Tibetan Text Clustering Based on Machine Learning

Gui-xian Xu[1], Li-rong Qiu[1], Lu Yang[2]
[1]College of Information Engineering
Minzu University of China, Beijing, China

[2]Faculty of Computer Science, University of New Brunswick
Fredericton, New Brunswick, Canada
xuguixian2000@sohu.com

JDIM
Journal of Digital Information Management

ABSTRACT: *Tibetan information processing technology has been obtained some achievements. But it falls behind Chinese and English information processing. It still needs to be paid more attention. Text clustering has the potential to accelerate the development of Tibetan information processing. In this paper, we propose an approach of Tibetan text clustering based on machine learning. Firstly, the approach is to execute Tibetan word segmentation with Tibetan texts. Then feature selection and text representation are conducted. Finally, K-means and DBSCAN are adopted to deal with the text clustering. The experimental results present that DBSCAN has better performance for Tibetan text clustering. Text clustering systems are designed based on proposed approach. The study is meaningful for the Tibetan text classification, information retrieval as well as construction of high-quality Tibetan corpus.*

Subject Categories and Descriptors
I.5.3 [Clustering]: Algorithm; **I.7 [Document and Text Processing]**

General Terms: Information Clustering, Knowledge Management

Keywords: Tibetan Information Processing, Machine Learning, Text Mining, Tibetan Text Clustering

Received: 17 December 2013, Revised 27 February 2014, Accepted 6 March 2014

1. Introduction

With the development of the information technology, a large number of Tibetan electric texts have appeared quickly. Tibetan text information processing has got some achievements such as the study of the Tibetan encoding conversion, word segmentation, Tibetan corpus tagging, machine translation, construction of the electric dictionary, text classification. For example, [1] conducted the research on the segmentation unit of Tibetan Word. [2] proposed the method to recognize the Tibetan person name. [3] showed the design of Tibetan-Chinese-English dictionary for Tibetan-Chinese-English translation. [4] studied the printed Tibetan character recognition technology. [5] used column information of web pages to classify Tibetan Web pages. [6] proposed the classification method based on the class feature dictionary for Tibetan web pages classification.

Obtaining the knowledge is meaningful from Tibetan texts. Tibetan text mining technology can extract useful things from the texts more conveniently. As a method of text mining, text clustering can organize the texts effectively. It is an unsupervised machine learning approach and the process which divides automatically the texts into some meaningful clusters or classes. In every class, the texts have the certain similarity. The algorithm of clustering involves in some subjects such as mathematics, computer science, statistics and so on. Text clustering is studied widely on English and Chinese [7-10]. Tibetan

text clustering has been shown to be helpful to find the knowledge and has the potential to accelerate the development of Tibetan information processing. However a few researchers are focused on Tibetan text clustering.

In this paper, our research focuses on clustering algorithm. Based on it, we construct an automated Tibetan text clustering system. Our goal is to benefit the development of Tibetan information technology, such as information extraction, text classification, hot topic tracking, pattern recognition [11] and so on.In the following, we first introduce the related background. Then we describe the proposed approach. At last, we present the experimental results, and conclude our work.

2. Background

Text clustering has been widely studied in the recent years. It is essential to many tasks in text mining such as information retrieval, spying the hot topics event and so on. Some clustering algorithms are well known such as K-means, K-MEDOIDS, DBSCAN, CLARANS [12].

Some researchers focused on improving the methods of feature selection so that the performance of clustering could be achieved. [13] proposed an "*Iterative Feature Selection* (IF)" method to iteratively select features and performed text clustering on unlabeled data. The experimental results were effective on Web Directory data. [14] proposed Text Clustering with Feature Selection (TCFS). It incorporated the statistical method CHIR to identify relevant features iteratively, and the clustering became a learning process. The experiments showed that TCFS with CHIR had better clustering accuracy in terms of the F-measure and the purity in various real data sets. [15] proposed a novel Harmony K-means Algorithm (HKA) that dealed with document clustering based on Harmony Search (HS) optimization method. It was proved by means of finite Markov chain theory that the HKA converged to the global optimum. The experiment compared the HKA with other meta-heuristic and model-based document clustering approaches. Experimental results showed that the HKA algorithm had the good quality of clusters. Some researchers focused on the semantic study so that the different relations of terms were calculated. The accuracy would be enhanced. [16] proposed that using nonnegative matrix factorization conducted document clustering. The proposed method was evaluated on a few benchmark text collections. It was proved the performance was good. [17] discussed a way of integrating a large thesaurus and the computation of lattices of result clusters into common text clustering. It overcame the disadvantage of traditional text clustering that did not relate semantically nearby terms and couldn't explain how resulting clusters are related to each other.

As far as we know, a few researchers conduct Tibetan text clustering. We propose the approach to cluster the Tibetan texts rapidly.

3. The Proposed Approach

3.1 Tibetan Encoding Conversion

Tibetan character has many encoding styles, such as Tongyuan, Sambhota, Ban Zhida, Unicode and so on. For clustering the same Tibetan encoding files, we need to standardize the Tibetan encoding style. So we need to convert txt format files with other encoding styles to txt format files with Unicode style.

3.2 Clustering Data Set

The Tibetan domain experts classify the texts into some classes. The classes include Sport, Economic, Religion, Army, Math-Physics, Politics and so on. We select five text sets as experiment data sets of the text clustering. The text number of every class is shown with each data set in the Table 1.

Data set 1		Data set 2		Data set 3		Data set 4		Data set 5	
Class	Text number	Class	Text number	Class	Text number	Class	Text number	Class	Text number
Religion	20	Religion	10	Religion	15	Army	29	Religion	10
Math-Physics	16	Math-Physics	10	Math-Physics	34	Agriculture-Forestry	11	Living	10
Living	29	Sport	10	Agriculture-Forestry	11	Sport	20	Sport	20
Army	30	Art	10	Army	8	Math-Physics	20	Army	10
Art	17	Politics	10	Environment	11	Politics	10	Medicine	10
Total	112	Total	50	Total	79	Total	90	Total	60

Table 1. The text number of every class of each data set

3.3 Word Segmentation

Computer can't understand unstructured text information. Before conducting machine learning, it is important to split the texts into the words which computers can handle. So word segmentation is indispensable for the information processing. Word segmentation is easier in English because spaces, punctuations can be used to split text

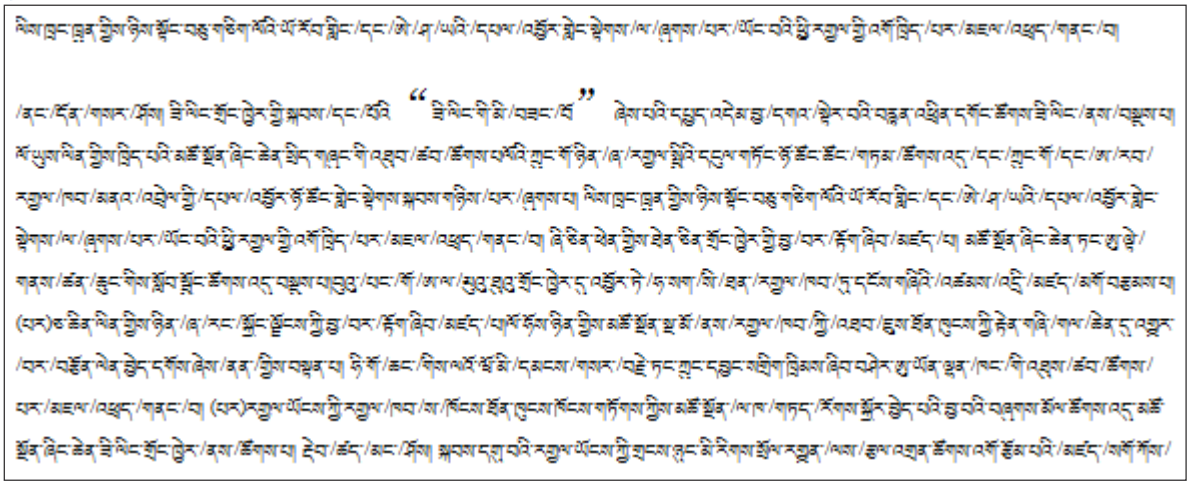

Figure 1. The example of word segmentation of a Tibetan text

information in English. However, Tibetan word segmentation is a difficult job and a great challenge because Tibetan doesn't contain the natural space separator.

Tibetan word unit is the syllables. The sentence is composed of the syllables. We use Tibetan electric dictionaryÿplace name dictionary, person name dictionary, case auxiliary words to split the Tibetan text. The word segmentation of a Tibetan text is shown in Figure 1.

3.4 Feature Selection and Text Representation

In order to use machine learning algorithm, we use vector space model to represent each text. Features are selected based on words appearing in the documents. We utilize the feature selection techniques to overcome the high-dimensionality disaster, e.g. Term Frequency (TF), Document Frequency (DF). We adopt TFIDF value as the feature weight. TF represents the frequency of word in a text, DF expresses the document number which contains the feature word, IDF is called inverse document frequency. $TFIDF = TF * IDF$.

Assume $D = \{d_1, d_2,\ldots, d_n\}$ is the text collection. $F = \{w_1, w_2,\ldots, w_{|F|}\}$ is the feature set of text collection. $|F|$ is the total number of features. Every text is transformed into a feature vector. Document d is represented as a vector $d = (d^{(1)}, d^{(2)},\ldots, d^{(|F|)})$. $d^{(i)}$ is the weight of feature w_i in document d. $d^{(i)} = \text{TFIDF}$.

3.5 Clustering Algorithm

We use the K-means and DBSCAN algorithms as the machine learning methods [12].

K-means is an approach based on the partitioning idea. K-means algorithm is as follows:

Input: k //k means the number of the clustering
Every object in the data set

Begin

1) Randomly select k objects to init the cluster centroids CEs [k].

//CEs [i] means the i-th cluster centroid, $1 < i <= k$.

2) Define Distance [i] [j], Distance [i] [j] = d. Distance [i] [j] is a distance matrix, it stores the distances of every object to each cluster centroid.

// Distance [i] [j] = d, it means that the distance is d between the object i and the cluster j.

// i expresses every object in the data set, $1 < j <= k$.

3) Belongs [i] expresses that the object i belongs to the cluster j. The reason is that the object i has the nearest distance with the j-th cluster centroid in all cluster centroids.

// i expresses every object number in the data set.

4) Update CEs according to step 3), recalculate the average centroid of each cluster.

5) If the CEs isn't changed, then stop the computation, output K clustering.

Else repeat step 2).

End.

DBSCAN is a density-based clustering algorithm. The basic idea of the algorithm is the data density is higher in the inner cluster than out the cluster.

DBSCAN algorithm is as follows:

Input: r //the radius length to determine if one data is a core point.

Num //the number of data in the circle whose radius is r.

Every object in the data set

Begin

1) Init the data set "DB": Each data is marked as non-core point and kept in DB. $DB = (D_1, D_2.... D_i,.... D_n)$.

Figure 2. The clustering tool

2) Init the core-point set DBC = null.

//DBC is the core-point set.

3) For all the data of DB

4) Compute the distances of D_i and all other points.

5) Count the number of the points that their distances from D_i is less than r in all other points.

// D_i expresses every object in the data set.

6) If (the points number of D_i) > Num

7) Then set D_i as a core point, put D_i into DBC.

8) End For

9) If DBC is not empty, then merge these core points when their densities can be connected, and form a new clustering.

10) Output the clustering result.

End.

3.6 Experiment Assessment

The assessment of clustering and classification is similar. For assessing the experiment results of clustering, we use the average F [18]. Assume i is a clustering result set, j is a classification text set. P is defined as $P = Precision(i,j) = \frac{N_{ij}}{N_i}$, $R = Recall(i,j) = \frac{N_{ij}}{N_i}$. N_{ij} is the text number of the class j in cluster i. N_i is the text number in the cluster i, N_j is the text number in classification j. F value of classification j is defined $F(j) = \frac{2PR}{P+R}$. $F(j)$ is the max value of all cluster i of classification j in one run. The total F value of the j-th class is calculated as $F = \frac{\sum_{j}[|j| \times F(j)]}{\sum_{j}[|j|]}$. $|j|$ is the number of all texts in the classification j.

4. Experiment and Results

We design the clustering system by utilizing the above

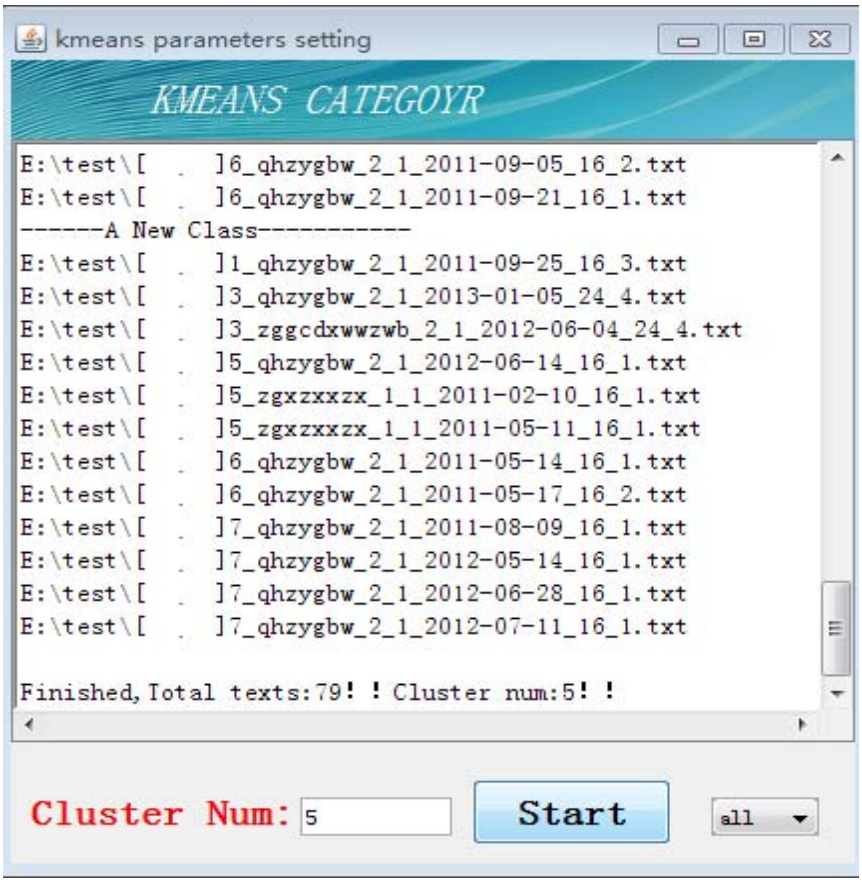

Figure 3. K-means interface

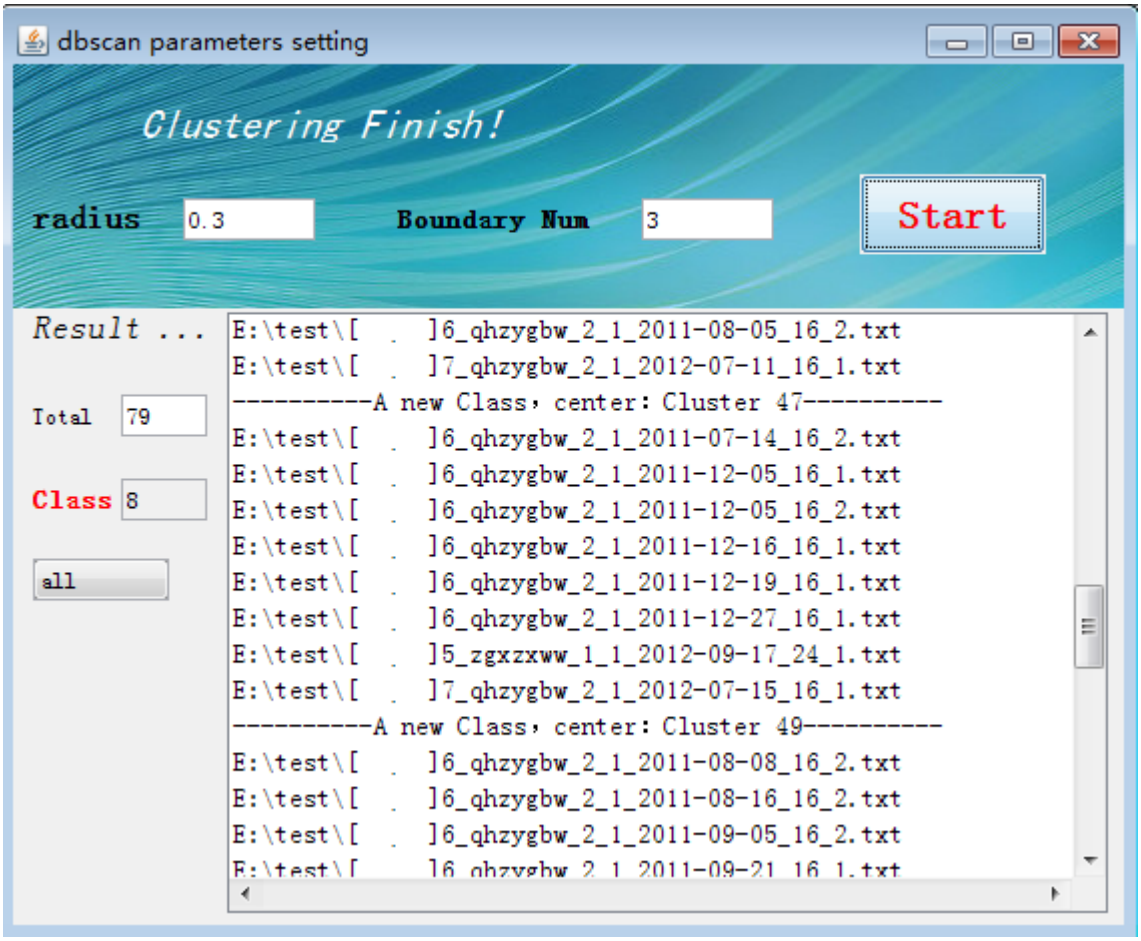

Figure 4. DBSCAN interface

approaches. Figure 2 shows the interface of clustering system. Using it, we can select the different clustering algorithm. Figure 3 shows the function graph of K-means. Through it, we can set K value of K-means. Figure 4 shows the function graph of DBSCAN.

We use five data sets shown in Table 1 to conduct the experiments. For K-means algorithm, we run respectively 5 times with $k = 5$ in every data set. For DBSCAN algorithm, we run respectively 5 times in every data set. The results are shown in the Table 2. Assessment method is F value introduced in 3.6.

Totally speaking, DBSCAN is better than K-means in terms of the performance. In K-means algorithm, F value of K-means is only 0.361 in the last data set, because the initial cluster centers are selected randomly, it affects the clustering results. The iterative process in this algorithm also makes every clustering result close to the optimal solution if the initial cluster centers are not very good. K-means will correct the cluster centers to some extent. From DBSCAN results, we can see that the best F value is up to 0.89. This method is good at finding the clusters of different shapes. The reason is that it is measured by the density of the relationship between clusters and can find the better clustering shape. It is not constricted by the clustering number.

	Data set 1	Data set 2	Data set 3	Data set 4	Data set 5	Average F
K-means ($k = 5$)	0.623	0.538	0.592	0.629	0.361	0.5486
DBSCAN	0.89	0.575	0.618	0.48	0.604	0.6334

Table 2. The F results of K-means and DBSCAN algorithm

We can also find the performance difference of the two algorithms in data set 4. In data set 4, the performance of DBSCAN is worse than K-means. This is because the texts have many differences in the Army class of data set 4. K-means algorithm is good at finding globular clusters while DBSCAN is showed that it is better at finding irregular shape clustering. So the difference is generated.

The clustering performance is affected by the Tibetan word segmentation. If the Tibetan word segmentation accaracy is improved, the clustering performance will be enhanced.

5. Conclusions

In this paper, we have reported the research on Tibetan text clustering based on machine learning. The results indicate that it can cluster a large number of unlabeled Tibetan texts into some classes rapidly. It is helpful and meaningful for Tibetan text classification, information retrieval as well as construction of high-quality Tibetan corpus.

6. Acknowledgements

This work is supported by "*the National Natural Science Foundation of China* (No.61309012)".

References

[1] Bai, Guan. (2010). Research on the Segmentation Unit of Tibetan Word for Information Processing. J*ournal of Chinese Information Processing*, 24 (3) 124-128. (In Chinese)

[2] Rong, Dou., Yangji, Jia., Wei, Huang. (2010). Automatic recognition of tibetan name with the combination of statistics and regular. *Journal of Changchun Institute of Technology* (Social Science Edition), 11 (2) 113-115. (In Chinese)

[3] Xiangzhen, He., Hongzhi, Yu., Jiang, Shen., Hui, Cao. (2012). Structural Design and Implementation of Tibetan-English-Chinese Electronic Dictionary. *Advances in Intelligent and Soft Computing*, 165 (1) 497-504.

[4] YongZhong, Li., YuLei, Wang., ZhenZhen, Liu. (2012). Study on printed Tibetan character recognition technology. Journal of Nanjing University(Natural Sciences), 48 (1) 55-62. (In Chinese)

[5] Guixian, Xu., Chuncheng, Xiang., Yu, Weng., Xiaobing, Zhao., Guosheng, Yang. (2011). Automatic Text Classification Approach of Tibetan Web Pages Based on Column. *Journal of Chinese Information Processing*, 25 (4) 20-23. (In Chinese)

[6] Guixian, Xu. (2013). Tibetan Web Pages Classification. *Journal of Convergence Information Technology*, 8 (1) 8-15.

[7] Changqiong, Shi., Hui, Huang., Dawei, Wang., Lalin, Jiang., Zongwen, Fu. (2009). Web text classification algorithm fused LSI and SVC. *Application Research of Computers*, 126 (112) 4523-4525.

[8] Jigui, Sun., Jie, Liu., Lianyu, Zhao. (2008). Clustering Algorithms Research. *Journal of Software*, 19 (1) 48-61.

[9] Jian, Ma., Wei, Xu., Yonghong, Sun., Turban, E. (2012). An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection. IEEE Transaction on Systems, *Man and Cybernetics*, 32 (3) 784-790.

[10] Owoputi, Olutobi., O'Connor, Brendan., Dyer, Chris., Gimpel, Kevin., Schneider, Nathan., Noach A. Smith. (2013). Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. *In*: Proceedings of The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p. 380-390.

[11] Nechval, K. N., Nechval, N. A., Purgailis, M., Strelchonok, V. F., Berzins, G., Moldovan, M. (2011). New Approach to Pattern Recognition Via Comparison of Maximun Separations. *Computer Modelling and New Technologies*, 15 (2) 30-40.

[12] Han Jiawei, Micheline Kamber. (2007). Data Mining: Concepts and Techniques(Second Editons). Copyright @2007 Elsevier (Singapore) Pte Ltd.

[13] Liu Tao, Liu Shengping, Chen Zheng, Ma Weiying. (2003). An Evaluation on Feature Selection for Text Clustering. *In*: Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC.

[14] Li Yanjun, Luo Congnan, Soon M. Chung. (2008). Text Clustering with Feature Selection by Using Statistical Data. *IEEE Transctions on Knowledge and Data Engineering*, 20 (5) 641 - 652.

[15] Mahdavi, Mehrdad., Abolhassani, Hassan.Harmony K-means algorithm for document clustering. *Data Mining and Knowledge Discovery*, 18 (3) 370-391.

[16] Farial Shahnaz, Michael W. Berry, V.Paul Pauca, Robert J. Plemmons. (2006). Document clustering using nonnegative matrix factorization, *Information Processing & Management*. 42 (2) 373-386.

[17] Andreas Hotho, Steffen Staab, Gerd Stumme. (2003). Explaining Text Clustering Results using semantic Structures. Knowledge Discovery in Databases:PKDD, Lecture Notes in Computer Science, 2838:217-228.

[18] Chen Jiayong. (2009). The Research and Implementation of Text Clustering Based on WEKA. *China Management Informationization,* 12 (2) 9-12. (In Chinese)