Clustering Research across Tibetan and Chinese Texts

G. X. Xu[1*], W. Sun[1], X. P. Peng[2]
[1] School of Information Engineering
Minzu University of China
Beijing, 100081, China
[2] University of Technolgy, Sydney, Australia

Journal of Digital
Information Management

ABSTRACT: *Tibetan text clustering has potential in Tibetan information processing domain. In this paper, clustering research across Chinese and Tibetan texts is proposed to benefit Chinese and Tibetan machine translation and sentence alignment. A Tibetan and Chinese keyword table is the main way to implement the text clustering across these two languages. Improved K-means and improved density-based spatial clustering of applications with noise (DBSCAN) algorithm are proposed. Experiments show that improved K-means algorithm gains stable text clustering result and performs better than traditional K-means after eliminating the limitation of random selection of initial k data. The improved DBSCAN algorithm obtains good performance through reasonable parameter setting. Improved DBSCAN performs better than improved K-means. The study is helpful and meaningful for the parallel corpus construction of Chinese and Tibetan texts.*

Subject Categories and Descriptors
I.5.3 [Clustering]: Algorithm; **I.2.7 [Natural Language Processing]:** Text Analysis

General Terms: Algorithm, Performance

Keywords: Text Clustering, Tibetan Information Processing, Cross-language Clustering

Received: 20 January 2015, Revised 24 February 2015, Accepted 5 March 2015

1. Introduction

With the development of information technology, Tibetan text information processing has achieved remarkable development. For example, Ref. [1] studied on the word segmentation of Tibetan. Ref. [2] proposed the method to recognize the Tibetan person name. Ref. [3] studied the printed Tibetan character recognition technology. Ref. [4] proposed the classification method of Tibetan web pages classification. However, compared with information technology in Chinese and English, the Tibetan version is relatively underdeveloped and needs to be explored. Tibetan word segmentation, text classification, clustering, hot topic tracking, and public opinion monitoring are areas that attract researchers and need further investigation.

Tibetan text mining technology can find implicit, previously unknown, and potentially useful knowledge from texts. Text clustering is a method of text mining that can organize the texts effectively [5]. This method is an unsupervised machine learning approach that automatically divides the texts into several meaningful clusters. Clustering algorithm is involved in various fields, such as mathematics, computer science, and statistics. Text clustering is studied widely in the English and Chinese languages. The clustering research across Chinese and Tibetan texts is helpful for Chinese and Tibetan machine translation and sentence alignment. However, only a few researchers have focused on the clustering research across Chinese and Tibetan texts.

In this paper, our research focuses on an improved clustering algorithm. Based on this algorithm, an automated Chinese and Tibetan text clustering system is constructed. Our goal is to develop a parallel corpus construction and machine translation method for Chinese and Tibetan texts.

In the following sections, we introduce the related background first. Then, we describe the proposed approach. Finally, we present the experimental results and conclusions of our work.

2. Related Work

With the increasing availability of electronic texts, text information processing is becoming an urgent task. Tibetan is an ancient language that has a long history [6]. To inherit this language effectively, developing Tibetan information processing is important. Text clustering has been widely studied in recent years. This method is useful for various tasks, such as retrieving information and discovering hot topics [7]. Several scholars have conducted research on Tibetan text clustering.

Ref. [8] studied the clustering algorithm that can find clusters in arbitrary shapes and can support the clustering of a large amount of texts. [9] 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. Ref. [10] investigated the clustering algorithm used in the field of architecture. Ref. [11] explored nonlinear clustering, and the result shows that calculating link matrix is a highly theoretical aspect in the development of clustering algorithm. Ref. [12] introduced a new clustering algorithm that has certain advantages over frequency term clustering (FTC) algorithm in terms of clustering results. Ref. [13] used clustering algorithms in image processing to recover archaeological relics, fragments of banknotes, and photos. Ref. [14] analyzed the economic and environmental level of the township region through clustering algorithms to infer the relevance of regional health expenditure and environmental protection. Therefore, the clustering research is highly useful in various fields. Chinese and Tibetan text clustering is more meaningful for constructing a parallel corpus of Chinese and Tibetan texts.

3. Proposed Approach

3.1 Construction of Keyword Table Across Chinese and Tibetan Languages

To conduct the Chinese and Tibetan text clustering, a keyword table in these two languages should be constructed first. Depending on subjects and thesaurus of China State Department documents, we select more than 3000 Chinese keywords from various classes, such as politics, law, history, sociology, economics, art, literature, and military. These Chinese words are then translated into Tibetan. A keyword table across the Chinese and Tibetan languages is composed of these translation pairs. Figure 1 shows an example of the keyword table across the Chinese and Tibetan languages.

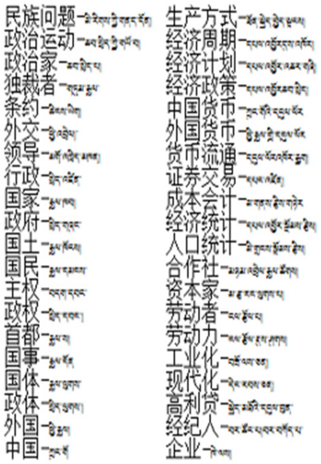

民族问题
政治运动
政治家
独裁者
条约
外交
领导
行政
国家
政府
国土
国民
主权
政权
首都
国事
国体
政体
外国
中国
生产方式
经济周期
经济计划
经济政策
中国货币
外国货币
货币流通
证券交易
成本会计
经济统计
人口统计
合作社
资本家
劳动者
劳动力
工业化
现代化
高利贷
经纪人
企业

Figure 1. Example of keyword table across Chinese and Tibetan languages

Part of speech	Suffix tag
Noun	/n
Time word	/t
Locative word	/s
Verb	/v
Quantifiers	/q
String	/x

Table 1. Part of speech and suffix tags left in Chinese word segmentation

3.2 Data Preprocessing

For Chinese texts, the stop words are first deleted from the text. The word segmentation is then conducted by ICTCLAS [15]. After segmentation, each word of the text is marked with the suffix of the part of speech. For example, the sentence "拉美国家庆祝五一国际劳动节" is divided into the following words and suffixes: "拉美/n," "国家/n," "庆祝/v," ""五一/m," "国际/n," and "劳动节/t." The parts of speech of the words include 16 types, such as "noun/n," "time word/t," "locative word/s," "verb/v," "direction word/f," "adjective/a," "adverb/d," "preposition/p," "pronoun/p," and "conjunction/c." However, only a few of them are meaningful to text clustering. The words with the suffixes "/n," "/t," "/s," "/v," "/q," and "/x" are retained to express the texts significantly. The rest of the words with other suffixes are deleted from the texts. Table 1 shows the suffix tags left and used to the text clustering. According to the rules, the above sentence can be saved as follows: "拉美/n," "国家/n," "庆祝/v," "国际/n," and "劳动节/t."

For Tibetan text segmentation, the processing procedure is simpler than that for Chinese text. The features are selected directly according to the Tibetan contents of the keyword table.

3.3 Text Representation

For Chinese texts, if the words are not in the keyword table, then these words are added into the keyword table to form the feature set. Vector space model is used to represent each text based on the feature set. *TFIDF* method is adopted to calculate the feature weight. *TF* represents the frequency of words in a text, *DF* denotes the document number that contains the feature word, and *IDF* is called inverse document frequency. $TFIDF = TF*IDF$.

We assume that $D = \{d_1, d_2, ..., d_n\}$ is the text collection.

$F = \{t_1, t_2, ..., t_{|F|}\}$ 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 $V(d) = (W(t_1, d), W(t_2, d), ...W(t_{|F|}, d))$. $W(t_i, d)$ is the weight of feature t_i in document d. Formula 1 shows how to compute $W(t_i, d)$.

$$W(t_i, d) = log\,(tf\,(t_i, d) + 1)\ \times log_2(n/_{df} + 0.01) \qquad (1)$$

n expresses the number of documents.

After normalization, Formula 1 is transformed into Formula 2 as follows:

$$W(t_i, d) = \frac{log\,(tf\,(t_i, d) + 1)\ \times log_2(\ n/_{df}\ + 0.01)}{\sqrt{\sum_{t \in d}[log\,(tf\,(t_i, d) + 1)\ \times log_2(n/_{df}\ + 0.01)]^2}} \qquad (2)$$

3.4 Clustering Algorithm

3.4.1 Improved K-means Algorithm

The traditional K-means algorithm has a few limitations [16]. One main drawback is the random initialization of cluster centers, which would cause the randomness and uncertainty of clustering results [17]. The improved algorithm mainly aims to improve the initial cluster centers. First, we sort the distances between the points in descending order. Weight of each point is calculated according to the result of the first step. The weight of word decides if the word becomes the initial cluster center. k top points are selected as the initial objects of the clustering center set. The distance of every selected point distance with other points must be greater than the distance threshold Q that is set by user. If the distance of the point does not meet the condition, then this object will be deleted from the clustering center set. Another point of the highest weight will then be added into the center set according to the distance threshold condition.

The main algorithm of the initial objects of clustering center set is as follows:

Input: k = number of clustering classes
Q = threshold value of the distance

Begin

1) Calculate the distance between each n object. The distance between objects x and x is 0. Given that the non-zero distance between two objects is symmetrical, the number of final distances is $\frac{n\,(n-1)}{2}$.

2) Sort the $\frac{n\,(n-1)}{2}$ distances in descending order.

3) The weights of the distances are given depending on the sorting order. The first distance is given the highest weight value n. The second distance is given the weight value $n-1$. Similarly, the last distance is given the weight value 1.

4) For n objects, calculate the weight sum between xi and other objects $i = 1, 2...n$.

5) Select k objects of the high weight sum as the initial objects of the clustering center set.

6) For k objects, check if each point distance between them is higher than the distance threshold Q. If the distance of the point does not meet the threshold condition, then this object will be deleted from the initial clustering center set. Another object with the highest weight then satisfies the distance threshold condition and will be added into the initial center set.

7) Output k initial clustering center.

End.

3.4.2 Improved DBSCAN Algorithm

Density-based spatial clustering of applications with noise (DBSCAN) algorithm is a density-based clustering algorithm based on the idea that the data density is higher in the inner cluster than in the outer cluster [18]. The DBSCAN algorithm does not need to input the class number. This structure may lead to scattered noises in various classes. In addition, more classes are produced. Improved DBSCAN attempts to avoid the influence of noise deviation on clustering center calculation [19]. The main idea is to dispatch the data of the noise classes to the higher-density classes to reduce the class number. This method compresses the distributed clustering results. The cluster center computing does not involve noises and does not cause clustering center deviation. This phenomenon results in a good clustering effect.

The improved DBSCAN method is as follows:

Input:
r = radius length to determine if one datum is a core point
$MinNum$ = number of data in the circle whose radius is r

Begin

1) In the data set "DB," each datum is marked as a non-core point and kept in DB. $DB = (D_1, D_2, ... D_i, ... D_n)$.

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

3) Count the number of points; their distances from D_i should be less than r in all other points.

4) If (the point number of D_i) > $MinNum$,

5) Then set D_i as a core point and input D_i into DBC. DBC is the data set of core points.

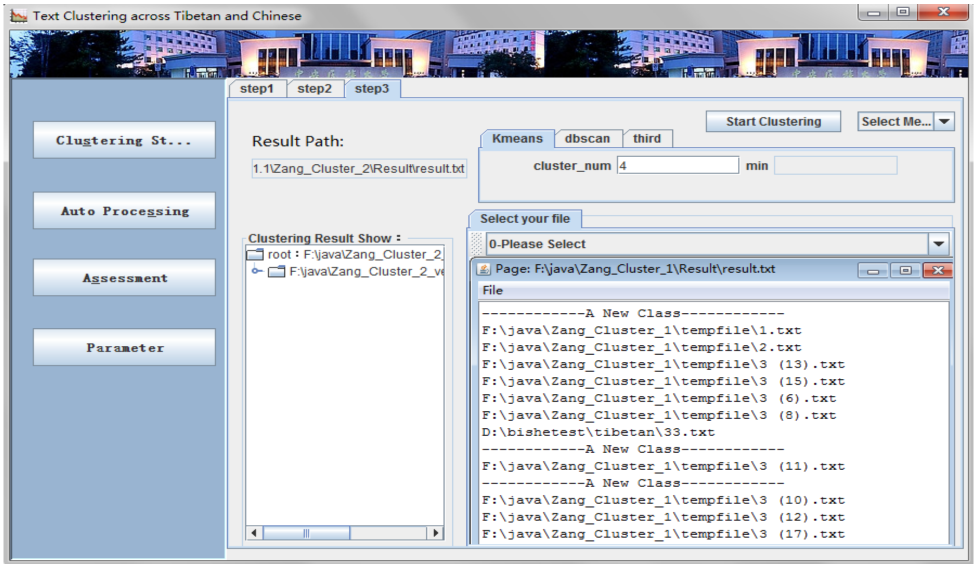
6) If D_i is not a border point,

7) Then input D_i into noise set NS.

8) If DBC is not empty, then merge these core points when their densities can be connected. Thus, clustering classes are formed and the cluster center $\overrightarrow{C_1}, \overrightarrow{C_1}, ...\overrightarrow{C_k}$.

9) For X_i in NS,

10) Compute X_i and $\overrightarrow{C_k}$ distance $(i = 1, 2, ... k)$, input X_i into the nearest class whose cluster center does not need to



Result Path: 1.1\Zang_Cluster_2\Result\result.txt

cluster_num 4

Clustering Result Show :

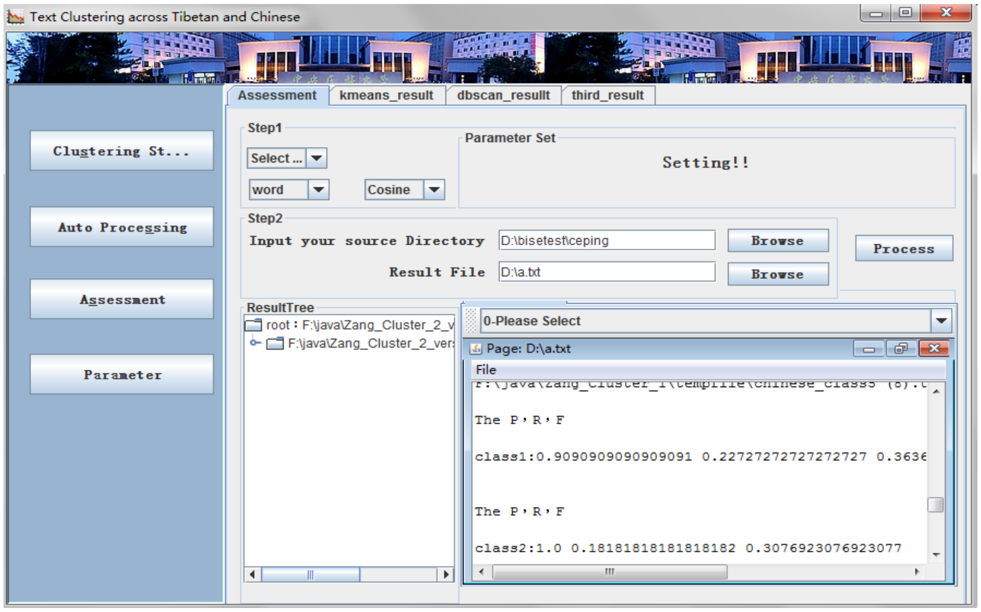
root : F:\java\Zang_Cluster_2_
F:\java\Zang_Cluster_2_ve

0-Please Select

Page: F:\java\Zang_Cluster_1\Result\result.txt

```
------------A New Class------------
F:\java\Zang_Cluster_1\tempfile\1.txt
F:\java\Zang_Cluster_1\tempfile\2.txt
F:\java\Zang_Cluster_1\tempfile\3 (13).txt
F:\java\Zang_Cluster_1\tempfile\3 (15).txt
F:\java\Zang_Cluster_1\tempfile\3 (6).txt
F:\java\Zang_Cluster_1\tempfile\3 (8).txt
D:\bishetest\tibetan\33.txt
------------A New Class------------
F:\java\Zang_Cluster_1\tempfile\3 (11).txt
------------A New Class------------
F:\java\Zang_Cluster_1\tempfile\3 (10).txt
F:\java\Zang_Cluster_1\tempfile\3 (12).txt
F:\java\Zang_Cluster_1\tempfile\3 (17).txt
```

Figure 2. Clustering software



Select ...

word

Cosine

Input your source Directory D:\bisetest\ceping

Result File D:\a.txt

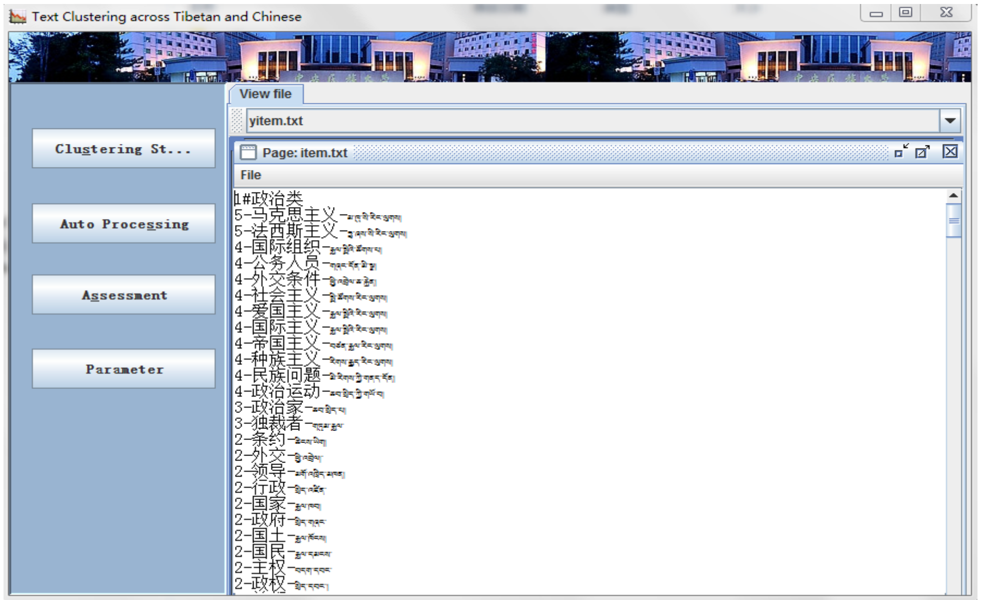
root : F:\java\Zang_Cluster_2_v
F:\java\Zang_Cluster_2_ver

0-Please Select

Page: D:\a.txt

```
F:\java\Zang_Cluster_1\tempfile\chinese_class5 (8).t

The P，R，F

class1:0.9090909090909091 0.22727272727272727 0.3636


The P，R，F

class2:1.0 0.18181818181818182 0.3076923076923077
```

Figure 3. Function of clustering assessment



yitem.txt

Page: item.txt

```
1#政治类
5-马克思主义 -མར་ཁེ་སི་རིང་ལུགས།
5-法西斯主义 -ཧྥ་ཤི་སི་རིང་ལུགས།
4-国际组织 -རྒྱལ་སྤྱིའི་ཚོགས་པ།
4-公务人员 -གཞུང་དོན་མི་སྣ།
4-外交条件 -ཕྱི་འབྲེལ་ཆ་རྐྱེན།
4-社会主义 -སྤྱི་ཚོགས་རིང་ལུགས།
4-爱国主义 -རྒྱལ་སྤྱིའི་རིང་ལུགས།
4-国际主义 -རྒྱལ་སྤྱིའི་རིང་ལུགས།
4-帝国主义 -བཙན་རྒྱལ་རིང་ལུགས།
4-种族主义 -རིགས་རྒྱུད་རིང་ལུགས།
4-民族问题 -མི་རིགས་ཀྱི་གནད་དོན།
4-政治运动 -ཆབ་སྲིད་ཀྱི་གཡོ་བ།
3-政治家 -ཆབ་སྲིད་པ།
3-独裁者 -གཏུམ་རྒྱལ
2-条约 -ཆིངས་ཡིག
2-外交 -ཕྱི་འབྲེལ།
2-领导 -མགོ་ཁྲིད་མཁན།
2-行政 -སྲིད་འཛིན
2-国家 -རྒྱལ་ཁབ།
2-政府 -སྲིད་གཞུང་
2-国土 -རྒྱལ་ཁོངས།
2-国民 -རྒྱལ་དམངས།
2-主权 -བདག་དབང་
2-政权 -སྲིད་དབང་།
```

Figure 4. The parameter setting interface

be computed again.

11) Output the clustering result.

End.

3.5 Experiment Assessment

The clustering assessment uses the method proposed in [20, 21]. Based on the assumption that i is a clustering result set, j is a classification text set. P is defined as $P = Precision\ (i, j) = \frac{N_{ij}}{N_i}$, $R = \text{Recall}\ (i, j) = \frac{N_{ij}}{N_j}$. Nij 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 as follows: $F(j) = \frac{2PR}{P+R}$. $F(j)$ is the maximum value of all cluster i of classification j in one run. The total F value of the j^{th} class is calculated as follows: $F = \frac{\sum_j [|j| \times F(j)]}{\sum_j [|j|]}$. $|j|$ is the number of all texts in the classification j.

4. Experiments and Results

According to the proposed algorithms, the application system software is developed. Figure 2 presents the clustering function of the software. Through this function, the algorithms of K-means, improved K-means, and improved DBSCAN can be conducted. Figure 3 shows the function of clustering assessment. The final clustering performance can be illustrated through this tool. Figure 4 shows the parameter setting interface. Through this function, the keyword table across the Chinese and Tibetan languages can be added, deleted, modified.

We use two data sets to conduct the experiment. Data set 1 includes 48 texts. Data set 2 includes 54 texts. Tables 2 and 3 show the detailed text distribution. In Tables 2 and 3, the data indicate the text number of each class (Chinese0text number + Tibetan text number). The total number of0classes is equal to the Chinese text number plus the Tibetan text number.

Table 4 shows the clustering performance of traditional K-means and improved K-means. Table 5 presents the clustering performance of improved DBSCAN. In Table 5, the first parameter of the bracket expresses radius and the second parameter expresses *MinNum*.

	political	law	history	social	economic
number	11 (6+5)	8 (0+8)	11 (6+5)	9 (5+4)	9 (4+5)

Table 2. Text number of every class in Data set 1

	political	law	history	social	economic
number	9 (4+5)	9 (0+9)	13 (4+9)	13 (4+9)	10 (4+6)

Table 3. Text number of every class in Data set 2

	Improved K-means	Traditional K-means (three times)		
P	0.5185	0.5185	0.5185	0.4629
R	0.5044	0.5262	0.4622	0.4180
F	0.5114	0.5223	0.4887	0.4393

Table 4. Clustering performance of K-means and improved K-means

	(0.1,3)	(0.2,3)	(0.09,3)	(0.08,3)
P	0.592	0.722	0.537	0.537
R	0.580	0.432	0.609	0.609
F	0.5859	0.6469	0.5707	0.5707

Table 5. Clustering performance of improved DBSCAN

	Improved K-means	Improved DBSCAN(0.1,3) (0.2,3)
P	0.5185	0.592
R	0.5044	0.580
F	0.5114	0.5859

Table 6. Comparasion of improved K-means and improved DBSCAN

As shown in Table 4, F values of traditional K-means are 0.5223, 0.4887, 0.4393 in three runs. F value of improved K-means is 0.5114. The clustering performance of the traditional K-means is unstable. The traditional K-means method obtains different results in three runs because of the randomness of the initial clustering center. One run of traditional K-means has the highest F value 0.5223 because the initial clustering centers are selected randomly and the distance of every pair of data is large in the initial clustering center. Its iterative process makes every clustering result close to the optimal solution. But the performances of other two runs are not good because the data in the clustering center are similar to some extent. This causes the iterative process not to find the optimal solution. The improved K-means algorithm eliminates the limitation of an uncertain initial clustering center and produces the explicit clustering center. The improved clustering result remains in a better level and is stable. P value is 0.5185. R value is 0.5044. F value is 0.5114. This algorithm has totally higher P, R, and F values than the traditional K-means.

The traditinal DBSCAN may produce more clustering classes for text clustering because of the data noise. After the improvement in noise processing, the result is not scattered. As shown in Table 5, the improved DBSCAN obtains a good clustering result. The parameter setting is important for the result. The optimal first parameter can be set through multiple testing. When the first parameter is set as 0.2, F is 0.6469 and the best value. But the difference value between P and R value is 0.29 and large, it isn't suitbale for the appilication. When the first parameter is set as 0.1, P value is 0.592, R value is 0.58, and F value is 0.5859. P, R, F values are ideal with the comparision of other parameter settings.

As shown in Tables 6, P, R, F values of improved DBSCAN are 0.592, 0.580, 0.5859 and P, R, F values of improved kmeans are 0.5185, 0.5044, 0.5114. It can be concluded that the performace of improved DBSCRAN is better than improved K-means through the comparision. Improved K-means still needs to input manually the class number before clustering. DBSCAN algorithm does not need to input the class number and conducts the clustering depending on the density of data. Improved DBSCAN avoids the influence of noise deviation on clustering center calculation and is better at finding irregular shape clustering. The improved DBSCAN is superior to the improved K-means in performance.

Through this study, the Chinese and Tibetan texts with similar topics are clustered. The clustering performance is affected by the quantity and quality of the Chinese and Tibetan keyword table. If the quantity and quality are improved, then the clustering performance is enhanced.

5. Conclusions

In this paper, we report the research on the clustering across Chinese and Tibetan texts. The improved K-means and the improved DBSCAN algorithms are proposed and used for text clustering across these languages. The experiment results indicate that improved DBSCAN approach has good performance. The study is helpful and meaningful for machine translation, particularly in information retrieval and sentence alignment across the Chinese and Tibetan languages.

Acknowledgement

This study is supported by "National Natural Science Foundation of China" (No. 61309012), "Beijing Social Science Fund"(No.14WYB040).

References

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

[2] Dou, R., Jia, Y. J., Huang, W. (2010). Automatic recognition of tibetan name with the combination of statistics and regular. *Journal of Changchun Institute of Technology*, 11 (2) 113-115.

[3] Li, Y. Z., Wang, Y. L., Liu, Z. Z. (2012). Study on printed Tibetan character recognition technology. *Journal of Nanjing University* (Natural Sciences), 48 (1) 55-62.

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

[5] Sun, J. G., Liu, J., Zhao, L. Y. (2008). Research on clustering algorithm. *Journal of Software*, 19 (01) 48-61.

[6] Chen, Y. Z., Yu, S. W. (2003). Tibetan information0processing technology research status and prospect.0*Chinese Tibetology*, (4) 97-107.

[7] Sambasivam, S., Theodosopoulos, N. (2006). Advanced data clustering methods of mining Web documents. *Issues in Informing Science and Information Technology*, (3) 563-579.

[8] Zengfang Yang., Xie, M. (2012). A Spatial Clustering Algorithm based on Neighbor Distribution. *Computer Science and Application*, 2 (2) 114-120.

[9] Mehrdad, M., Hassan, A. (2009). Harmony K-means algorithm for document clustering. Data Mining and Knowledge Discovery, 18 (3) 370-391.

[10] Yang, B. (2014). Analysis and algorithm of Web information clustering. *Silicon Valley*, (6) 53-53.

[11] Wang, C. D., Lai, J. H. (2013). Support Vector and Multi-Exemplar: Two Main Approaches for Nonlinear Clustering. *Hans Journal of Data Mining*, 3 (4) 41-49.

[12] Ponmuthuramalingam, P., Devi, T. (2010). Effective Term Based Text Clustering Algorithms. *International Journal on Computer Science and Engineering*, 2 (5) 1665-1673.

[13] Li, J. H., Kong, X. F., Fu, C. M. (2014). Scraps of

Paper Splicing Recovery Model. *Hans Journal of Data Mining*, (4) 1-8.

[14] Tang, Z., Gan, W. (2014). Influence Analysis of Environmental Quality on Health Expenditure of Urban Residents Based on Hierarchical Cluster. *Statistics and Applications*, (3) 76-84.

[15] Liu, Q., Zhang, H. P., Yu, H. K., Cheng, X. Q. (2004). Chinese Lexical Analysis Using Cascaded Hidden Markov Model. *Journal of Computer Research and Development*, 27 (1) 85-91.

[16] Zhou, L. J., Wang, H., Wenbo Wang, W. B., Zhang N. (2012). Parallel K-Means algorithm for huge amounts of data. *Journal of Huazhong University of Science andTechnology*, 40 (S1) 150-152.

[17] Han, J. W., Kamber, M. Fan, M., Xiaofeng Meng, X. F. (translators) (2001). Concepts and techniques of data mining. p. 288-289.

[18] Wang, G. Z., Wang, G. L. (2009). The improved DBSCAN algorithm. *Journal of Computer Applications*, 29 (9) 2505-2508.

[19] Feng, S. R., Xiao, W. J. (2008). A new method to improve the quality of DBSCAN clustering algorithm. *Journal of Xidian University*, 35 (3) 523-529.

[20] Chen, J. Y. (2009). The Research and Implementation of Text Clustering based on WEKA. *China Management Informationization*, 12 (2) 9-12.

[21] Xu, Guixian., Qiu, Lirong, Yang, Lu. (2014). Tibetan Text Clustering Based on Machine Learning. *Journal of Digital Information Magement*, 12 (3) 176-182.