen).

Using Algorithm 2, the sentence is segmented to (主 禮 嘉) (主 禮 嘉 賓) (guests of ceremony) (將) (will)

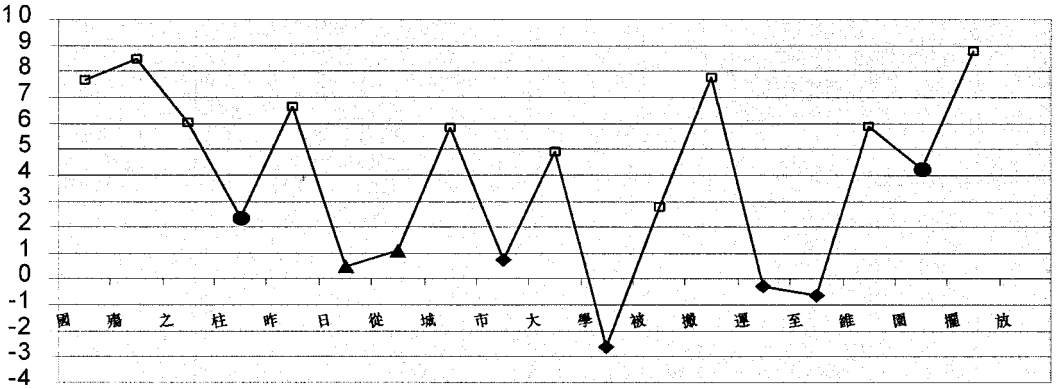

FIG. 2. Mutual information values of the sentence 國殤之柱昨日從城市大學被搬運至維園擺放.

(包括) (include) (國家主席) (chairman of country) (江澤民) (Jiang Zemin) (及) (and) (副總) (理錢) (其琛) (副總理) (vice president) (錢其琛) (Qian Qichen) (副總理錢). Noise segments (主禮嘉) (副總) (理錢) (其琛) (副總理錢) are also segmented although the correct segments, (副總理) (vice president) and (錢其琛) (Qian Qichen) that are not segmented by Algorithm 1, are retrieved.

Example:

Sentence: 搬運至維園擺放

Sentence in English: move to and place at the Victoria Park.

Using Algorithm 1, the sentence is segmented to (搬運) (至) (維園擺放) ([move] [to] [place at the Victoria Park]). Indeed, (維園擺放) should be segmented as (維園) (擺放) because 維園 is the name of a park and 擺放 is a verb meaning "place at." It is combined in Step 4 because the mutual information of (維園) and (擺放) is higher than the threshold.

Wrong segmentation and word phrases are not able to decompose into shorter word phrases using the combination approach because it is difficult to obtain an appropriate threshold in Steps 3 and 4 of the algorithms. The average of the mutual information values for combining bi-gram and uni-gram or bi-gram and bi-gram or tri-gram and uni-gram are relatively higher than the mutual information values for combining uni-gram and uni-gram. It is because the frequency of a bi-gram is usually lower than the frequency of an uni-gram in a corpus. In calculating the mutual information for two adjacent uni-grams, we divide the probability of the combined bi-gram by the product of the probabilities of two uni-grams. In calculating the mutual information for other combination of uni-gram, bi-gram or tri-gram, we divide the probability of the combined *n*-grams by the product of the probabilities of uni-gram, bi-gram, or tri-gram. Determining an appropriate threshold in Step 3 and Step 4 for both Algorithm 1 and Algorithm 2 is not easy. In some cases, uni-gram and bi-gram, bi-gram and bi-gram, tri-gram and uni-gram are not supposed to be combined. They are combined because their mutual information values are higher than the threshold. However, if a higher threshold is selected, some other combinations, which are supposed to be combined, may not be able to form a word phrase. This difficulty does not exist in the boundary detection approach because the boundary detection approach considers not only threshold on mutual information but also the changes in mutual information.

The boundary detection approach produces good segmentation result. The accuracy is 92%. In Figures 2 to 7, ◆ represents the segmentation points detected by threshold, ● represents the valley points, and ▲ represents points of bowl shape curve. Inaccuracy may occur in the following situations:

- Missing segmentation points occur when there is an uni-gram located between two *n*-grams, e.g., (方向) (的) (中間) ([direction] [*adjective marker*] [center]) is segmented as (方向) (的中間). Only one segmentation point, between 向 and 的, is detected but another segmentation point, between 的 and 中, is missed.
- If the first character or the last character is an uni-gram, it may not be able to be segmented because the left (right) reference point is not available for the first (last) character.
- Some names are segmented into smaller segments because the names are composed of words that always appear separately.
- Overall, the performance on extracting unknown words is excellent. We have tested with several unknown names, events, and terms in the corpus. Only a few of them cannot be extracted. However, these missed segments could be retrieved when the size of the corpus increases.

For the sentence in Figure 2, 國殤之柱昨日從城市大學被搬運至維園擺放 (The Pillar of Shame has been moved from the City University and placed at the Victoria Park yesterday), it is segmented to (國殤之柱) (昨日) (從) (城市) (大學) (被搬運) (至) (維園) (擺放) ([The Pillar of Shame] [yesterday] [from] [City] [University] [is moved] [to] [Victoria Park] [placed]). (城市) (大學) should be combined since 城市大學 is the name of a university, however, both 城市 (city) and 大學 (university) are words that appears as separate words frequently. It is difficult to determine 城市大學 as a word using only mutual information. Yet, the name of the statue, 國殤之柱 (the Pillar of Shame), which was made as a memorial for the June 4th event that happened at Beijing in 1989, are segmented without any problems.

For the sentence in Figure 3, 長江實業於上周六開售的青衣機鐵站盈翠半島 (The Cheung Kong (Holdings) Limited are selling the Tierra Verde at Tsing Yi

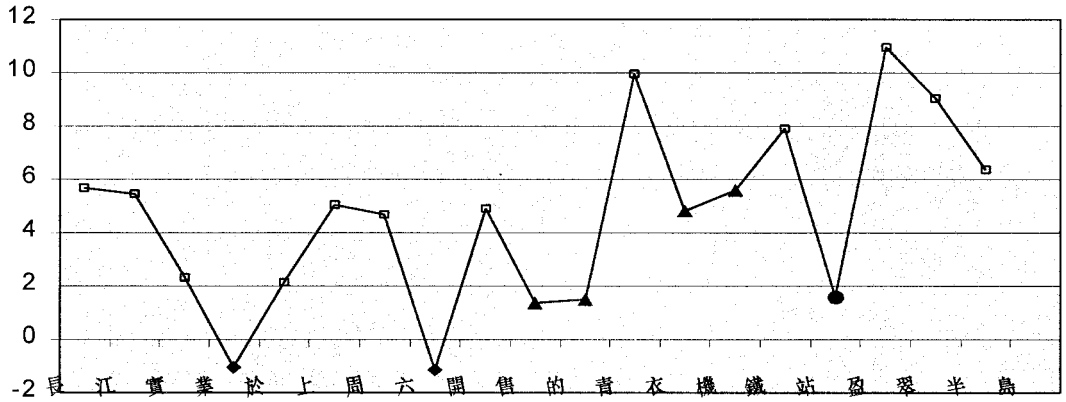

FIG. 3. Mutual information values of the sentence 長江實業於上周六開售的青衣機鐵站盈翠半島.

Mass Transit Railway station on last Saturday.), it is segmented to (長 江 實 業) (於 上 周 六) (開 售) (的) (青 衣) (機 鐵 站) (盈 翠 半 島) ([Cheung Kong (Holdings) Limited] [on last Saturday] [sell] [*adjective marker*] [Tsing Yi] [Mass Transit Railway station] [Tierra Verde]). Unknown words, such as, company name, 長 江 實 業 (Cheung Kong (Holdings) Limited), location name, 青 衣 (Tsing Yi), and transportation name, 機 鐵 站 (MTR station), which are not available in the dictionary have been successfully segmented.

For the sentence in Figure 4, 他 在 黃 大 仙 地 鐵 站 一 號 月 台 往 觀 塘 方 向 的 中 間 位 置 (He is located at Wong Ta Shin MTR station Platform Number One's central location heading to Kwung Tong direction), it is segmented to (他 在) (黃 大 仙) (地 鐵 站) (一 號) (月 台) (往 觀 塘) (方 向) (的 中 間) (位 置) ([He] [at] [Wong Ta Sin] [MTR station] [number one] [platform] [to Kwung Tong] [direction] [central] [location]). The first character, 他 (He), should be segmented. However, it is not able to be retrieved because the mutual information of 他 在 is barely higher than the threshold, 1.0. Moreover, it does not have a reference point on its left, therefore, it is not able to determine it as a valley point or a point on the bowl shape curve.

For the sentence in Figure 5, 各 車 行 近 日 引 入 的 東 南 亞 版 平 治 及 寶 馬 等 房 車 (Every dealers import south east Asia version of Benz and BMW etc. cars lately.), it is segmented to (各) (車 行) (近 日) (引 入) (的) (東 南 亞 版) (平 治 及) (寶 馬) (等) (房 車) ([Every] [dealers] [lately] [import] [*adjective marker*] [south east Asia version] [Benz and] [BMW] [etc.] [cars]). In this sentence, 及(and) is an uni-gram between two bi-grams, 平 治 (Benz) and 寶 馬 (BMW). The segmentation point between 及 and 寶 is detected but the segmentation between 治 and 及 is not detected.

For the sentence in Figure 6, 就 投 訴 現 任 高 級 助 理 警 務 處 長 李 明 逵 於 去 年 回 歸 當 晚 (For complaining current senior associate chairman of police department, Lee Ming-kwai at last year returning evening), it is segmented to (就) (投 訴) (現 任) (高 級) (助 理) (警 務 處) (長) (李 明 逵) (於) (去 年) (回 歸) (當 晚) ([For] [complaining] [current] [senior] [associate] [chairman of police department] [Lee Ming-kwai] [at] [last year] [returning] [evening]). Names and other words are segmented without errors in this sentence.

For the sentence in Figure 7, 主 禮 嘉 賓 將 包 括 國 家 主 席 江 澤 民 及 副 總 理 錢 其 琛 (The guests of ceremony include the country chairman, Jiang Zemin, and vice president, Qian Qichen), it is segmented to (主 禮 嘉 賓) (將) (包 括) (國 家) (主 席) (江 澤 民) (及) (副 總 理) (錢 其 琛) ([The guest of ceremony] [will] [include] [country] [chair-

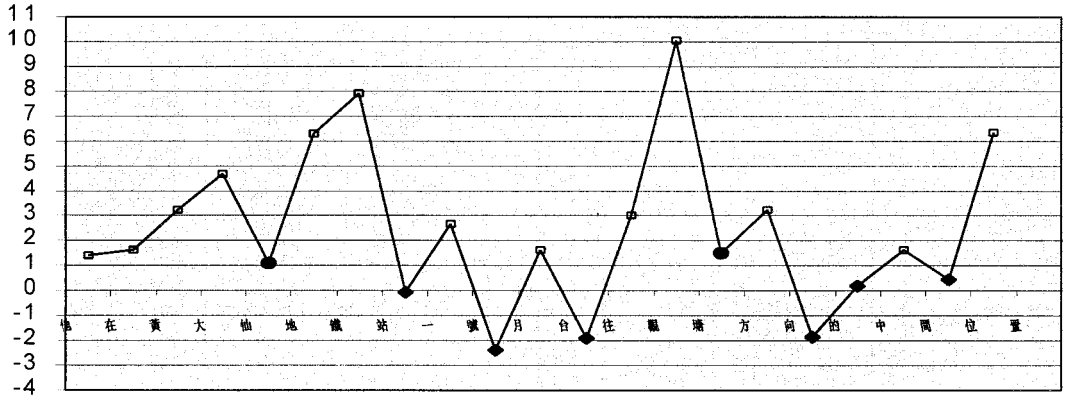

FIG. 4. Mutual information values of the sentence 他在黃大仙地鐵站一號月台往觀塘方向的中間位置.

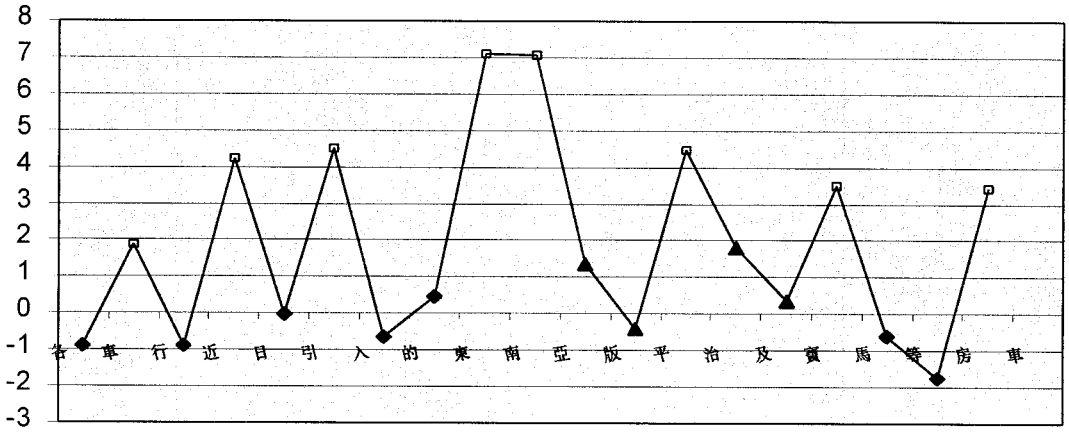

FIG. 5. Mutual information values of the sentence 各車行近日引入的東南亞版平治及寶馬等房車.

man] [Jiang Zemin] [and] [vice president] [Qian Qichen]). In this sentence, the names, 江 澤 民 (Jiang Zemin) and 錢 其 琛 (Qian Qichen), and their positions, 國 家 主 席 (country chairman) and 副 總 理 (vice president) are segmented without difficulty whiles errors occur using Algorithm 1 of the combination approach.

Experiments are also conducted to test if these approaches are corpus dependent. We have tested Chinese sentences that originated from news articles not within our corpus. These articles were published either earlier or later than our corpus. The results show that the accuracy drops but not significantly. Most of the words are segmented without problems. However, new unknown words that do not appear in our corpus can no longer be identified. The conclusion is that both of the combination and boundary detection approaches are corpus dependent.

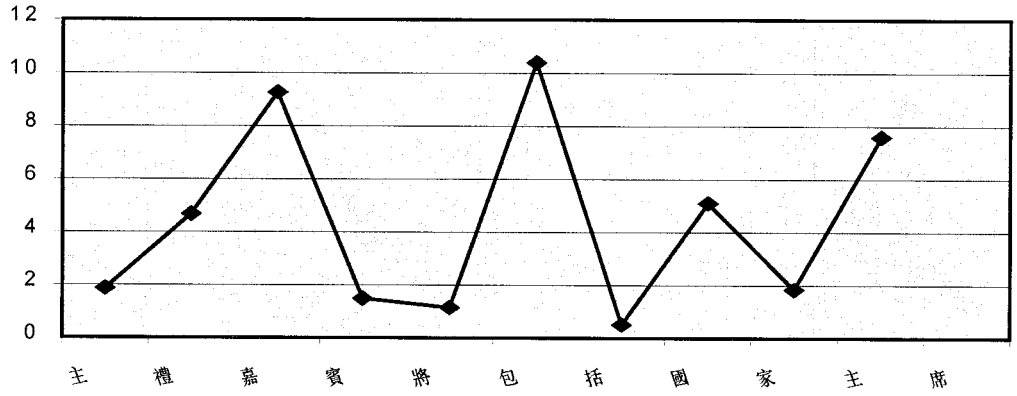

FIG. 6. Mutual information values of the sentence 就投訴現任高級助理警務處長李明逵於去年回歸當晚.

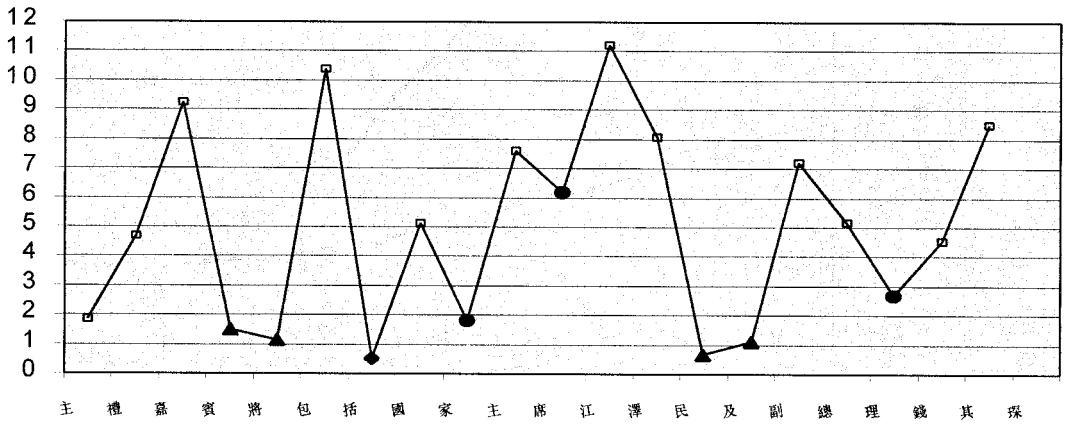

FIG. 7. Mutual information values of the sentence 主禮嘉賓將包括國家主席江澤民及副總理錢其琛.

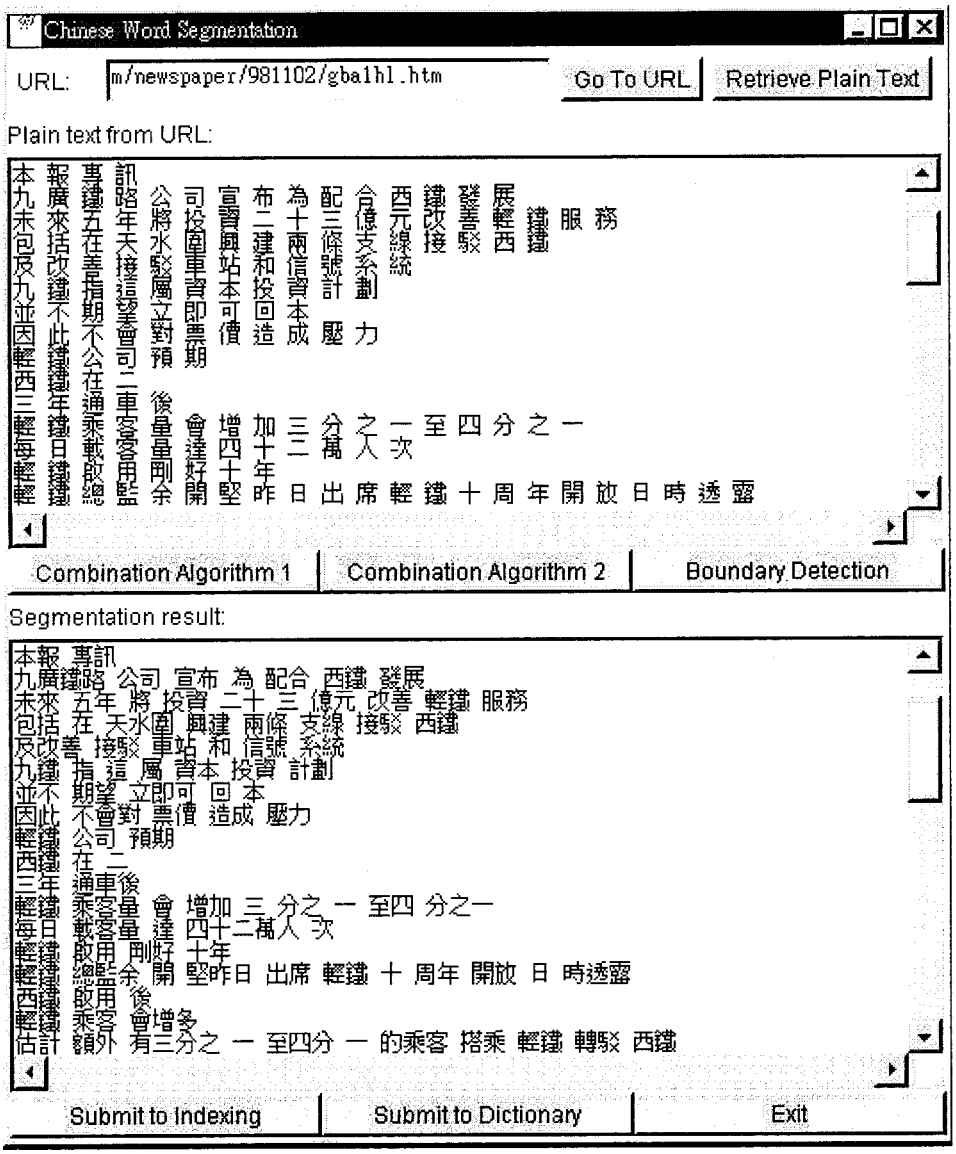

FIG. 8. System interface of the combination and boundary detection approaches for Chinese segmentation.

6. System Interface

The combination and boundary approaches for Chinese segmentation are implemented as a subsystem of a Chinese information retrieval system. The system interface is shown in Figure 8.

In this subsystem, user can input a web page address, in which a Chinese article is presented. Two buttons are available for the user to open the web page on a user-defined browser and retrieve the plain text to a text field in the interface. In Figure 2, an example of an URL address is submitted, and the corresponding web page is opened on a Netscape browser as shown in Figure 9 by clicking on the button "Go to URL." User can also click on the button "Retrieve Plain Text" to retrieve the plain text to the below text field labeled, "Plain text from URL." In the retrieved plain text, the punctuation is replaced by a line break.

When the plain text is retrieved, user can use any one of the three segmentation techniques, Combination Algorithm 1, Combination Algorithm 2, or Boundary Detection, to segment the text into *n*-grams and displayed on the lowest text field labeled, "Segmentation result:." In Figure 2, it shows the segmented *n*-grams of the corresponding web page.

The segmentation result can then be submitted to other processing. In Figure 2, it shows that user can click on the button "Submit to indexing" to index the processing homepage or click on the button "Submit to Dictionary" to identify the unknown words that do not exit in the current dictionary and add them to the dictionary. The "Exit" button is for the user to get out of the system.

7. Conclusion

Chinese indexing is an important task for searching in a Chinese digital library. However, Chinese text does not have a delimiter as in English text. Statistical approach for

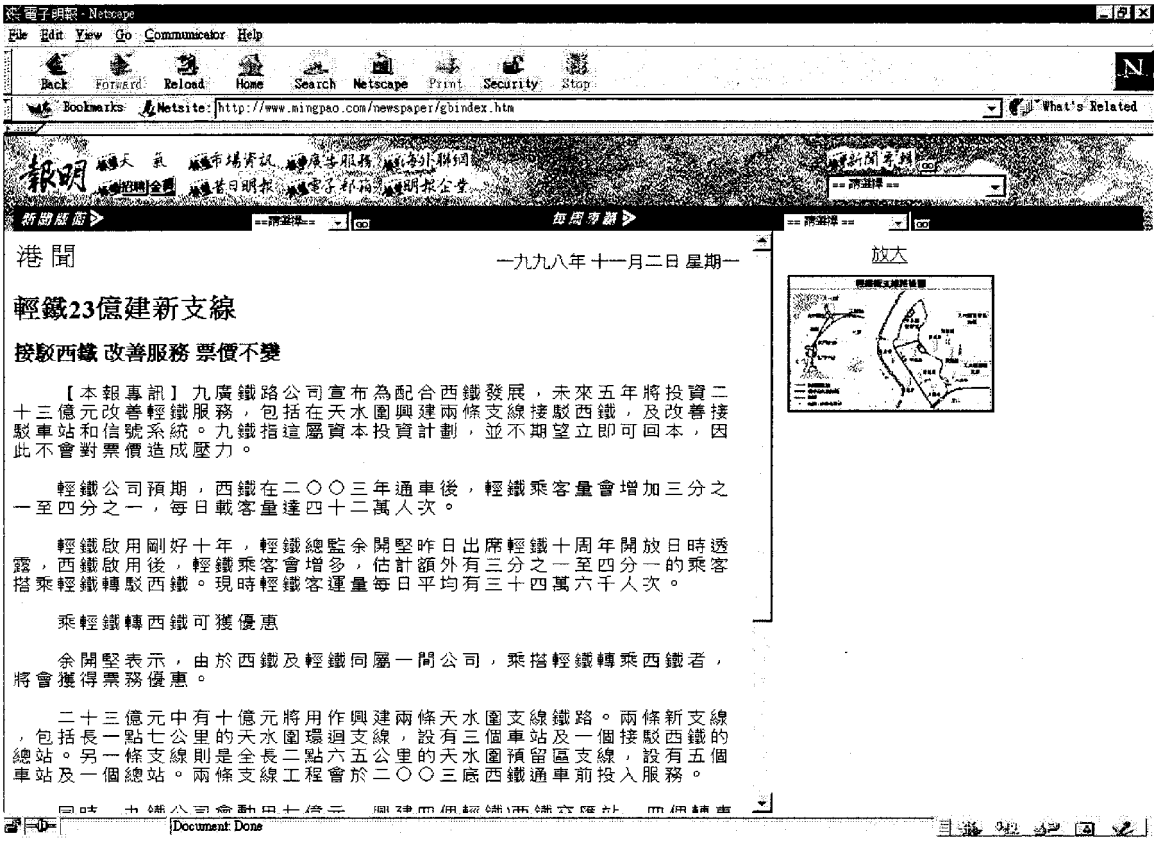

FIG. 9. A browser showing the webpage used in the example of system interface in Figure 2.

Chinese segmentation is known to have good performance in identifying unknown words, such as names, places, events, and specialized terms, which do not exist in a dictionary but appear quite often in newspapers.

In this article, we propose two approaches based on mutual information. Algorithm 1 and Algorithm 2 in the combination approach combine *n*-grams with fewer numbers of characters to form longer *n*-grams. Algorithm 2 allows combinations between subsets of an extracted *n*-gram while Algorithm 1 does not. The boundary detection approach segments a sentence based on the threshold and change of mutual information values. We also conduct an experiment to test their performance. Algorithm 2 of the combination approach produces noise segments as by-products and it is not avoidable. The result also shows that the boundary detection does not have the incorrect segmentation as Algorithm 1 and Algorithm 2 of the combination approach do. On the other hand, the boundary detection approach still suffers in poor segmentation in some situations, such as an uni-gram is located between two *n*-grams ($n \geq 2$) or names are formed by several words that always appear separately. However, these errors can be minimized when the size of the corpus increases. Overall, the boundary detection approach is more accurate and efficient than the combination approach.

References

Chen, A., He, J., Xu, L., Gey, F.C., & Meggs, J. (1997). Chinese text retrieval without using a dictionary. Proceedings of ACM SIGIR (pp. 42–49), Philadelphia, PA.

Chen, H., Ng, D.T., Martinez, W., & Schatz, B.R. (1997). A concept space approach to addressing the vocabulary problem in scientific information retrieval: An experiment on the worm community system. Journal of the American Society for Information Science, 48, 17–31.

Chen, K.-J., & Liu, S.-H. (1992). Word identification for Mandarin Chinese sentences. Proceedings of COLING, Nantes, France (pp. 101–107).

Chien, L.F. (1995). Csmart—A high-performance Chinese document retrieval system. Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages, ICCPOL '95, Hawaii (pp. 176–183).

Chien, L. F. (1997). PAT-tree-based keyword extraction for Chinese information retrieval. Proceedings of ACM SIGIR (pp. 50–58), Philadelphia, PA.

Davis, L.S. (1975). A survey of edge detection techniques. Computer Graphics and Image Processing, 4, 248–270.

Fan, C.K., & Tsai, W.H. (1988). Automatic word identification in Chinese sentences by the relaxation technique. Computer Processing of Chinese & Oriental Languages, 4, 33–56.

Gan, K.W., Palmer, M., & Lua, K.T. (1997). A statistically emergent approach for language processing: Application to modeling context effects in ambiguous Chinese word boundary perception. Computational Linguistics, 22(4), 4531–553.

Ikehara, S., Shirai, T.S., & Kawaoka, T. (1995). Automatic extraction of uninterrupted and interrupted collocations from very large Japanese

corpora using *n*-gram statistic. Transactions of the Information Processing Society of Japan, 36, 2584–2596.

Kwok, K.L. (1997). Comparison representations in Chinese information retrieval. Proceedings of ACM SIGIR (pp. 34–41), Philadelphia, PA.

Leun, C.H., & Kan, W.K. (1996). Parallel Chinese word segmentation algorithm based on maximum matching. Neural, Parallel and Scientific Computations, 4, 291–303.

Leung, C.H., & Kan, W.K. (1996). A statistical learning approach to improving the accuracy of Chinese word segmentation. Literary and Linguistic Computing, 87–92.

Lua, K.T. (1990). From character to word—An application of information theory. Computer Processing of Chinese & Oriental Languages, 4(4), 304–313.

Nie, J.Y., Brisebois, M., & Ren, X. (1996). On Chinese text retrieval. In: H.P. Frei, D. Harman, P. Schauble, & R. Wilkinson (Eds.), Proceedings of 19th Annual International ACM SIGIR Conference on R&D in IR (pp. 225–233), Zurich, Switzerland: ACM.

Nie, J.Y., Hannan, M.L., & Jin, W. (1994). Combining dictionary, rules and statistical information in segmentation of Chinese. Computer Processing of Chinese and Oriental Languages, 9(2), 125–143.

Nie, J.Y., Jin, W., & Hannaan, M.L. (1994). A hybrid approach to unknown word detection and segmentation of Chinese. Proceedings of International Conference on Chinese Computing, 326–335.

Ogawa, Y., & Iwasaki, M. (1995). A new character-based indexing method using frequency data for Japanese documents. Proceedings of 18th ACM SIGIR Conference on R&D in IR, Seattle, WA (pp. 121–129).

Prewitt, J.M.S. (1970). Object enhancement and extraction. Picture Processing and Psychopictorics, New York, NY: Academic Press.

Roberts, L.G. (1977). Machine perception of three-dimensional solids. Computer Methods in Image Analysis, Los Angeles: IEEE Computer Society.

Schatz, B., & Chen, H. (1996). Building large-scale digital libraries. IEEE Computer, Special Issue on Digital Library Initiative, 29(5), 22–27.

Schatz, B., & Chen, H. (1999). Digital libraries: Technological advances and social impacts. IEEE Computer, Special Issue on Digital Libraries, 32, 45–50.

Schatz, B., Mischo, W., Cole, T., Bishop, A., Harum, S., Johnson, E., Neumann, L., Chen, H., & Ng, D. (1999). Federated search of scientific literature. IEEE Compute, Special Issue on Digital Libraries, 32, 51–59.

Schatz, B., Mischo, W., Cole, T., Hardin, J., Bishop, A., & Chen, H. (1996). Federating diverse collections of scientific literature. IEEE Computer, Special Issue on Digital Library Initiative, 29, 28–36.

Sproat, R., & Shih, C. (1990). A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese and Oriental Languages, 4, 336–351.

Sproat, R., Shih, C., Gale, W., & Chang, N. (1996). A stochastic finite-state word segmentation algorithm for Chinese. Computational Linguistic, 22, 377–404.

Sun, M., Shen, D., & Huang, C. (1997). CSeg&Tag1.0: A practical word segmenter & POS tagger for Chinese texts. Proceedings of 5th Conference on Applied Natural Language Processing, Washington, D.C. pp. 119–124.

Wang, L.J., Pei, T., Li, W.C., & Huang, L.R. (1991). A parsing method for identifying words in Mandarin Chinese sentences. Natural Language (pp. 1018–1023).

Wu, Z., & Tseng, G. (1995). ACTS: An automatic Chinese text segmentation system for full text retrieval. Journal of the American Society for Information Science, 46, 83–96.

Wu, Z., & Tseng, G. (1993). Chinese text segmentation for text retrieval: Achievements and problems. Journal of the American Society for Information Science, 44, 532–542.

Yeh, C.L., & Lee, H.J. (1991). Rule-based word identification for Mandarin Chinese sentences—A unification approach. Computer Processing of Chinese and Oriental Languages, 5, 97–118.