Pergamon

Information Processing & Management, Vol. 32, No. 2, pp. 227–237, 1996

0306-4573/96 $15 + 0.00

0306-4573(95)00039-9

OPTIMAL WEIGHT ASSIGNMENT FOR A CHINESE SIGNATURE FILE

TYNE LIANG,[1] SUH-YIN LEE[1] and WEI-PANG YANG[2]

[1]Institute of Computer Science and Information Engineering and [2] Institute of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan 300, R.O.C.

(Received December 1994; accepted June 1995)

Abstract—The performance of a character-based Chinese text retrieval scheme (the combined scheme) is investigated. In the scheme both the monogram keys (singleton characters) and bigram keys (consecutive character pairs) are encoded into document signatures such that half of the bits in every signature are set. For disyllabic queries, an analytical expression of the false hit rate that accounts for both random false hits and adjacency false hits is proposed. Then optimal monogram and bigram weight assignments together with the corresponding minimal false hit rate are derived in terms of signature length, storage overhead of the combined scheme, and the occurrence frequency and the association value of a disyllabic query. The theoretical predictions of the optimal weight assignments and the minimal false hit rate are tested and verified in experiments using a real Chinese corpus for disyllabic queries of different association values. Satisfactory agreement between the experimental results and theoretical predictions is found.

1. INTRODUCTION

Many text access mechanisms have been proposed in the literature (Faloutsos, 1985, 1990) such as inversion method, clustering method, and signature method. Among them, the signature-based method has the advantages of low space overhead and relatively simple insertion of data (Faloutsos, 1985). However, a document retrieved with the signature method may not actually contain the words in the query. Such a document is called a false hit. In general the number of false hits increases with the number of documents in the database and the system performance may be seriously affected, especially for large databases (Leng & Lee, 1992). Therefore a lot of effort has been spent on reducing the false hit rate of signature-based text retrieval systems (Faloutsos, 1987, 1990; Faloutsos & Christodoulakis, 1985; Leng & Lee, 1992).

In the signature method, each indexing word of a document is transformed into a fixed-length word signature (a binary bit pattern) by hashing functions and the signatures of all indexing words in a document are ORed to create a document signature. To retrieve relevant documents with respect to a query, a query signature is generated as a filter. Since the bits in a word signature are set to one by hashing functions, a document signature will be filtered out if any of its bits corresponding to the set bits of the query signature is not set. On the other hand, a document with matched signatures may not actually contain the search word. This is because the corresponding bits in document signature can be set by irrelevant words in the document other than the search word during signature creation. Thus, documents accessed by matched signatures may be returned to users as false hits.

Experimental and theoretical analysis shows that the number of false hits are directly controlled by the weight (i.e. the number of bits set to one) in a document signature and that in a query signature. A query signature with more weights will result in less false hits since more bits in the document signature have to be matched in order to pass through the filtering process. On the other hand, a document signature with more weights will have a higher chance to be retrieved as a false hit. Since both the weight of document signature and that of query signature increase with the weight assigned to a word, there is an optimal condition for the smallest

number of false hits. It has been shown that at the optimal condition, the weight assigned to a word is such that the weight of the document signature equals half the signature length. This corresponds to maximum entropy in the document signature from an information theory point of view (Faloutsos, 1987).

Roberts (1979) and Faloutsos (1987) implement an optimal weight assignment scheme for minimizing false hit rate such that the signature of each word is assigned the same weight. For documents with a fixed number of distinct words, the optimal weight assigned to each word can be calculated in terms of the signature length (Stiassny, 1960). Hence, to achieve the optimal condition a large-size document is traditionally divided into document blocks containing a fixed number of unique words such that its signature is expected to contain half number of set bits. Then these block signatures are concatenated to form the document signature. In this way, document signatures with too high or too low weight should be eliminated. However, due to the problem of collision (i.e. different words may set the same bit in the signature) in non-perfect hashing, it is only the expected signature weight that can be controlled by adjusting the number of distinct words within the block. Variation of the document signature weight is unavoidable and the optimal condition is only satisfied on statistical grounds. In view of such a problem, Leng and Lee (1992) proposed a fixed-weight method in which the weight of each block signature is fixed instead of fixing the number of distinct words (the fixed-size block method) in the block. Their simulation shows that the fixed-size block method is indeed superior to the fixed-size block method in reducing false hits under certain conditions.

In addition to fixing the document weight for optimal condition, the false hits can be reduced further by assigning different weights to different classes of words. Faloutsos and Christodoulakis (1985) proposed an optimal weight assignment which takes into account both the non uniform query and occurrence frequencies of words. If words with high discriminatory power (i.e. high query frequency and low occurrence frequency) receive more weights than those words with low discriminatory power (i.e. low query frequency but high occurrence frequency), substantial savings in false hits can be achieved in fixed size block method. Leng and Lee (1992) then extended the above optimal weight assignment to include the signature generation by fixed weight block method.

The schemes previously described are basically designed for single word queries. Although they are effective for English text retrieval, they are not suitable for Chinese text retrieval. This is because words in Chinese text are not bounded by spaces as in English text. Automatic word identification for large collection of Chinese textual documents becomes time-consuming and it is difficult to achieve 100% accuracy of word identification (Wu & Tseng, 1995). Some authors claim that it is impractical to do word-based text retrieval for large-size textual documents (Tseng *et al.*, 1989; Liang *et al.*, 1994a, b; Chien, 1994) since many special terms or names occurring in texts and query may not be included in a dictionary which is essential in word-identification. One can resort to character-based algorithms with the multi-syllabic (multi-character) words treated as combination of their constituent characters. Using character as the basic indexing unit will eliminate the need of a word or term dictionary. Besides, the total number of unique Chinese characters is much smaller than that of Chinese words. So, it is easier to manage an exhaustive list of Chinese characters. Furthermore, character processing is more efficient than word processing because a Chinese character is represented by two bytes (in Big-5 code) while the number of bytes to represent a Chinese word depends on the number of characters in the word.

However, when character becomes a basic indexing unit of text, exact ordering of characters in a document becomes indispensable information to retrieve correctly a multi-syllabic word. Otherwise, a document which contains the components of the multisyllabic query will be retrieved even though these component characters are not in the exact sequence as in the query. Such a retrieved document is called an adjacency false hit.

To reduce adjacency false hits, additional space overhead for storing character sequence information in a character-based index file is required. Nevertheless, the signature method can easily support the adjacency operation by encoding positional information of characters without requiring too much additional space overhead. In our previous work (Liang *et al.*, 1994a) the combined scheme is implemented in such a way that both monogram keys (singleton characters)

and biogram keys (consecutive character pairs) within a document are encoded in a document signature. The experimental results showed that the combined scheme outperforms the traditional monogram signature scheme in terms of lower false hit rate.

Another character-based signature scheme for Chinese text retrieval has been proposed by Chien (1994). He divides a document signature into two parts. The first part encodes commonly-used characters in such a way that characters with the lowest similarity value will set the same single bit in the signature through a lookup table generated by a character-grouping algorithm specially designed to reduce false hits. The second part of the signature accommodates rarely-used characters through random hashing. His experiments on database of book titles shows a substantial improvement in certain circumstances if the special lookup table is used instead of random hashing. However no analytical formula is given for estimating the number of false hits. Hence, his findings remain empirical.

In this paper, the combined scheme proposed by Liang *et al.* (1994a) is analyzed in detail. A theoretical expression of the false hit rate in terms of the weights assigned to monogram keys and bigram keys is derived. An optimal weight assignment is proposed to minimize the number of false hits for disyllabic queries which contribute the largest part of Chinese word distribution (Lua, 1990). In addition, the proposed assignment accounts for both the word occurrence frequency and the association value (Sproat & Shih, 1990) which indicates the correlation of the word and its character components. In the end an experiment to test and verify the model is described. It is found that the experimental results tested on a real corpus agree well with the theoretical predictions.

2. CHARACTER-BASED CHINESE TEXT ACCESS BY SIGNATURE FILE

Due to the complexity of automatic word identification in Chinese texts, Chinese text retrieval is generally based on the character-indexing rather than word-indexing. The character-based indexing is fast. However, it introduces two possible kinds of false hits (Liang *et al.*, 1994d). One is due to the lack of correct parsing. For example, a document with '開發中國家' (developing country) will be treated as a qualifying document with respect to the query '中國' (China). Such a false hit is intrinsic and cannot be eliminated in character-based indexing. Another kind of false hits, known as adjacency false hits, is due to the loss of character sequence. For example, a document with '法治國家' (constitutional country) is an adjacency false hit with respect to the query '法國' (France). One can see that such false hits are more significant in Chinese text retrieval than in English. This is because the occurrence frequencies of commonly-used Chinese characters are extremely high (Liang *et al.*, 1994c). Furthermore, character combinations in a Chinese term can be constructed through reduplication, affixation and compounding processes, and character sequence is significant in the interpretation of a term. For example, '中國' (China), '日本' (Japan), '科學' (science), have very different semantics from the following terms which use the same characters: '國中' (middle school), '本日' (today), '學科' (study field), respectively. In addition, Chinese names for persons or places are constructed in a rather random manner and they are frequently used as query terms. Nevertheless, the adjacency false hits can be reduced if the character sequence is encoded in the character-based index file.

2.1. The combined scheme

It is noted that the space overhead for storing character sequences increases with the size of a textual database. Among various indexing schemes, the signature method can easily support adjacency operations without much space overhead (Sacks-Davis *et al.*, 1987; Liang *et al.*, 1994a). Hence, Liang *et al.* (1994a) proposed a combined scheme based on superimposed coding which supports fast searching and hardware implementation on parallel machines (Stanfill & Kahle, 1986).

In the combined scheme based on superimposed coding, every singleton character

(monogram) and every adjacency key (bigram) are encoded into a document signature by hashing functions. Below is an example to illustrate the procedure of the creation of a document signature. Let us consider a simple Chinese text document block '法治國家' (constitutional country). Since there are four Chinese characters, the document block size is $D=4$. Let the length of the signature be $b=32$ bits. Then a hashing function should return an integer between 1 and 32 for each Chinese character. A simple hashing function h such that $h=((A_1*B_1+A_2*B_2) \bmod 32)+1$ will be sufficient for most cases. Here A_1 and A_2 are prime numbers: B_1 and B_2 are respectively, the first and second byte of a Chinese character. Using different values of A_1 and A_2, one can construct many hashing functions. Suppose m_1 ($=3$) of such hashing functions are used to set the bits in each monogram signature. Then the monogram signatures could be:

法 0000 0100 0000 0000 1000 0000 0000 1000,
治 0100 0000 0100 0000 0000 0000 1000 0000,
國 0000 0000 0000 0010 0001 0000 0100 0000,
家 0000 0000 0010 0000 0000 0011 0000 0000.

With the traditional singleton-character (monogram) scheme the document signature for '法治國家' will be

DS1 0100 0100 0110 0010 1001 0011 1100 1000

by ORing the monogram signatures. However, with the combined scheme, all the bigrams of a document will be also treated as keys and encoded in the document signature. Since a bigram contains two characters, each bigram is represented by 4 bytes $B_1B_2B_3B_4$ with the B_1B_2 and B_3B_4 representing the first and the second characters, respectively. The hashing functions g for bigrams can be similar to those for monograms such that $g=((A_1*B_1+B_2+A_2*B_3+B_4) \bmod 32)+1$. Suppose we use m_2 ($=2$) hashing functions to set the bits in each bigram signature and the bigram signatures are:

法治 0000 0000 0100 0000 0000 0000 0000 0001
治國 0010 0000 0000 0000 0000 0000 0000 0010
國家 0000 0000 0000 0000 0000 0000 0000 0101.

Then the document signature for '法治國家' (constitutional country) will be

DS_2 0110 0100 0110 0010 1001 0011 1100 1111

after storing all the monogram and bigram signatures. It is obvious that the combined scheme is capable of reducing the number of adjacency false hits whereas the traditional monogram signature scheme cannot no matter how many hashing functions are used during signature creation. For example, the query signature with monogram signature scheme for '法國' (France) is

QS_1 0000 0100 0000 0010 1001 0000 0100 1000.

Then the document block '法治國家' (constitutional country) will be retrieved as an adjacency false hit since the document block '法治國家' (constitutional country) is treated as a collection of four characters ('法', '治', '國', '家') and the character sequence existed in the document is lost by the traditional monogram signature scheme.

However, in the combined scheme the additional bigram signature for '法國', say,

0001 0000 0000 0000 0001 0000 0000 0000.

will be encoded and the query signature for '法國' (France) becomes

QS_2 0001 0100 0000 0010 1001 0000 0100 1000.

Therefore by checking the bits set by bigram hashings (i.e. the 4th and 20th bits) in QS_2, the document signature DS_2 for '法治國家' ('constitutional country') will be rejected.

2.2. *False hit rate*

Based on the signature file an unqualified document may be filtered out immediately by comparing its signature to that of the query. However, among those documents whose signatures

pass through the filtering process, some of them may not actually contain the query. This is because those set bits in the document signature with respect to the set bits in the query signature may be set by some other irrelevant keys after superimposed coding. For example, if the query '台灣' (Taiwan) has a signature

0100 0100 0100 0000 1000 0000 1000 1001

the document '法治國家' (constitutional country) will be retrieved even though the document does not contain the query. In this situation, we call the document '法治國家' (constitutional country) a random false hit with respect to the query '台灣' (Taiwan).

The random false hit rate RFHR, which is defined as the probability that an unqualified document emerges as a random false hit, can be estimated (Shepherd *et al.*, 1989) to be

$$\mathrm{RFHR} = p^m, \tag{1}$$

where p is bit-set density of a document signature and m, is the weight in a query signature. The above expression for RFHR is in fact the probability that all the m bits in a document signature corresponding to those set bits in a certain query signature are accidentally set. For a query which contains only one character c_1, $m = m_1$. For a disyllabic query "c_1c_2" which contains two characters c_1 and c_2, $m = 2m_1$ in traditional monogram signature scheme. As mentioned before, the adjacency false hits will not be filtered. Therefore the total false hit rate TFHR_t of the monogram scheme should include the adjacency false hit probability AFHR. In the lowest order approximation (Liang *et al.*, 1994c)

$$\mathrm{AFHR} = D(D-1)\rho_1\rho_2 - (D-1)\rho_1\rho_2 = (D-1)^2\rho_1\rho_2 \tag{2}$$

where D is the number of characters in a document block and ρ_1, ρ_2 are respectively the occurrence probabilities of the characters c_1, c_2. The term $D(D-1)\rho_1\rho_2$ expresses the probability that a D-character document block contains at least one C_1 and one c_2, whereas the term $(D-1)\rho_1\rho_2$ indicates the probability that the D-character document block contains the character sequence c_1c_2. Then adding equations (1) and (2), we obtain the total false hit rate of a disyllabic query in the traditional scheme:

$$\begin{aligned}\mathrm{TFHR}_t &= \mathrm{RFHR} + \mathrm{AFHR} \\ &= p^{2m_1} + (D-1)^2\rho_1\rho_2.\end{aligned} \tag{3}$$

In the combined scheme, there are additional m_2 set bits that have to be matched in the filtering process before coming out as false hit. Therefore the total false hit rate TFHR for the combined scheme is

$$\begin{aligned}\mathrm{TFHR} &= \mathrm{TFHR}_t\, p^{m_2} \\ &= \{p^{2m_1} + (D-1)^2\rho_1\rho_2\}p^{m_2} \\ &= p^{2m_1+m_2} + (D-1)^2\rho_1\rho_2 p^{m_2}.\end{aligned} \tag{4}$$

The TFHR as given in equation (4) is expressed in terms of the occurrence probabilities of the characters in the text database. As mentioned before, a stand-alone character in a Chinese textual document usually is not treated as a meaning-complete entity. It is the word that is actually significant. Therefore, the total false hit rate should be expressed in terms of the occurrence probability ρ of the disyllabic word c_1c_2 instead of the occurrence probabilities of its components. Now ρ can be expressed in terms of ρ_1 and ρ_2 through their correlation relation

$$S = \log\frac{\rho}{\rho_1\rho_2} \tag{5}$$

where S is the association value (Sproat & Shih, 1990) of the disyllabic term c_1c_2 and log denotes logarithm of base 2. From equation (5) we have

$$\rho_1\rho_2 = \rho 2^{-S}. \tag{6}$$

Substitute the above equation (4), the total false hit rate of the combined scheme is

$$\mathrm{TFHR} = p^{2m_1+m_2} + (D-1)^2\rho 2^{-S}p^{m_2}. \tag{7}$$

In this expression, one can see that the total false hit rate for the combined scheme is very similar to that of the traditional monogram signature scheme because it can be considered as the sum of the random false hit rate and the adjacency false hit rate. The first term in equation (7) associates with those false hits which do not contain the query at all. So all the $2m_1 + m_2$ set bits in the query have to be matched. As for the adjacency false hit, only the m_2 set bits have to be matched because the presence of the query components guarantees that the $2m_1$ bits are set. Note that similar to the random false hit rate of the traditional monogram signature scheme, the "random" false hit rate of the combined scheme depends explicitly on the weight of the query signature and the bit-set density p only. In contrast, the adjacency false hit rate depends not only on p and m_2 but also explicitly on the occurrence probability and the association of the query, as well as the number of characters D in a document block. In the combined scheme with fixed p, D is not fixed and it depends implicitly on the parameters b (signature length), m_1 and m_2.

3. OPTIMAL FALSE HIT RATE AND OPTIMAL WEIGHT ASSIGNMENT

As discussed in the previous section, the total false hit rate TFHR of the combined scheme depends on the parameters m_1, m_2, and b. Therefore an optimal weight assignment to minimize total false hit rate may exist when the constraints on m_1, m_2, b are given. The most important constraint is on the storage overhead B, which is the size of the signature file. Denoting the number of block signatures in the signature file by N_s and the total number of characters in the corpus by N_c, we have

$$B = bN_s = \frac{bN_c}{D}. \tag{8}$$

For a document block containing D characters, the number of bigrams is $D-1 \approx D$. Then the weight of the block signature is about $Dm_1 + Dm_2 = D(m_1 + m_2)$ if no collision occurs. Since the bit-set density is fixed at 1/2 during signature extraction, then $D(m_1 + m_2) = b/2$ or

$$D = \frac{b}{2(m_1 + m_2)}. \tag{9}$$

When the effects of collision and the presence of repeated characters and punctuation marks are taken into account, one would expect D to be larger than the above expression. So D can be written as

$$D = \frac{\beta b}{2(m_1 + m_2)} \tag{10}$$

where β is a constant greater than unity. Then from equation (8), the storage overhead is

$$B = 2(m_1 + m_2)N_c/\beta. \tag{11}$$

This shows that fixing the storage overhead is equivalent to fixing the sum of the monogram and bigram weights, i.e.

$$m_1 + m_2 = C \tag{12}$$

with $C = \beta B/(2N_c)$ being a constant. Under this constraint the average number of characters in a document block is, from equations (10) and (12),

$$D = \frac{\beta b}{2C}. \tag{13}$$

Given the constraint $m_1 + m_2 = C$ being a constant, the contribution of random false hit rate to the total false hit rate can be rewritten as p^{2C-m_2} [see equation (7)] which is an increasing

function of m_2 because the bit-set density of each block signature $p = 0.5 < 1$. This is a direct consequence of the fact that there are two monograms but only one bigram in a disyllabic term. In contrast, the contribution due to the adjacency false hit is proportional to p^{m_2} which is a decreasing function of m_2. Therefore it is not difficult to see that there may be an optimal combination of m_1 and m_2 such that the total false hit rate TFHR acquires a minimum value. To find this optimal condition we proceed by writing

$$\text{TFHR} = p^{C+m_1} + \alpha p C - m_1 \tag{14}$$

in which

$$\begin{aligned}\alpha &= (D-1)^2 \rho 2^{-S} \\ &\simeq \left(\frac{\beta b}{2C}\right)^2 \rho 2^{-S}\end{aligned} \tag{15}$$

under the approximation $D-1 \simeq D$ when D is large. Then the optimal monogram weight assignment m_1^* can be obtained by setting the derivative of TFHR to zero, i.e. $d/m_1^*\,\text{TFHR} = 0$. Therefore we have:

$$p^{C+m_1^*} \ln p - \alpha p^{C-m_1^*} \ln p = 0. \tag{16}$$

After rearrangement, equation (16) becomes

$$p^{2m_1^*} = \alpha. \tag{17}$$

Taking logarithm (base 2) of the above equation, we have

$$\begin{aligned}2m_1^* \log p &= \log \alpha \\ &= 2 \log \frac{\beta b}{2C} - \log \rho - S.\end{aligned} \tag{18}$$

Since $p = 1/2$, $\log p = -1$. After simplifying the above equation, the optimal monogram weight m_1^* is

$$m_1^* = \frac{S}{2} - \log \frac{\beta b}{2C} - \frac{1}{2} \log \rho \tag{19}$$

and the corresponding optimal bigram weight is

$$\begin{aligned}m_2^* &= C - m_1^* \\ &= C - \frac{S}{2} + \log \frac{\beta b}{2C} + \frac{1}{2} \log \rho.\end{aligned} \tag{20}$$

The above result shows that the optimal weight m_2^* assigned to the bigram should increase with the signature length b. In other words for large block signature length, one should assign more weight to the bigram instead of the monogram. This is because larger b implies that there will be more characters encoded in the block signature [see equation (13)], thus increasing the number of adjacency false hits according to equation (4). On the other hand, the optimal weight m_2^* decreases with the association of the disyllabic query. This is reasonable because a bigram of high association should have small number of adjacency false hits. Then the contribution of the m_2 bigram weight in the query signature to reduce total false hits is less significant and the shift to smaller m_2 at the optimal condition is expected.

From equations (14) and (17), the optimal total false hit rate is,

$$\begin{aligned}\text{TFHR}^* &= p^{C+m_1^*} + p^{2m_1^*} p^{C-m_1^*} \\ &= 2p^{C+m_1^*}\end{aligned} \tag{21}$$

Substitute m_1^* from equation (19) into the above equation, we have

$$\begin{aligned}\text{TFHR}^* &\propto 2p^{S/2 - \log b} \\ &\propto b^{-S/2}.\end{aligned} \tag{22}$$

since $p = 0.5 = 2^{-1}$. Therefore the optimal total false hit rate is proportional to b and decreases exponentially with S. However, it should be pointed out that the optimal conditions as given by equations (19), (20) and (21) may not be realized in practice because m_1 and m_2 have to be non-negative integers.

4. EXPERIMENTS AND ANALYSES

To test the theoretical results for the combined scheme, a corpus with articles from Freedom Times ('自由時報'), Taipei, Taiwan were examined. There were 3601 articles, 1,426,199 Chinese characters in which 4340 characters were unique. Since character '的' (a particle used for signaling a possessive/genitive relation in a noun phrase or nominalizing a verb phrase or a sentence) occurs in Chinese texts very frequently (36,908 occurrences in our corpus) and it seldom appears in a disyllabic term, it was treated as a word bounding symbol (i.e. similar to punctuation marks) instead of a key character in the text. This corpus was partitioned into N textual blocks using the combined scheme such that the weight in each block signature is $b/2$ where b is the signature size. Therefore $p = 0.5$ for each of the block signatures. In the experiment, for each $b = 80, 160, 240, \ldots, 800$, $C = m_1 + m_2$ was varied from 2 to 6 with all possible combinations of m_1 and m_2. With these parameters 250 text files and their associated signature files were generated.

The test queries were obtained by extracting all valid disyllabic terms from the corpus. The statistics ρ_1, ρ_2, and ρ of these terms were recorded and their association values S were calculated using its definition [see equation (5)]. Then the disyllabic terms were sorted into groups according to S and 100 terms were picked from those groups of $1 < S < 2$, $3 < S < 4$, $5 < S < 6$, $7 < S < 8$, $9 < S < 10$, $11 < S < 12$ as the test queries. With each query, the textual file was searched through the signature file for each combination of b, m_1 and m_2. Each of the retrieved document blocks was scanned to find the number of actual hits (A) and the number of false hits (R). Then the false hit rate was calculated by the ratio $R/(N-A)$, where N is the total number of document blocks in the corpus. To illustrate the existence of the optimal combination of m_1 and m_2, the mean value of the false hit rates for all the queries in each query group was calculated. Table 1 tabulates the mean false hit rates of different query groups at different combinations of m_1 and m_2 when $b = 800$, and $C = 6$. One can observe that the minimal mean false hit rate, which is printed in bold face, indeed takes place at non zero value of m_2.

To compare quantitatively the theoretical prediction, the minimal false hit rate (FHR*) and the corresponding optimal bigram weight m_2^* were looked for among the false hit rates of different combinations of m_1 and m_2 under the constraint $m_1 + m_2 = C$ for each query separately. Then the mean FHR* (MFHR*) and the mean m_2^* were obtained by taking the average values of FHR* and m_2^* for all the 100 queries within the group. It was found that the numerical values of MFHR* agree to within 20% of the theoretical prediction given by equation (21). When MFHR* is plotted against the signature length b for each query group (see Fig. 1), MFHR* increases linearly with b and decreases with increasing S. Such behaviour is again consistent with relation (22).

Table 1. Mean false hit rate of different association (S) query groups for different monogram-bigram weight combinations at $C = m_1 + m_2 = 6$ and $b = 800$. The minimal mean false hit rate for each query group is printed in **bold face**

S	$m_1=6$ $m_2=0$	$m_1=5$ $m_2=1$	$m_1=4$ $m_2=2$	$m_1=3$ $m_2=3$	$m_1=2$ $m_2=4$	$m_1=1$ $m_2=5$	$m_1=0$ $m_2=6$
1.5000	0.1738	0.0856	0.0489	0.0290	0.0186	**0.0176**	0.0182
3.5000	0.0952	0.0516	0.0301	0.0202	0.0144	**0.0141**	0.0185
5.5000	0.0418	0.0237	0.0157	0.0106	**0.0098**	0.0114	0.0173
7.5000	0.0116	0.0079	0.0058	**0.0055**	0.0067	0.0106	0.0181
9.5000	0.0027	**0.0021**	0.0027	0.0031	0.0048	0.0085	0.0178
11.500	0.0017	**0.0014**	0.0018	0.0029	0.0046	0.0091	0.0179

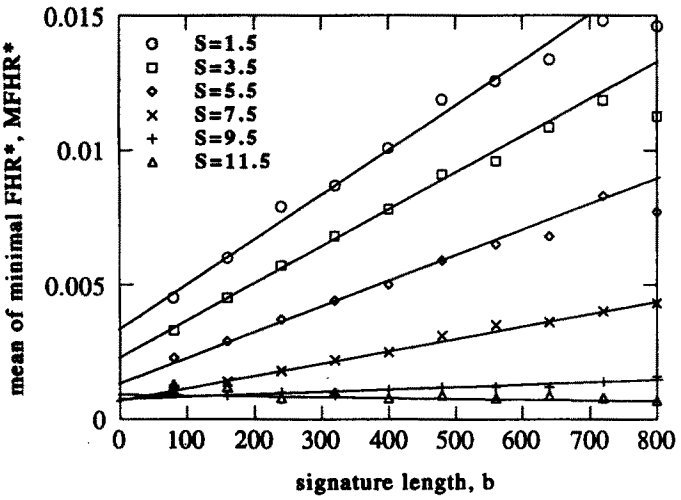

Fig. 1. Variation of the minimal false hit rate vs the signature length at $C=6$. Symbols are the experimental data and solid lines are there to show the linear relation of the minimal false hit rate with the signature length.

The theoretical optimal bigram weight m_2^* can be calculated by equation (20) if β is known. To obtain β, the total number of characters N_c in the corpus was divided by the number of signatures N_s to obtain the average number D of characters within each document block. From equation (13), β is the slope of the graph $2DC$ vs b. Figure 2 shows such a plot in which the quantity $2DC$ was found to vary quadratically with b. The solid curve in this graph is the best fit with equation $2DC = 9 + 1.49b + 0.0004b^2$. Nevertheless the quadratic contribution is very small as seen from the coefficient of b^2. Therefore β can be taken as the coefficient of the linear term in the fitted equation, i.e. $\beta = 1.49$. Using this value of β the theoretical m_2^* was calculated and plotted together with the experimental values at different signature length b for different query groups in Fig. 3 where $C=6$. One can find that the experimental results agree satisfactorily with theoretical prediction. In particular, when m_2^* is rounded off to an integer in practical design, the agreement between the theoretical and the experimental results is nearly perfect.

The agreement of the experimental MFHR* and m_2^* with their theoretical predictions are equally good in experiments in which $C=5$, 4, 3, 2. In other words, the experimental results

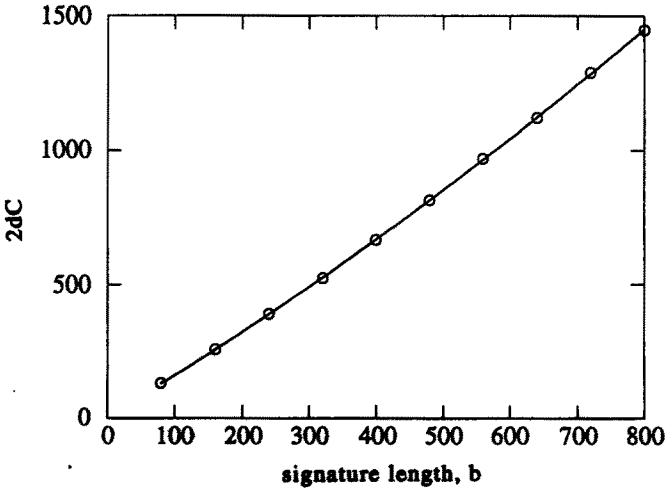

Fig. 2. Variation of $2DC$ vs b for $C=6$. The solid curve is the best fit of the equation $2DC = 9 + 1.49b + 0.0004b^2$.

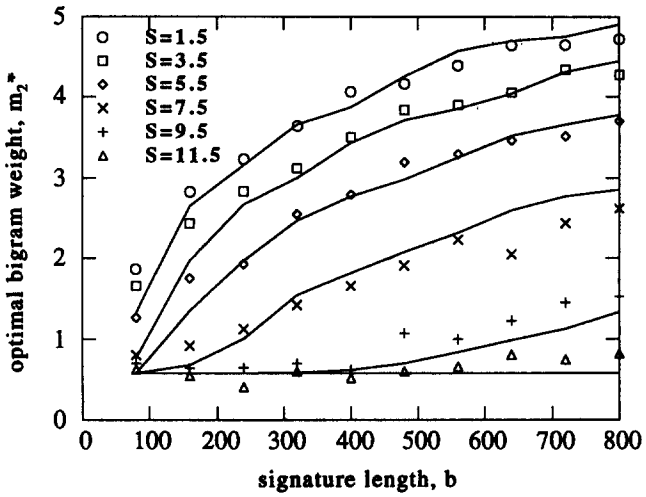

Fig. 3. Variation of optimal bigram weight vs signature length at $C=6$. Symbols are the experimental data and solid lines are the theoretical predictions.

show that the trade off between monogram weight and bigram weight can indeed improve the overall performance of character-based Chinese text retrieval through signature file. Furthermore, the improvement due to the addition of bigram weight is most significant for disyllabic queries of low association. This is due to their high adjacency false hit rate. It is interesting to find that for queries of high association in which the number of adjacency false hits is expected to be insignificant, the effect of bigram weight is still observable, i.e. $m_2^*>0$. It is only when both C and b are small that the bigram weight should be set to zero for high association queries.

5. CONCLUSION

As a conclusion, a theoretical basis has been formulated for a combined scheme in which bigram hashing is incorporated in addition to monogram hashing when generating the signature files for character-based Chinese text retrieval. The false hit rate of the combined scheme has been derived and the optimal weight assignment obtained. The theoretical predictions of the minimal false hit rate and optimal bigram weight have been tested and shown to be in close agreeement with experimental results using a real Chinese corpus for different system parameters for disyllabic queries of different associations.

Acknowledgement—This research was partly supported by the National Science Council of Taiwan, R.O.C. under contract NSC83-2213-E-009-127(1994).

REFERENCES

Chien, L. F. (1994). Fast and efficient full-text retrieval for large Chinese document database. In *Proceedings of the International Conference on Computer Processing and Oriental Languages*, Seoul, Korea.

Faloutsos, C. (1985). Text retrieval methods. *ACM Computing Survey, 17*(1), 49–74.

Faloutsos, C. (1987). Description and performance analysis of signature file methods for office filing. *ACM Transactions on Office Information Systems, 5*(3), 237–257.

Faloutsos, C. (1990). Signature-based text retrieval: a survey. *IEEE Data Engineering, 13*(1), 25–32.

Faloutsos, C., & Christodoulakis, S. (1985). Design of a signature file method that accounts for non-uniform occurrence and query frequencies. In *Proceedings of 9th International Conference on Very Large Data Bases* (pp. 165–170), Stockholm.

Leng, C. R., & Lee, D. L. (1992). Optimal weight assignment for signature generation. *ACM Transactions on Database Systems, 17*(2), 346–373.

Liang, T., Lee, S. Y., & Yang, W. P. (1994a). On the design of effective Chinese textual retrieval based on signature method. *Computer Processing of Chinese and Oriental Language, 8*(1), 87–96.

Liang, T., Lee, S. Y., & Yang, W. P. (1994b). A signature file for Chinese textual retrieval. In *Proceedings of International Conference on Chinese Computing* (pp. 412–417), Singapore.

Liang, T., Lee, S. Y., & Yang, W. P. (1994c). The analysis of disyllabic terms in Chinese text retrieval. *Computer Processing of Chinese and Oriental Language.*

Liang, T., Lee, S. Y., & Yang, W. P. (1994d). Approximating false hits of disyllabic terms in a Chinese signature file. *International Journal of Information Science and Engineering.*

Lua, K. T. (1990). From character to word—an application of information theory. *Computer Processing of Chinese and Oriental Language, 1*(4), 304–313.

Roberts, C. S. (1979). Partial-match retrieval via method of superimposed codes. *Proceedings of IEEE, 67*(12), 1624–1642.

Sacks-Davis, R., Kent, A., & Ramamohanarao, K. (1987). Multikey access methods based on superimposed coding techniques. *ACM Transactions on Database Systems, 12*(4), 655–696.

Shepherd, M. A., Phillips, W. J., & Chu, C. K. (1989). A fixed-size bloom filter for searching textual documents. *Computer Journal, 32*(3), 212–219.

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

Stanfill, C., & Kahle, B. (1986). Parallel free-text search on the connection machine system. *Communications of the ACM, 29*(12), 1229–1239.

Stiassny, S. (1960). Mathematical analysis of various superimposed coding methods. *American Documentation, 11*(2), 155–169.

Tseng, S. S., Yang, C. C., & Hsieh, C. C. (1989). On the design of Chinese textual database. *Computer Processing of Chinese and Oriental Language, 4*(2&3), 240–273.

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*(2), 83–96.