Chinese Keyword Spotting Using Knowledge-based Clustering

Yong Xia Kuanquan Wang Mingwei Li
School of Computer Science and Technology
Harbin Institute of Technology
Harbin, China
xiayong@hit.edu.cn

***Abstract*—Content-based document image retrieval is a new and promising research area. Without OCR, document indexing directly based on image content is more general and convenient. However content-based Chinese document retrieval is difficult for the complexity of Chinese character structure and large class numbers. Few papers cover this issue, and this paper will focus on it. This paper presents a novel algorithm of knowledge-based clustering and gives a mechanism of serial batch clustering for large data set. Knowledge derives from an artificial document image collection. Chinese characters with high frequency are edited and synthesized to images automatically. Cluster IDs are adopted to index the characters. A Dream of Red Mansions, a famous classical Chinese literature work including near one million characters, is used to evaluate the performance of Chinese keyword spotting. Experimental results confirm the effectiveness of knowledge-based clustering and its application on Chinese keyword spotting.**

Keywords-component; Content-based Chinese keyword spotting; Knowledge-based clustering; Serial batch clustering; Document image synthesis

I. INTRODUCTION

At present, there are mainly two ways to retrieve document image[1][2]. The one is retrieval based on OCR and the other is retrieval based on image content. The former is a traditional and classical way, the basic idea is that document image is recognized firstly and then OCR'd text is adopted for retrieval[3]. The latter is a relatively new way that character recognition is omitted. Character feature is extracted firstly and then similarity comparison is done for retrieval or a special feature coding is generated for retrieval[4]-[8]. As for document image of low recognition rate, content-based retrieval is more effective.

Although a lots of researches on document image retrieval are reported, most of that focus on English document and very few researches refer to Chinese document[9][10]. Tseng et al. retrieve Chinese document based on OCR'd text in [9] and Lu and Tan give a way of content-based retrieval for Chinese document in [10]. In [10], weighted Hausdorff is adopted to search similar character. However these methods of content-based retrieval only evaluate the performance in some very small collections.

Recently, document image retrieval for large collection becomes a focus[11][12]. Especially, Google and Yahoo both declare that retrieving scanned document on large scale will be provided to the public in near future, and these scanned documents include historic books, historic handwritten documents, etc.. In [12], a very large collect including over ten million document pages is introduced and according to the latest news from [13], the number of document is over eleven million and that of document pages is over sixty million. It will be a big challenge for content-based document image retrieval.

Text retrieval on large scale has been used in market for many years, for example, the web retrieval platform from Google only takes about one-tenth of a second to search 10 billion documents[14]. So a simple and fast way to retrieve large image collection is to convert image to text. OCR is the most popular way for the conversion. But OCR requires complicated offline training and giant character image samples. The collecting and marking these samples are very time-consuming and labor-intensive. Besides, in application, many documents, especially historic documents, include some special fonts or character degradation that doesn't occur in the training samples. So designing a special text coding based on actual character image may be more general compared with retrieval based on OCR. As for English documents, some coding mechanisms are introduced in [5] and [6]. The robust font-insensitive feature is extracted and several ASCII codes are connected to mark a character segment. But it will become invalid for Chinese document. For the sake of complicated structure of Chinese character, font-insensitive feature is difficult to extract. In [15] and [16], a more general and simple character indexing based on clustering is introduced. The cluster ID is taken as character index. Two clustering algorithms including SOM(Self-Organizing Map) and BSAS(Basic Sequential Algorithmic Scheme) is used to index document image respectively in [15] and [16].

Compared to English character, the class number of Chinese character is much larger and the structure of Chinese character is also much complicated, therefore content-based indexing and retrieval of Chinese document image is more difficult and also a challenge for researches. This paper will focus on this area and a new way of document indexing using knowledge-based clustering is provided for Chinese document image collection. Prior knowledge is added to

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.162

actual document data in order to promote precision and speed.

The next section will introduce the mechanism of clustering and document indexing. Experimental results and conclusion will be given in section Ⅲ and Ⅳ.

II. KNOWLEDGE-BASED CHARACTER CLUSTERING

As for unsupervised clustering, there are three key problems, namely number of clusters K, cluster initialization and distance metric. Because of the complicated character structure and large class number, Chinese character clustering is very difficult. K-means algorithm is a famous partition-based clustering. It is simple, fast and especially efficient for clustering of large data set. The algorithm complexity of K-means is O(nkt), where n is the number of data points, k is the number of clusters, and t is the number of iteration. The number of Chinese characters used frequently is more than 3000, namely k>3000, which makes much trouble for clustering. Obviously, the dimension of feature also affects the computation complexity greatly. So low feature dimension is expected. But for the sake of the complicated structure and large class number of Chinese character, low feature dimension cannot distinguish all Chinese characters very well. In general, the dimension of feature is over 50, even several hundreds. Therefore, algorithm complexity of Chinese character clustering is very high. In addition, large data will cause more iteration circulations. In [17], an algorithm of single pass K-means is introduced, large data can be divided into lots of patches, and standard K-means clustering is done in different patch independently. In this paper, we present a similar algorithm, and a novel prior knowledge-based clustering is adopted and inserted to the head of patch queues. The clustering process is shown in Figure 1. The standard K-means is also used for clustering in patch in this paper.

A. Determination of Cluster Number

We can see from Figure 1 that the mechanism of clustering requires identical cluster number for each data patch. Otherwise, information of patch clustering cannot pass backward validly. Automatic detection of cluster number is very difficult and in this paper we will avoid to solve this problem directly. Prior knowledge will be adopted to alleviate this difficulty. Large document collection can cover almost all common characters, so national standard of character coding such as GB2312-80 is a very important clue for cluster number. The standard of GB2312-80 is prevalently accepted for common Chinese characters, and lots of commercial Chinese OCR engines obey this standard. According to this standard, the class number of Chinese character is 6763. But among them, 45 Chinese radicals are included, so the class number of valid Chinese character is 6718. Although the true number of cluster in actual document collection may less than 6718, initial cluster centers from prior knowledge can guarantee the performance of clustering doesn't deteriorate obviously.

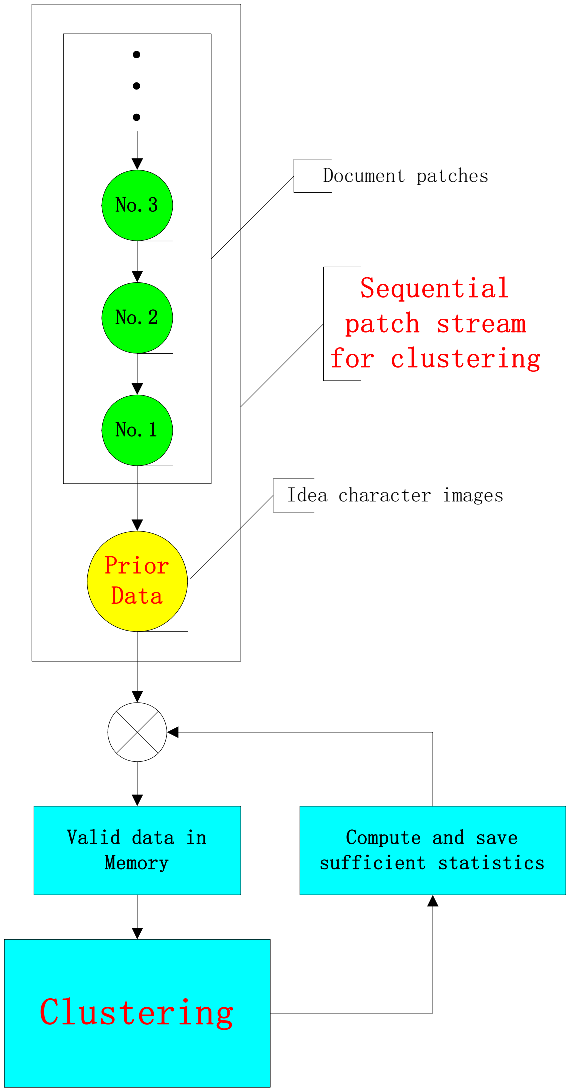


Figure 1. Knowledge-based serial batch clustering

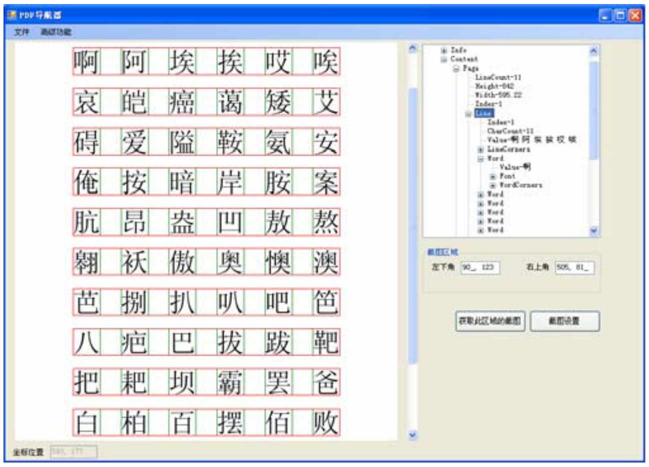
B. Knowledge-based Chinese Character Clustering

As for the algorithms of local optimization, such as K-means and so on, good original cluster points can improve the precision and decrease the iteration circulations. Prior knowledge can be used to improve the clustering initialization.

An artificial Chinese character image collection is built in this paper. Various Chinese characters of different fonts, sizes and resolutions can be converted to images automatically. These artificial character images are ideal and no-noise, which is different from actual character image. But they are still very similar. Furthermore, knowledge is adopted softly and clustering initialization is expected to be improved based on knowledge guidance.

A software platform is built to automatically convert text to image and extract and encode character image. The basic procedures are as follows. Firstly, all valid Chinese characters are input or copied to text editor, and special fonts, sizes and resolutions are set. Secondly, text is converted to PDF document based on PDF Plug-in. Thirdly, the coordinates of bounding box of character are extracted from PDF document and then character's snapshot is generated and saved in the corresponding character class lists. Figure 2 is a simple interface of synthesizing character image. The left of the figure is document image and the right is character's groundtruth. Because the synthesis of character image is fully automatic, the expense of time is very little, which is greatly different from collecting actual character images.

Because the groundtruth of artificial character image is valid, supervised clustering is done for synthesized data. All samples belonging to one class are averaged and the mean of each class is set as cluster center.



啊 阿 埃 挨 哎 唉
哀 皑 癌 蔼 矮 艾
碍 爱 隘 鞍 氨 安
俺 按 暗 岸 胺 案
肮 昂 盎 凹 敖 熬
翱 袄 傲 奥 懊 澳
芭 捌 扒 叭 吧 笆
八 疤 巴 拔 跋 靶
把 耙 坝 霸 罢 爸
白 柏 百 摆 佰 败

```
Info
Content
  Page
    LineCount=11
    Height=842
    Width=595.22
    Index=1
    Line
      Index=1
      CharCount=11
      Value=啊 阿 埃 挨 哎 唉
      LineCorners
      Word
        Value=啊
        Font
        WordCorners
      Word
      Word
      Word
      Word
      Word
```

截图区域

左下角 90_ 123　　右上角 505, 81_

Figure 2. Automatic generation of character image

C. Mechanism of Serial Batch Clustering

Both batch learning and online learning are used prevalently for clustering in application. Convergence of batch learning is much faster than that of online learning. K-means algorithm belongs to batch learning. But batch learning requires all data in memory and which is infeasible for large data set. So a mechanism of serial batch learning is built for clustering large document collection. The mechanism is shown in Figure 1.

All document data is divided into lots of patches according to a fixed data length. One patch once is loaded to memory and batch clustering is done. Sufficient statistics of this patch are computed and saved. Then this patch is unloaded and next patch is loaded to memory. The statistics from previous patch is also added to current patch for clustering. So information of clustering in previous patch can be passed to current patch. Therefore, the latest patch covers all data and clustering in this patch is approximately identical to clustering in whole data set.

Prior knowledge data can be seen as a special patch and is inserted to the head of the patch queues. This prior data patch is firstly loaded to memory and sufficient statistics are updated real time and saved. After accomplishing the patch clustering, the statistics are passed to next patch. In fact, this process is similar to cluster initialization. Taking prior data as a patch makes the system structure simpler.

Two sufficient statistics, including sum of character feature vectors in a patch and the number of feature vectors, are updated real time and passed to next patch. Based on these two statistics, we can get accumulated mean of cluster in any patch. When a data point is assigned to a cluster A, the sufficient statistics ($Sum^{(A)}, n^{(A)}$) will be updated as follows:

$$\begin{cases} Sum_j^{(A)} := Sum_j^{(A)} + x_j \\ n_j^{(A)} := n_j^{(A)} + 1 \end{cases}$$

Where j is the index of data update.

Furthermore, each patch clustering is independent, so an easy parallelization of algorithm is feasible. But in this paper, we will not discuss this problem too much and only attach importance to knowledge-based clustering.

D. Indexing and Retrieval of Document Image

After clustering, cluster centers will be saved as the cluster prototypes. Based on nearest neighbor algorithm, all character in document will be indexed with a cluster ID. Because the maximum of cluster ID is 6718, a cluster ID can be described by two bytes, which is identical to ASCII code of character. So document images are converted to ASCII-like text, and then text retrieval algorithms can be used for keyword spotting or document retrieval.

III. EXPERIMENTAL RESULT

A Dream of Red Mansions, also called as the Story of the Stone, is one of the four most famous Chinese classical literature works and multi-language versions are available all the world. Chinese version of this book is adopted to evaluate the performance of keyword spotting based on clustering. The number of Chinese characters in this book is near 1 million. The document is scanned to imaged document based on 200 dpi. Then the book is divided into lots of patches based on a fixed patch size of 50000 characters.

Layout analysis and character segmentation are done firstly, and then multi-scale directional element feature of 440 dimensions is extracted, which is an efficient feature and used in our previous work about OCR[18]. Only the directional element feature of pixel in contour is considered and four directions, including horizontal, vertical and ±45°, are adopted. The extraction of feature is described as follows. First, the character image is linearly normalized to a grid size of 64×64. Second, extract the character contour image. Third, the contour image is partitioned into 7×7 zones. For each zone, the elements with the same direction are accumulated. The dimension of feature based on this partition is 4×7×7=196. Next, similar to the above step, the contour image is partitioned into 6×6 and 5×5 respectively and subsequently the dimension of features is 4×6×6=144 and 4×5×5=100 respectively. Finally, combine serially all the feature vectors into a vector of 440 elements.

Prior knowledge is composed of documents of various fonts, sizes. Six fonts are considered in this paper, that is Song, Fang Song, Kai, Hei, Lishu, Youyuan. Prior knowledge data is inserted to patch queues from the head, and then the serial batch clustering for patches is done. Prior data clustering is supervised, other patch clustering is unsupervised and K-means is used for batch clustering.

Chinese keyword spotting is evaluated by F metric, and F is defined as follows:

$$F = \frac{2}{1/R + 1/P} = \frac{2RP}{R+P}$$

Where R is recall and P is precision.

As for prior knowledge, each font and font combination is respectively used for evaluation. In order to compare performance, random initialization rather than knowledge guidance is also tested. The performance of clustering for the first patch of the collection is shown in TABLE Ⅰ. A computer with Intel Core 2 Duo CPU 2.1GHz is used for evaluation.

R(Recall), P(Precision) and F are used to evaluate the performance for one character spotting. The cost of time for patch clustering is also given in this table. From this table, we can see that knowledge guidance-based clustering is much better than clustering without knowledge and random initialization. When appropriate prior data is provided, both recall and precision are improved greatly and the speed of clustering increases to three times of clustering without knowledge.

The sixth font, Song, is the best for this collection. The reason is that this font is similar to actual font used in the book. Font combination is the second best. Therefore, when the font in scanned document is recognized or known beforehand, the synthesized document with this font is appropriate for building prior data. But in fact, the fonts in document are often unknown and multiple fonts tend to scatter in document. So the knowledge from the font combination is more general and effective for application.

TABLE I. RESULT OF KEYWORD SPOTTING WITH ONE CHARACTER

Knowledge	R(%)	P(%)	F(%)	Time(s)
Hei	94.01	90.78	92.37	1191
Kai	95.17	93.50	94.33	1029
Fang Song	92.56	62.74	74.79	1904
Li Shu	93.91	87.36	90.52	1334
You Yuan	93.09	70.67	80.35	1830
Song	97.31	97.31	97.31	732
Font Comb.	**97.07**	**96.79**	**96.93**	**876**
Rand. Init.	66.08	66.07	66.075	2542

Fifty Chinese words of high frequency are submitted to retrieval system for performance evaluation, and the length of word is 2 to 4. The knowledge from all samples of six fonts is used for keyword spotting. The results of keyword spotting are shown in TABLE Ⅱ.

From the table, longer the word is, less the recall is. This is reasonable and understandable because longer word requires more character matching. The average performance of keyword spotting is acceptable with recall 96.7%, precision 98.3% and F 97.5%.

The time of clustering and indexing the whole book is about five fours. The time can be decreased if computer with high speed is used.

TABLE II. RESULT OF KEYWORD SPOTTING WITH VARIOUS WORD LENGTH

Word length	2	3	4	Average
Recall	97.8	97.1	95.3	96.7
Precision	97.3	98.2	99.5	98.3
F	97.5	97.6	97.5	97.5

IV. CONCLUSION

Algorithm complexity of clustering is closely relative to cluster number. Class number of Chinese character is very large and which makes much trouble to character clustering. Furthermore, large data set for clustering increases the difficulty. In this paper, we present a mechanism of serial batch clustering for large data set. Besides, prior knowledge is adopted to promote the precision and speed of clustering. Experimental results show that clustering-based indexing is effective for Chinese keyword spotting and knowledge-based clustering is better than clustering with random initialization.

Parallelization of algorithm isn't discussed deeply in this paper, but it is effective according the mechanism of patch clustering and will be considered in future research. In addition, only one simple clustering algorithm of k-means is considered for batch learning. Our future work will cover more clustering algorithms and verify the feasibility of knowledge-based clustering and the mechanism of serial batch clustering. Furthermore, retrieval of degraded Chinese document and larger document collection will be considered in future research.

ACKNOWLEDGMENT

This project was supported by China Postdoctoral Science Foundation (20090450994), Heilongjiang Postdoctoral Grant (LBH-Z09150), and the Fundamental Research Funds for the Central Universities (Grant No.HIT. NSRIF.2009152).

REFERENCES

[1] A.Murugappan, B.Ramachandran, P.Dhavachelvan. A survey of keyword spotting techniques for printed document images. Artificial Intelligence Review, 2010, pp.1-18

[2] M.S.Shirdhonkar, M.B.Kokare. Document Image Retrieval: An Overview. International Journal of Computer Applications, 2010, 1(7):128-130

[3] H. Cao, A. Bhardwaj,V. Govindaraju. A probabilistic method for keyword retrieval in handwritten document images. Pattern Recognition, 2009,(42):3374-3382

[4] T. Zant, L. Schomaker, K. Haak. Handwritten-Word Spotting Using Biologically Inspired Features. IEEE Trans. PAMI, 2008, 30(11):1945-1957

[5] S. Lu, L. Li, C.L. Tan. Document image retrieval through word shape coding. IEEE Trans. PAMI, 2008, 30(11):1913-1918

[6] S. Lu, C.L. Tan. Retrieval of Machine-printed Latin Documents through Word Shape Coding. Pattern Recognition. 2008, 41(5):1816-1826

[7] K.Zagoris, K.Ergina, N.Papamarkos. A Document Image Retrieval System. Engineering Applications of Artificial Intelligence, 2010, 23(6):872-879

[8] A.I.Wagan, S.Bres, H.Emptoz. Word spotting in Alice's adventures underground using multi scale integral orientation features. Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, 2010, pp.417-423

[9] Y.H. Tseng, D.W. Oard. Document Image Retrieval Techniques for Chinese. Proceedings of the Fourth Symposium on Document Image Understanding Technology. Columbia Maryland, April 23-25th, 2001, pp.151-158

[10] Y. Lu, C.L. Tan. Chinese word searching in imaged documents. International Journal of Pattern Recognition, 2004, 18(2):229-246

[11] L. Vincent. Google Book Search: Document Understanding on a Massive Scale. Proceedings of 9th International Conference on Document Analysis and Recognition, Sept. 2007, Curitiba, Brazil, pp.3-7

[12] G. Agam, S. Argamon, O. Frieder, D. Grossman, and D. Lewis. Content-based document image retrieval in complex document collections. Proceedings of Document Recognition and Retrieval XIV, 2007, pp. 65000S-1 - 65000S-12

[13] The Legacy Tobacco Document Library (LTDL), University of California, San Francisco, 2011, http://legacy.library.ucsf.edu/

[14] Anil K. Jain. Data Clustering: 50 Years Beyond K-Means. Pattern Recognition Letters, June 2010, 31(8): 651-666

[15] S. Marinai, E. Marino, G. Soda. Font Adaptive Word Indexing of Modern Printed Documents. IEEE Trans. PAMI., 2006, 28(8):1187-1199

[16] W. Magdy, K. Darwish, and M. El-Saban. Efficient Language-Independent Retrieval of Printed Documents without OCR. Proceedings of the 16th edition of the Symposium on String Processing and Information Retrieval , August 2009, Finland, pp.334-343

[17] F. Farnstrom, J. Lewis, C. Elkan, Scalability for clustering algorithms revisited, SIGKDD Explorations, 2000, 2 (1) :51–57

[18] Y. Xia, B.-H. Xiao, C.-H. Wang, R.-W. Dai. Integrated Segmentation and Recognition of Mixed Chinese/English Document. The Ninth International Conference on Document Analysis and Recognition, 2007, vol. 2, pp.704-708