A Novel PAT-Tree Approach to Chinese Document Clustering

Kenny Kwok, Michael R. Lyu, Irwin King
Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
{clkwok,lyu,king}@cse.cuhk.edu.hk

Contact Person: Kenny Kwok, clkwok@cse.cuhk.edu.hk
Fax: (852) 2603 5024

A Novel PAT-Tree Approach to Chinese Document Clustering

Kenny Kwok, Michael R. Lyu, Irwin King
Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
{clkwok,lyu,king}@cse.cuhk.edu.hk

ABSTRACT

In this paper, we propose a novel Chinese document clustering technique based on the suffix tree clustering (STC) approach with the PAT-Tree data structure and our own ranking algorithms. By utilizing the PAT-tree data structure, we are able to segment Chinese documents easily. We then cluster this result using the STC method. We detail the PAT-tree data structure within the STC framework. In particular, we define the Essential Node, a type of node in the PAT-tree, which can be used directly as base clusters. Experimental results from a collection of Chinese newspaper articles with variations on document type demonstrate that our approach is feasible and effective.

Keywords

Clustering, Chinese Information Processing, PAT-tree

1. INTRODUCTION

Since the number of electronics Chinese documents is growing very fast, efficient techniques for Chinese information retrieval are needed. Much of the current research focuses on Chinese keyword extraction and sentence segmentation [1] [4] [5] [9]. PAT-tree [8] is a variation of Patricia Tree [2] developed by Gonnet in 1992. It is widely used in handling Chinese information in these areas [5] [9] since PAT-tree can handle word segmentation without a clear word boundary. Suffix tree clustering (STC) [6] is a clustering technique developed by Zamir and Etzioni in 1998. Although the STC is an efficient clustering technique that is suitable for clustering a large number of documents in real-time, it cannot handle documents without word boundary.

In this paper, we propose to combine PAT-tree data structure and the STC for Chinese document clustering. We use the PAT-tree data structure to overcome the problem of handling Chinese information in suffix tree. We use the STC framework to ensure the efficient clustering performance.

In the next section, we detail the Chinese PAT-tree data structure with the added definition of Essential Node. The STC algorithm is described in Section 3. We present our experimental set-up and results in Section 4. We then draw a conclusion in the final section.

2. CHINESE PAT-TREE STRUCTURE

2.1 PAT-Tree

PAT-tree [3] is a Patricia Tree [2]. It stores every semi-infinite string (sistring) [8] of a document into leaf nodes. Leaf nodes of a PAT-tree are called *external nodes*. They have a link to a sistring. Non-leaf nodes of a PAT-tree are called *internal node*s. Internal nodes are used as index to insert and locate sistrings efficiently from the PAT-tree in log (*n*) time.

Given a document, which contains a sequence of characters, a sistring is a subsequence of characters that goes from a given point on the sequence to the end of the sequence. Conceptually, sistrings are padded with special null characters at its end such that sistrings are of equal length for comparison from one to another.

Figure 1. An example of sistrings from a document

Document	NUMBER
sub-strings	N NU NUM NUMB NUMBE NUMBER U UM UMB UMBE UMBER M MB MBE MBER B BE BER E ER R
sistring 1	NUMBER
sistring 2	UMBER0
sistring 3	MBER00
sistring 4	BER000
sistring 5	ER0000
sistring 6	R00000

By the nature and definition of sistrings, when we have a set of all sistrings from a document, any sub-strings of that document can be located in a prefix of one of those sistrings. In other words, sistrings can efficiently represent all the possible sub-strings. An example of sistrings is given in Figure 1. It shows a simple document with 6 characters has 21 sub-strings that can be represented by 6 sistrings. Therefore, with this compressed tree data structure, searching of sub-strings inside a document can be done efficiently with the aid of PAT-tree.

2.2 Chinese PAT-Tree

PAT-trees are widely used for Chinese keyword extraction. The difference of Chinese PAT-tree is that sistrings are extracted at the sentence level instead of at the document level. This is because Chinese words

are packed together without delimiters. We cannot separate Chinese words with space as in English. However, Chinese sentences have punctuation marks as separators in the sentence level. Therefore, we can use the sistrings from the set of sentences of a Chinese document to construct the corresponding Chinese PAT-tree.

One of the problems resulting from the sistrings in Chinese document processing is that two or more sistrings obtained from the sentences may be identical because they may have equal length. Special attention is needed by having the frequency count in tree nodes to indicate the number of occurrence of identical sistrings.

To represent Chinese phrases for our clustering need, we carefully define a variation of the Chinese PAT-tree structure. Our proposed structure is very similar to the original PAT-tree [8] with some enhancements. First, the concept of external nodes and internal nodes in the PAT-tree will not play an important role in the clustering process. Both external nodes and internal nodes will be more uniform, and every node has a link to a sistring as well as a check bit value. Moreover, we define a new type of node called *Essential Node* (*EN*) such that information in *EN* can be used to identify the base clusters. This modification can simplify the clustering process with the Chinese PAT-tree. A summary of the differences among the PAT-tree, the Chinese PAT-tree and our modified Chinese PAT-tree is shown in Table 2.

Table 2. Summary of differences among three variations of PAT-tree

	PAT-tree	**Chinese PAT-tree**	**Our Modified Chinese PAT-tree**
Usage	For generic character sequence	For Chinese sentences on keyword extraction	For Chinese sentences on document clustering
Sistrings	in document level	in sentence level	in sentence level
External node	Have link to a sistring No check bit	Have link to a sistring No check bit	Have link to a sistring Have a check bit
Internal node	No link to sistrings Have a check bit	No link to sistrings Have a check bit	Have link to a sistring Have a check bit
Essential Node (*EN*)	-	-	Yes

2.3 Tree Node Structure

In our modified Chinese PAT-tree structure, each of the tree node consists of the following component: (1) a check bit, (2) a link to a sistring, (3) a *frequency* count of the phrase which is represented by the current node, and (4) left and right pointers to the child nodes. This is illustrated in Figure 3.

2.3.1 Check Bit

Check bit is the key information of the internal nodes in PAT-tree. It indicates the first different bit of the sistrings in the left and right subtrees. Branching decision in the internal nodes depends on this check bit.

Although branching is not required in external nodes, we still have check bit in external nodes for consistency. In this case, we define the check bit of external node equals to the bit length of the content of the sistring it is linked to.

2.3.2 Link to a sistring

Each external node can represent a sistring uniquely by this link. Hence, the same sistring can only have exactly one external node linking to it.

For internal nodes, we do have a link to a sistring. This link is pointed to any of the sistring its children pointed to. For simplicity, we can assume that it is always the same as the link of its left child. Therefore, all these links in internal nodes are redundant links.

When a node is identified as an *EN*, it can represent a Chinese phrase, which equals to a prefix of the sistring it is linked to.

2.3.3 Frequency count

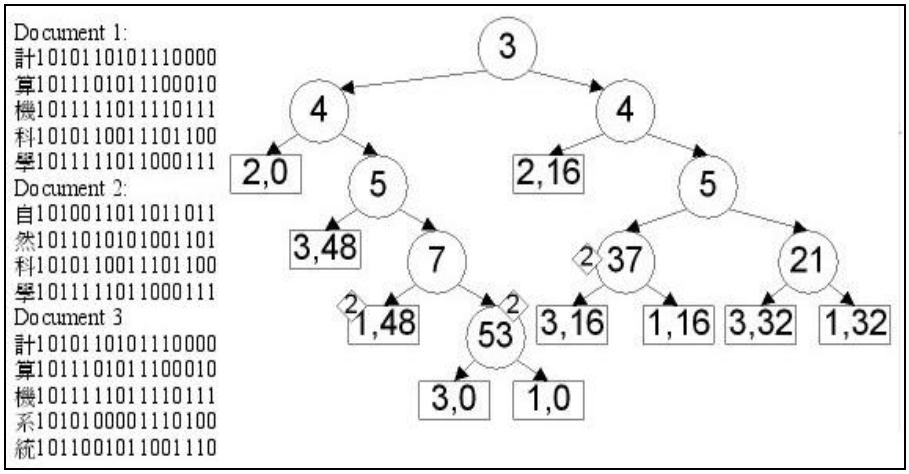
The frequency count indicates the frequency of the corresponding character string (Chinese phrase) that the node can represent. For the external nodes, it is simply the occurrence of their associated sistring, and for the internal nodes, it is defined as the total sum of the frequency count of their left and right child.

2.3.4 Pointers to the child nodes

Each internal node must have pointers to both left subtree and right subtree. The left subtree contains sistrings with value 0 in the check bit position, and the right subtree, on the other hand, contains sistrings with value 1 in the check bit position.

External nodes are leaf nodes. They do not have any child nodes and should not have such pointers. In other words, their pointers are always pointing to *null*.

Figure 3. The modified Chinese PAT Tree of three documents



2.4 Essential Node

Essential Nodes (*EN*) provide key information in the modified Chinese PAT-tree to identify the base cluster of the documents. A node (x) in the modified Chinese PAT-tree is an *EN* if and only if

(1) $Essential\ Length_x \geq 32$; and

(2) $Essential\ Length_x - Essential\ Length_y \geq 16$ (if such y exists),

where node (y) is the nearest ancestor of x such that y is also an *EN*.

$Essential\ Length_x$ is equal to the check bit of x truncated to the nearest Chinese character (16 bits).

Each *EN* represents a Chinese phrase, p, which equals to its linked sistring from bit-0 up to the *Essential Length*. By definition of the *EN*, we assume that each p contains at least two Chinese characters (32 bits) and is represented by at most one *EN*.

2.5 A Chinese PAT-tree Example

Figure 1 shows an example of the Chinese PAT-tree after inserting sistrings from three documents. Each of the documents contains the word "Computer Science" (計算機科學), "Natural Science" (自然科學), and "Computer System" (計算機系統), respectively. Circles are internal nodes with check bits indicated. Their links to the sistring are indicated in the content in their leftmost descendant. Rectangles are external nodes with a link to a sistring. Their check bit is equal to the length of their linked sistrings. For example, |3,48| indicates an external node with a link to the bit-48 (4th character) of document 3. Internal nodes with check

bit 37 and 53 are *ENs*. Their *Essential Length* is 32 and 48, representing a two-character phrase "Compute" (計算) and a three-character phrase "Computer" (計算機), respectively. Both of them have frequency count of 2 because these phrases exist in document 1 also. All external nodes here are *ENs*. The external node |1,48| represents a phrase, *p*, "Science" (科學) that has $frequency_p = 2$ and the others have $frequency_p = 1$.

In this example, we do not have sistring of single Chinese character in the modified Chinese PAT-tree. For example, |1,64|, |2,48| (學) and |3,64| (統) are not found in our modified Chinese PAT-tree example. This can greatly reduce the size of our data structure. Since those sistrings of single Chinese character is not informative for our clustering need, this can improve the searching time of the modified Chinese PAT-tree without affecting our clustering performance.

2.6 Discussion

Our Chinese PAT-tree design has several differences from the original PAT-tree design. We keep the structure of external nodes and internal nodes more uniform by introducing the check bit for external nodes and the link for internal nodes. With this uniformity, we introduce a new type of node called *Essential Node* (*EN*). With the link and check bit information, an *EN* can represent a unique Chinese phrase, *p*, for the clustering process. Furthermore, by our observation, single Chinese character is irrelevant for document clustering and they would never be found in *ENs*. Therefore, the size of our modified Chinese PAT-tree can be greatly reduced by ignoring those sistrings of single Chinese character. It is justified that our modified Chinese PAT-tree data structure is carefully design and optimized for the Chinese document clustering purpose.

3. CHINESE PAT-TREE CLUSTERING

Our modified Chinese PAT-tree clustering algorithm is a new method for Chinese document clustering. Although the framework is similar to STC [6], we use our proposed modified Chinese PAT-tree data structure as well as our ranking function in this framework. Our algorithm has four steps:

(1) The removal of punctuation and non-Chinese elements,

(2) The construction of the modified Chinese PAT-trees for each document (Document PATs) and a modified Chinese PAT-tree with all documents (Global PAT),

(3) The extraction of base clusters from Global PAT based on our own ranking function, and

(4) The grouping of base clusters into clusters resulting in a set of categorized documents.

Our algorithm is a linear algorithm. We assume that, with a set of *N* documents, the number of words per document is bounded by a constant. Since the construction and searching of a PAT-tree is logarithmic to the number of words in the document, they perform in linear time. The extraction and grouping processes are similar to those of STC, which is also a linear algorithm.

In the following, we focus on Step 2 to Step 4 of the algorithm. Step 1 is obvious and depends on the format of the raw documents.

3.1 Document PATs and Global PAT

We construct a Document PAT for each document. It is a modified Chinese PAT-tree that similar to the one in Figure 3, but it contains sistrings from a single document only. In a Document PAT, we define the term frequency, tf_p, [7] for each phrase, *p*, such that,

$$tf_p = \frac{frequency_p}{frequency_{max}},$$

where $frequency_p$ is the frequency of the effective node that represents *p*, and $frequency_{max}$ is the maximum of $frequency_p$ in the Document PAT.

Global PAT, as the example in Figure 3, contains sistrings of the set of *N* documents. For each phrase, *p*, in Global PAT, we define the inverse document frequency idf_p, such that,

$$idf_p = \log(\frac{N}{n_p}),$$

where n_p is the number of documents containing *p* [7].

Moreover, we define the total term frequency (ttf_p) and weight (w_p), such that,

$$ttf_p = \sum (tf_p),$$
$$w_p = ttf_p \times idf_p.$$

The calculation of w_p can be done in linear time because the value of n_p and tf_p can be found in the Document PATs in linear time.

3.2 Extraction of Base Clusters

The w_p in Global PAT indicates the relevancy of the word phrase, p. Common phrases usually have low idf_p. Insignificant phrases in the set of documents have low ttf_p. These phrases will produce low w_p.

There is a pitfall if p appears in one single document and has a high tf_p. Such p is not suitable for clustering, but may score a high ttf_p, idf_p, and w_p. Because of this, we further define an adjusted weight (w'_p), such that,

$$w'_p = w_p \times s_p,$$
$$s_p = \begin{cases} 1 & \text{if } n_p > 1 \\ 0 & \text{otherwise.} \end{cases}$$

We rank phrases in effective nodes with w'_p and select the top k phrases to form base clusters.

3.3 Grouping of Base Clusters

We use the single-link clustering algorithm to combine the base clusters. This method is similar to the one used in STC [6].

We denote the set of documents belongs to the base cluster m as B_m and the size of that cluster as $|B_m|$. When there are two base clusters B_m and B_n with size $|B_m|$ and $|B_n|$ respectively, they will be merged into a new cluster if and only if

$$\frac{|B_m \cap B_n|}{|B_m|} > 0.5,$$
$$\frac{|B_m \cap B_n|}{|B_n|} > 0.5.$$

After the above procedure, similar base cluster will be connected together and the result forms the final clusters for the set of documents.

3.4 Discussion

Since the framework of the algorithm is from STC, the performance of Chinese document clustering is comparable to the STC. First, our algorithm utilizes the Chinese PAT-tree structure that can efficiently identify the potential base cluster from the essential nodes. Moreover, we introduce the ranking function using the concept of term frequency and inverse document frequency. This ranking function can highlight the important phrase from frequency measurement but avoid some common phrases that appear everywhere. Furthermore, the ranking function is easy to change and modify. We introduce the binary variable s_p to further avoid unwanted phrase being rank top.

One preprocessing step we can take is to apply a stop word list to filter out some common words or domain-specific words that are not interesting for our clustering result. For example, the phrase "say" (表示) is a meaningless phrase that may always appear in some type of documents, the phrase "news" (新聞) is a domain-specific phrase that would not be interesting when the set of documents is actually a set of newspaper articles. In this case, result adjustment from stop word list can be easily applied before the ranking of the base clusters.

4. CLUSTERING EXPERIMENTS

We conduct a series of three clustering experiments to illustrate that our proposed modified Chinese clustering algorithm is feasible to organize a set of documents into groups. With the same collection of news documents, we prepare three data sets. The first data set (S_1) contains documents with full contents, the second data set (S_2) contains the headline and the first paragraph of the documents, and the third data set (S_3) contains the headline of the documents only. These three data sets illustrate the result of our clustering algorithm when it is applied on the set of documents, paragraphs, and sentences respectively. After we justify the results of these three experiments, we notice S_2 can gives the best clustering result since the computer is faster than S_1 and the clusters results are more relevant than S_3. Hence, we conclude that our algorithm can cluster the set of documents more efficiently without much affecting the clustering result when we use the headline and the first paragraph.

4.1 Experiments Set-up

Our experiments are performed on an Ultra Sparc 5 Sun workstation with 128MB of memory running on SunOS v5.6. Our clustering algorithm is programmed in C language without any code optimization. Further experiments on computation time would be conducted in future.

Our experiments use the daily news articles from a Hong Kong local Chinese newspaper, MingPao. The set of articles is obtained from their Hong Kong local news category that published between February 11-17, 2001.

Table 4 shows the number of documents we collect on each date. To estimate the number of *ENs* in the document set of each day, we build the Global PAT for each of them. As a result, there are roughly 15,000 *ENs* in the set of articles on each day. In our experiments below, we use these articles as one set of raw information to perform the clustering. The set of articles has totally 251 documents and its corresponding Global PAT-tree contains 112,471 *ENs*. We notice that the number of *ENs* for the whole set of articles is much larger than the sum of the sets of articles on each day. It is because the union of the set of articles has more variations on phrases, *p*, than the individual daily sets.

Table 4. Number of articles in the collections

Date of Collection	Number of articles	Number of *Ens*
February 11	28	14,826
February 12	35	14,652
February 13	36	16,140
February 14	37	17,151
February 15	41	17,444
February 16	39	16,322
February 17	35	16,003
TOTAL	251	112,471

To prepare the data set for each of our three experiments, we pre-process the raw set of news articles by extracting the desired sections of the articles. The first data set (S_1) contains the full documents of the set of all the 251 news articles. The second data set (S_2) contains the news headline and the first paragraph of the raw articles. The third set (S_3) contains only the sentences of the headlines only. The number of *ENs* in Global PAT for each of the data set is shown in Table 5.

Table 5. Number of Essential Nodes in each set

Set	Description	Number of *ENs*
S_1	All sentences	112,471
S_2	Headline and 1st Paragraph	23,621
S_3	Headline Only	2,792

We perform our modified Chinese PAT-tree clustering with the procedures described in the last section. First, we remove all the non-Chinese information from the input data set in Step 1. Then, we build the Document PATs for each of the articles in the data set and a Global PAT for all articles in the data set in Step 2. From the Document PATs and the Global PAT, we prepare the necessary parameters for our ranking function to obtain the weight (w_p) of each phrase (p) in the *ENs* of the Global PAT. Result of the base clusters can be ranked and extracted according to the modified weight (w'_p), which is derived from w_p, in Step 3. In the last Step, final clusters are obtained by combining the based clusters with high similarities. Applying the input data set S_1, S_2, and S_3 in our algorithm, we obtain the clustering results as shown from Table 6 through Table 8.

4.2 Clustering Result on The Full Contents of The Articles

The result of clustering experiment using S_1 shows that the top cluster with topic "Police Investigation" (警方調查) that contains 77 articles, followed by the cluster with topic "Hong Kong" (香港) that contains 88 articles, and the cluster with topic "Ma-Yingjiu Taiwan" (馬英九 台灣) that contains 18 articles. Detailed results of the top 20 clusters are shown in Table 6. These results demonstrate the feasibility of our Chinese clustering technique for full-length documents. Interesting news topics are grouped together which formed clusters.

Table 6. Top 20 clusters using the full contents of the articles (S_1)

	Cluster Key / Topic	Cluster Size	New Articles
1	警方 調查	77	77
2	香港	88	62
3	馬英九 台灣 台北	18	3
4	梁錦松 曾蔭權 公務員 司司長 財政	39	18
5	司機	23	9
6	政府	66	18
7	銀行	16	1
8	學生 學校	36	5
9	問題	63	9
10	被告	11	2
11	消防	16	2
12	工作	52	5
13	服務	38	0
14	記者	72	6
15	市民	47	7
16	朋友	26	2
17	醫生 病人	19	1
18	認爲	75	4
19	行動	32	2
20	影響	37	3
	Others	15	15

4.3 Clustering Result on The Headline and The First Paragraph of The Articles

Instead of using the full contents of the set of articles, we conduct another clustering experiment with the same set-up using only the sentences in the headline and the first paragraph. As a result, the cluster with topic "Government" (政府) that contains 27 articles ranks top, followed by the cluster with topic "Hong Kong" that contains 41 articles, and the cluster with topic "Police Investigation" (警方 調查) that contains 46 articles. Detailed results of the top 20 clusters are shown in Table 7. Although the ranking of result of S_2 is different from S_1, clusters with interesting topics still remain. Comparing the clusters in Table 6 and Table 7, e.g. the cluster with topic "Police Investigation" (警方 調查), we notice that the cluster size obtained from S_2 is reduced. This means that some of these articles do not have the keyword in the headline or in the first paragraph. Although the result from S_2 may have chances to miss some relevant articles, we can effectively eliminate those irrelevant articles that mention the keywords only in some minor paragraphs. As the main idea of an article is usually presented in its title and the first paragraph, this experiment result demonstrate the possibility of using a part of articles for clustering to improve the computation time that can still obtain a reasonable clustering result.

Table 7. Top 20 clusters using the headline and the first paragraph (S_2)

	Cluster Key / Topic	Cluster Size	New Articles
1	政府	27	27
2	香港	41	36
3	警方 調查	46	38
4	司機	15	28
5	梁錦松	12	10
6	銀行	10	6
7	工作	15	2
8	馬英九	12	5
9	學生	12	7
10	情人節	11	3
11	交通	13	5
12	市民	18	8
13	行動	14	4
14	立法會	15	5
15	問題	17	4
16	病人	7	3
17	網上	6	0
18	曾蔭權 司司長 財政	16	4
19	台灣	9	0
20	貨櫃	8	0
	Others	76	76

4.4 Clustering Result on The Headline of The Articles Only

The last experiment is conducted with the same set-up but using only the headline of the articles. Detailed result of the top 20 clusters is shown in Table 8. Interestingly, the result of this experiment can only form very small size clusters. The top ranked cluster is with the topic "Ma-Ying-jiu" (馬英九) that contains 5 articles, followed by the cluster with the topic "Driver" (司機), and then the cluster with the topic "Youth" (青年). We discover that results from this experiment cannot efficiently form large clusters. Therefore, 188 articles cannot be included in any of the top 20 clusters. From this observation, we justify that by the document title alone, such as the headline of news articles, is not informative enough for our clustering process. A title is usually short and it is unable to fully describe the main idea of an article for our clustering purpose.

Table 8. Top 20 clusters using the headline only (S_3)

	Cluster Key / Topic	Cluster Size	New Articles
1	馬英九	5	5
2	司機	5	5
3	青年	5	5
4	少年	4	4
5	委會	4	4
6	網上	4	3
7	保安	4	2
8	病人	4	3
9	梁錦松	4	4
10	內地	4	4
11	櫃車	3	3
12	輻射 華仁	3	2
13	大磡村	3	3
14	公司	3	3
15	燒炭	3	3
16	衛署	3	0
17	首批 智能身分證	3	3
18	小馬哥	3	3
19	回收	3	2
20	項提名 奧斯	3	2
	Other	188	188

4.5 Discussion

Our experiments results show that the modified PAT-tree clustering approach we propose is feasible for the Chinese information clustering. The method is sensible to the common topics from the set of documents and it can form clusters of documents with the same topic.

We obtain positive results in the set-up using document digests such as S_2. Due to the lack of details from the news headline alone, we can only produce small size clusters in the particular experiment on S_3. By reducing the number of *ENs*, the runtime performance in experiment on S_2 is much better than that on S_1 while both results contain similar clusters in their top 20 ranks.

These results provide the support in adopting our Chinese information clustering technique on the Chinese web documents. Since web search engines supply results that contain the title and some concise descriptions,

our clustering algorithm can be applied to the post-processing part of the search engine to group the web searching results into clusters.

5. CONCLUSIONS

This paper proposes the modified Chinese PAT-tree structure that is optimized for Chinese document clustering. We define the *Essential Node* in our modified Chinese PAT-tree, which can be used to identify base clusters. We define and justify a weight function to rank and extract the base clusters from the modified Chinese PAT-tree.

Our proposed approach applies the modified Chinese PAT-tree on the framework of STC to perform clustering efficiently in Chinese documents. Experiment results from a collection of Chinese newspaper articles with variations on document type demonstrate the capability of our novel algorithm in Chinese document clustering.

We plan to integrate our algorithm into a Chinese search engine for clustering Chinese web documents, and perform evaluations on the web document clustering results in near future.

6. REFERENCES

[1] A. Chen, J. He and L. Xu, “Chinese Text Retrieval Without Using a Dictionary,” *SIGIR'97*, 42-49, 1997.

[2] D. Morrison, “PATRICA-Practical Algorithm to Retrieval Information Coded in Alphanumeric,” *JACM*, 15, 514-534. 1968.

[3] G.H. Gonnet, and R. Baeza-Yates, *Handbook of Algorithms and Data Structures in Pascal and C*, 2nd Ed., 1991.

[4] J.-Y. Nie, M. Briscbois, "On Chinese Text Retrieval," *SIGIR'96*, 1996.

[5] L.-F. Chien, “PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval,” *SIGIR'97*, 50-58, 1997.

[6] O. Zamir and O. Etzioni, “Web Document Clustering: A Feasibility Demostration, *SIGIR'98*, 1998.

[7] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

[8] T. Gaston, “New Indices for Text: PAT Trees and PAT Arrays,” in *Information Retrieval Data Structures & Algorithms*, Frakes and Baeza-Yates (eds.) Prentice Hall, 66-82, 1992.

[9] T.-H. Ong and H. Chen, “Updateable PAT-Tree Approach to Chinese Key Phrase Extraction Using Mutual Information: A Linguistic Foundation for Knowledge Management,” *Proceedings of the Second Asian Digital Library Conference*, Taipei, Taiwan, 1999.