A New Text Clustering Algorithm Based on Improved K_means

Li Xinwu
Electronic Business department, Jiangxi University of Finance and Economics, Nanchang, Jiangxi, China
Email: liyue7511@163.com

***Abstract*-Text clustering is one of the difficult and hot research fields in the internet search engine research. A new text clustering algorithm is presented based on K-means and Self-Organizing Model (SOM). Firstly, texts are preprocessed to satisfy succeed process requirement. Secondly, the paper improves selection of initial cluster centers and cluster seed selection methods of K-means to improve the deficiency of K-means algorithm that the K-means algorithm is very sensitive to the initial cluster center and the isolated point text. Thirdly the advantages of k-means and SOM are combined to a new model to cluster text in the paper. Finally the experimental results indicate that the improved algorithm has a higher accuracy compared with the original algorithm, and has a better stability.**

***Index Terms*—Text Clustering, K-means clustering, Self-Organizing Model**

I. INTRODUCTION

With the popularization and application of Internet, network has become an important part of the people's working and living, and various search engines have been an indispensable tool to retrieve the necessary resources for the people. However, the Internet search engine can often find thousands of search results. Even if some useful information is obtained, it is often mixed with a lot of "noises" to waste the users' time and money. Therefore, in order to efficiently and economically retrieve the resource subset relevant to the given search request and with the appropriate number, the Text clustering is performed and becomes one of important and hot research fields in data mining[1].

Text clustering is different from Text classification. The latter has them for each category while Text clustering has no category annotates in advance. The Text clustering is to divide the text sets into several clusters according to the Text contents, and requires the similarity of the Text contents in the clusters as great as possible and that of different clusters as small as possible. It can organize the Web Text effectively, but also form a classification template to guide the classification of the Web Text. Therefore, the Text clustering can do the online information clustering based on the contents to facilitate the retrieval and reading[2].

Manuscript received January 1, 2011; revised June 1, 2011; accepted July 1, 2011.

Since the 1950s, a variety of clustering algorithms has been proposed, is divided into partition-based and hierarchy-based algorithms broadly and also includes the mixed clustering algorithm combined with two ideas. In the partition-based clustering algorithm, K-means algorithm is the most famous. K-means algorithm includes K-means, K-Modes and K-Prototypes basically, of which, K-means algorithm is used for numerical data, K-modes for attribute data, and K-prototypes for mixed numerical and attribute data[3,4,5,6,7].

K-means algorithm features quick clustering and easy operation, and is applied to the cluster analysis of several data such as Texts, images and others; but this algorithm tends to terminate iterative process quickly to only obtain a partial optimal results, and fluctuate the clustering result because of random selection of the initial iterative center point. Due to the fact that the clustering is often applied to the data of the cluster quality the end-users can't judge and this fluctuation is difficult to be accepted in the application, it is of great significance to improve the quality and stability of clustering results in the analysis of the Text cluster[3,4].

For K-means problems concerning the selection of initial point and the sensitivity of isolated point in the Text clustering, this paper combines the advantages of SOM and K-means algorithms to enhance the stability and quality of the Text clustering of the algorithm.

II. TEXT PREPROCESSING

Text clustering can be described as: a given Text set $D=\{d_1,d_2,...,d_n\}$ eventually gets a cluster's set $C=\{C_1,C_2,...C_n\}$, $\Upsilon_{i=1}^{k} C_i = D$ derives $\forall d_i (d_i \in D)$, $\exists C_j (C_j \in C)$ and $d_i \in C_j$), and also makes the objective function $Q(C)$ reach the minimum or maximum value, of which, n is total Text number, k is final clustering number, and $C_j \cap C_i \neq \phi, j \neq i$.

A. Characteristic Selection and Expression of Text

Vector space model (VSM) is commonly adopted to express each Text. In this model, each Text d is considered as a vector in a vector space. $tfidf$ is used as a measure of characteristic vector in this paper, and this measure gives the weight of each word t. See Formula 1 for the calculation of the weight.

doi:10.4304/jsw.7.1.95-101

$$tfidf(d,t) = tf(dt) * \log_2 \frac{N}{df(t)} \quad (1)$$

In formula 1, $tf(d,t)$ is the word frequency of word t in the Text d, $df(t)$ is all the Text numbers of word t contained in the Text set D, and N is total Text number. After the characteristic selection, Text $d \in D$ is the form of the vector, and the value of each dimension is the corresponding $tfidf(d,t)$ weight value, so the Text can be expressed as Formula 2.

$$d = \{(t_i, tfidf(d,t_i)) | 1 \le i \le m\} \quad (2)$$

Of which, t_i is the lexical entry, and m is the dimension of the characteristic vector. However, after the characteristic selection, m is still very large, thousands of dimensions at least and tens of thousands of dimensions at most while non-zero word frequency of each corresponding Text vector is very few, which makes Text VSM show the high-dimension and sparsity of the model.

B. Definition of Similarity

In this paper, cosine distance is used to measure the similarity between the texts and defines the similarity of two texts d_1 and d_2 in Formula 3.

$$Sim(d_1, d_2) = \cos(d_1, d_2) = d_1 \cdot d_2 \| d_1 \| \cdot \| d_2 \| \quad (3)$$

In order to reduce the impact of different length of the Texts on calculating the Text similarity, each Text vector has been integrated to the unit length. See Formula 4.

$$d = d / \|d\| = \frac{\{tfidf(d,t_1), fidf(d,t_2), ...fidf(d,t_m)\}}{\sqrt{\{tfidf(d,t_1)^2, fidf(d,t_2)^2, ...fidf(d,t_m)^2\}}} \quad (4)$$

Thus, $\|d\| = 1$, and the similarity of the cosine is the dot product of two Text vectors, that is, $Sim(d_1, d_2) = d_1 \cdot d_2$.

III. Combination of K-means and SOM

A. K-means Algorithm Principle

Steps for K-means clustering algorithm are[4] (see Fig.1):

(1) Select n objects as the initial cluster seeds on principle;
(2) Repeat (3) and (4) until no change in each cluster;
(3) Reassign each object to the most similar cluster in terms of the value of the cluster seeds;
(4) Update the cluster seeds, i.e., recompute the mean value of the object in each cluster, and take the mean value points of the objects as new cluster seeds.

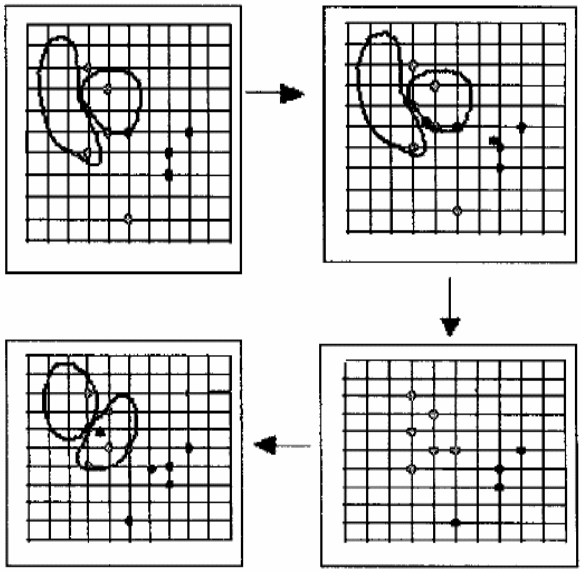

Figure 1. K-means algorithm procedures

B. Self-Organizing Model (SOM) Algorithm

SOM is the Self-Organizing feature Map proposed by Kohonen. Kohonen believed that, a nervous network's outer input receiving model was to divide the nervous network into different regions, these regions have different corresponding features to the input model, and such an input process is finished automatically[2]. The connecting weights of various neurons have a certain distribution, the nearest neurons excite each other, while the distant neurons inhibit each other, and the more distant neurons have a relatively weak inter-inhibition effect. In a word, Self-Organizing feature Map method is a teacher-free clustering method, and compared with the traditional model clustering methods, its former cluster centers could be mapped on a contour or plane, with the topological structure maintained original. Competitive Study refers to that various neurons at the same neuron level compete with each other and the winner neurons modify the connecting weights related with them. Competitive Study is a kind of study without supervision, and only some studying samples are required providing for the network during the study process, rather than the ideal output. The network finishes the self-organization according to the input samples and partitions them into the corresponding model categories. Due to no demand for the presentence of the ideal output samples, the supervised model classification method is promoted. Fig 2 represents the structural model of a competitive network.

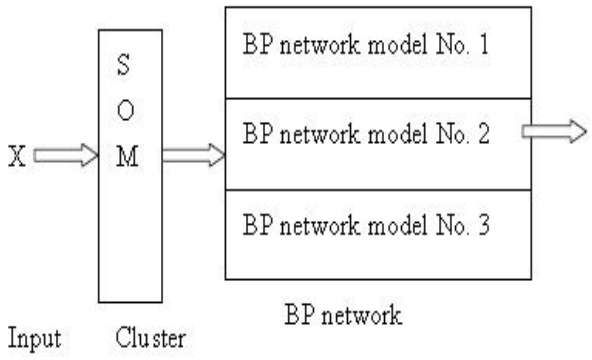

Figure 2.Competitive Network Topology

The competitive network consists of two levels, respectively input level and competition level. The input level of the competitive study network is used to receive the input samples, while the competition level is used to finish the classification of the input samples. The neurons at these two levels are fully interconnected, that is, each neuron at one level is connected separately with all neurons at the other level. At the competition level, the neurons compete with each other and eventually only one or several neuron activities adapt(s) to the current input samples. The neurons winning in the competition represent the classification model of the current input samples, each input node and each output node are connected through the connecting weight w, and the connecting weight $w_{i,j}$ of the node j at the output level to the node $x_i(i=1,2...N)$ at the input level is the cluster center of the j category. The model studying sample is composed by the actually measured samples of N classification indications. Proposed these study samples are all the points in N-dimensional space, it is obvious that the samples in the same categories or having some similar features are relatively close to each other in N-dimensional space. These relatively close samples compose a category and form a cluster in N-dimensional space. If the input samples belong to various categories, N-dimensional space will have a feature of several-cluster distribution. Each cluster represents one category, and the center of the cluster is the cluster center. The distance between the samples in the same category and the cluster center of this category is smaller than that between these samples and the cluster center of another category. Eyclid distance could be used to represent the distance, see Formula (5). In Formula (5), D_j represents Eyclid distance, x_i represents classification indication, w_{ij} represents the cluster center of the j category, and k represent iteration times.

$$D_{j(k)} = \sum_{i=1}^{N} (x_i - w_{ij(k)})^2 \qquad (5)$$

C. SOM & K-means Combination-Based Text Clustering

In the algorithm, the commonly used vector spatial model representation method is used to represent the texts, that is, use the characteristic items and their weights to represent the text information. The vector $d = (w_1, w_2, ...w_m)$ represents the characteristic items and corresponding weights of the text d, of which m represents the numbers of all items in the text set, $w_i(i=1,2,...m)$ represents the weight of the item t_i in the text d. To obtain the characteristic items, it is needed to firstly delete the unusable words from the text's characteristic set and then simplify the characteristic items according to TF-DF rules. In order to avoid the situation that an item obtains a large weight only due to its high appearance frequency (a tf larger value) in one text, Formula (6) is used to calculate he weights. In Formula (6), w_{ij} represents the weight of the j item in the i text, and $coef_{ij}$ could be obtained through Formula (7). While in Formula (7), tf_{ij} represents the appearance frequency of the j item in the i text.

$$w_{ij} = (coef_{ij}) \cdot (\log N - \log df_i) \qquad (6)$$

$$coef_{ij} = \begin{cases} 1 & if \quad tf_{ij} = 1 \\ 1.5 & if \quad 1 \pi tf_{ij} \le 5 \\ 2 & if \quad 5 \pi tf_{ij} \le 10 \\ 2.5 & if \quad tf_{ij} \phi 10 \end{cases} \qquad (7)$$

Thus, a group of vectors representing the text set is obtained, that is, the model sets to be clustered. The distance between the text vectors is represented by adopting the cosine distance, defined as Formula (8).

$$d(doc_i, doc_j) = 1 - sim(doc_i, doc_j) \qquad (8)$$

In Formula (8), $sim(doc_i, doc_j)$ could be calculated through Formula (9). $sim(doc_i, doc_j)$ is called as cosine similar function, and the bigger its value is, the more similar the text i and j are, thus, the smaller the cosine distance between these two texts is.

$$sim(doc_i, doc_j) = \frac{\sum_{k=1}^{m} (w_{ik} \cdot w_{jk})}{\sqrt{\sum_{k=1}^{m} (w_{ik})^2 \cdot \sum_{k=1}^{m} (w_{jk})^2}} \qquad (9)$$

The main process to cluster the texts by using the clustering combination algorithm of SOM and K-means (SOMK) is below: firstly apply the commonly used vector spatial model to represent the text information, delete the unusable words with the conventional method,

and simplify the characteristic items according to TF-DF rules to obtain the text's characteristic set, secondly calculate the weights of various characteristic items and express the text in the form of vectors, thirdly input the vectors of the text set for SOM algorithm and cluster the texts through SOM (the number of SOM network's input nodes equals to the dimension of the text vectors, while the number of SOM network's output nodes equals to the number of the texts' categories) to obtain a group of output weights, and finally initialize K-means algorithm's cluster centers with this group of weights and implement K-means algorithm to cluster the text sets.

Ⅳ. IMPROVING K-MEANS

A Limitation of Initial Algorithm

When K-means algorithm is used to cluster data, the stability of the clustering results is still not good enough, sometimes, the clustering effect is very good (when the data distribution is convex-shaped or spherical), while sometimes, the clustering results have obvious deviation and errors, which lies in the data analysis. It is unavoidable for the clustered data to have isolated points, referring to the situation that a few data deviate from the high-dense data intensive zone. The clustering mean point (geometrical central point of all data in the category) is used as a new clustering seed for the K-means clustering calculation to carry out the next turn of clustering calculation, while under such a situation, the new clustering seed might deviate from the true data intensive zone and further cause the deviation of the clustering results. Therefore, it is found that using K-means algorithm to process the data of isolated points has a great limitation.

When use combination of K-means and SOM, the paper improves K_means including improving selection methods of initial cluster centers and improving cluster seed selection to enhance the stability and quality of the text clustering of the algorithm..

B Improving Selection Methods of Initial Cluster Centers

When use combination of K-means and SOM, the paper improves K_means including improving selection methods of initial cluster centers and improving cluster seed selection to enhance the stability and quality of the text clustering of the algorithm..

The basic idea of new selection method of initial cluster centers is that on the assumption that the distribution of the Text sets has been known, a good initial cluster center should satisfy the follows in the paper:

1) The selected initial centers belong to different clusters respectively, that is, any two initial centers can not be the same cluster;

2) The selected initial cluster centers should represent this cluster, that is, be as close as possible to the cluster centers. To select k Texts as initial cluster centers and at the same time ensure that k Texts just belong to different clusters, such strict constraints are difficult to be achieved through random sampling as much as possible, so it is thought: in order to minimize the sampling's effect on initial cluster centers, m times of samplings are taken and the sample size is n/m, of which, n is the number of the Text in the Text sets, the value of m is that each sample size should be put into the main storage and as far as possible satisfies the fact that the sum of the samples taken for m times is equivalent to the original Text set. Each sample Text taken is clustered by k-means algorithm to produce a group of Text clusters with k cluster centers respectively; m times of sampling operation produce $m \times k$ cluster centers in all, and then agglomerative hierarchical clustering algorithm Single - link algorithm is used to do the clustering to obtain k clusters, of which, the average value is the final k initial cluster centers. Different from the division strategies taken by k-means algorithm, the agglomerative hierarchical clustering algorithm does not exist in the selection of the initial cluster centers. It regards each Text as a cluster at first, the Text is the centre of this cluster, and each step of clustering combines the two most similar clusters into a cluster until all the Texts are integrated into a cluster or only k clusters. With clustering, the similar Text is integrated into a cluster gradually and the hierarchical clustering is able to automatically generate different hierarchical clustering model.

In the combination of agglomerative hierarchical clustering algorithm and k-means algorithm, a hierarchical clustering algorithm based on k-means is addressed to select the initial cluster centers, that is, the cluster centers produced by k-means method restrain the agglomerative space of the agglomerative hierarchical clustering algorithm. The selection method of the initial cluster centers is generally described as followings:

1) m times of sampling are taken for the Text sets, which are divided into m sample sets $\{S_1, S_2, ..., S_m\}$;

2) Each sample set performs k-means algorithm respectively to produce m groups of k cluster centers;

3) Another clustering is done for $m \times k$ cluster centers by the agglomerative hierarchical clustering algorithm(Single-link algorithm is used here)until only having k clusters, and the average value of each cluster is taken as the initial cluster centers of next step of k-meanalgorithm.

From the above algorithm it is seen that the Text set of the sample taken is smaller than the original Text set, so the search process amount of the initial cluster centre is less, the iterative number is less and the speed is faster; at the same time it is also ensured that the final cluster centers belong to different clusters and have adequate representation.

C Improving cluster seed selection

When calculating the K turn of clustering seeds with the improved algorithm, those data in the cluster having a great similarity to the K-1 category seeds should be adopted to calculate their mean points (geometrical center) as the clustering seed of the K turn and the specific calculation method is below:

(1) For the cluster $c_{i(k-1)}$ obtained through the K-1 turn of clustering, the minimum similarity $sim\_\min_{i(k-1)}$ of the data in the cluster to the clustering seed $s_{i(k-1)}$ of the cluster is calculated;

(2) The data in the cluster $c_{i(k-1)}$ is calculated that has a similarity of more than $1-\beta*(1-sim\_\min_{(k-1)})$ to the clustering seed $s_{i(k-1)}$ (among, β is a constant between 0-1), and the data set is recorded as $cn_{i(k-1)}$;

(3) The mean points of the data in $cn_{i(k-1)}$ are calculated as the clustering seed of the K turn.

In Fig.3, (a) shows cluster i of the K-1 turn and its seeds, (b) shows cluster i of the K turn and the new seeds (initial algorithm), and cluster i of the K turn and the new seeds (improved algorithm).Indication of the symbols in Fig 1 can be listed as follows:▲ means data point i in the cluster,☆ means seed of the cluster i in the K-1 turn, ★ means new seed the cluster i in the K turn, ● means other data points, ○ means the points within its range are used to calculate the new seeds.

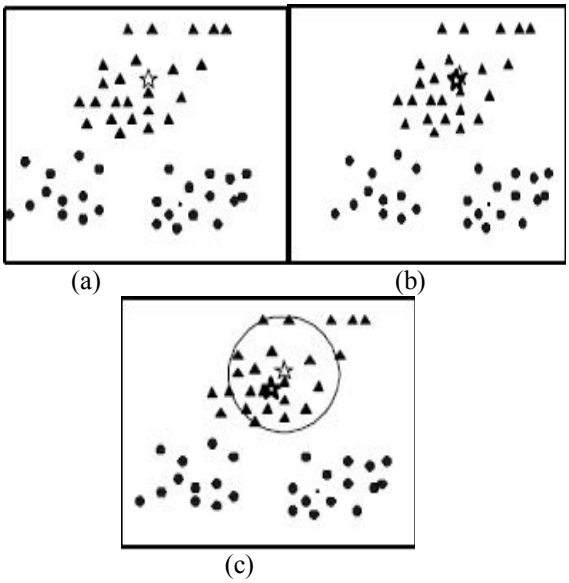

Figure 3 Comparing Pictures of k-means Algorithm and Its Improved Algorithm

As seen in Fig. 3, the new clustering seeds are obviously moving toward the data intensive zone. The improved algorithm could achieve a good clustering effect on the cluster sets containing isolated points. For the processing of big sets, this improved algorithm, as same as k-means algorithm, is relatively flexible and high-effective. Its time complexity is $O(nkt)$, of which, n is the number of all objects, k is the number of the clusters, while t is the iteration number of the algorithm, and generally, $k \langle\langle n$ and $t \langle\langle n$.

V. Experimental Verification

To test the effectiveness of the improved algorithm, the original algorithms[5] and the improved algorithms are compared. The experiment is made on the computer of the Celeron (R) 2.0G, 512M memory by VC++ and the experimental data is from www.China.com.cn and www.sohu.com. See Tab. Ⅰ for the corresponding experimental results of www.China.com.cn and www.sohu.com, of which, M_i is total Text number of category i; N_j is total Text number of cluster j; $M(n_{ij})$ is total Text number of category i included in cluster j when category j reaches the maximum F-measure value; $M(F(i,j))$ is the maximum value in category i and F-measure value of different clusters.

In Tab. 1, the data shows that adopting the improved clustering algorithm improves the accuracy of the clustering results. In order to verify the stability of the improved clustering algorithm result, multi-group data is used to perform the comparative experiment respectively by two algorithms to obtain 30 groups of experimental data. The F-measure value distribution in the experimental results is shown in Tab. Ⅱ

It can be seen from Tab. 2 that there is poor stability in the clustering results obtained by ordinary k-means algorithm and scattered F-measure value; but the improved clustering algorithm has better stability of the clustering results, more concentrated F-measure value and higher F-measure average value. The experiment shows that the improved clustering algorithm improves its accuracy and stability greatly. In the use of ordinary k-means algorithm, F-value of the clustering results scatters from 0.60 to 0.75; in the use of the improved algorithm, the stability of its value is from 0.75 to 0.85.

VI. Conclusion

Regarding the shortcoming that K-means algorithm has poor clustering quality and stability in the application of the Text clustering, this paper addresses a new Text clustering algorithm based on combination of k-means and SOM. The experimental results show such a clustering combination algorithm not only maintains the self-organizing features of SOM network, but also makes up the disadvantages of SOM network's overlong convergence duration and the bad clustering effect caused by the inadequate selection of K-means algorithm's initial cluster center.

TABLE I.
A SET OF COMPARISON CLUSTERING RESULTS

Cluster	Original K_means Algorithm				Improved Algorithm		
	M_i	N_j	$M(n_{ij})$	$M(F(i,j))$	N_j	$M(n_{ij})$	$M(F(i,j))$
Arts	44	40	20	0.48	39	28	0.68
Politics	83	80	53	0.65	81	61	0.74
Health	41	39	31	0.71	39	34	0.85
Sports	65	63	38	0.59	64	57	0.88
News	89	74	59	0.73	81	77	0.91
Culture	104	95	63	0.64	94	88	0.89
Education	112	98	74	0.71	102	96	0.90
Military	132	104	89	0.77	123	119	0.93
Science	144	123	98	0.74	137	132	0.94
Average	$F = 0.67$				$F = 0.86$		

TLBLE II
COMPARISON EXPERIMENTAL CLUSTERING RESULTS

F-measure interval	F-measure typical value	Original K-means algorithm F-measure value falls into the experimental frequency of this interval	Improved K-means algorithm F-measure value falls into the experimental frequency of this interval
[0.15,0.25]	0.20	0	0
[0.25,0.35]	0.30	1	0
[0.35,0.45]	0.40	2	0
[0.45,0.55]	0.50	4	0
[0.55,0.65]	0.60	6	0
[0.65,0.75]	0.70	9	8
[0.75,0.85]	0.80	1	12
[0.85,0.95]	0.90	0	8
[0.95,1.00]	1.0	0	0

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under the grant No.60963012 and is supported by the education department of Jiangxi Province (2007259).

REFERENCES

[1] A Likas, N Vlassis, J J Verbeek, "The global k - means algorithm," *Pattern Recognition*, vol. 36, pp. 451–462, February, 2007.
[2] Tao Li, "Document clustering via Adaptive Subspace Iteration," *Proceedings of the 12th ACM International*

Conference on Multimedia. New York: ACM Publishe, vol. 124, pp. 364–367, 2004.

[3] Hamerly G, Elkan C., "Learning the K in K_means," *Proceedings of the 17th Annual Conference on Neural Information Processing Systems(NIPS)*, vol. 231, pp. 281–289, 2008.

[4] Tang Yong， Rong Qiusheng, "An Implementation of Clustering Algorithm Based on K-means," *Journal of Hubei Institute for Nationalities*, vol. 22, pp. 69–71, January 2009.

[5] Gong Jing，Li Anming, "An Implementation of Clustering Algorithm Based on K-means," *Journal of Hunan University of Technology*, vol. 22, pp. 52–54, February, 2008.

[6] Yi S, "Global Optimization for NN Training," *IEEE Computers,* vol. 19, pp. 45–54, March, 2009.

[7] Zhang Yufang, Mao Jiali, "An improved K-means Algorithm," *Computer Applications,* vol. 23, pp. 31–33, August, 2010.

Li Xinwu, August 1973, male,1999 received MS from Xi'an Jiaotong University in Xian in china , majored in Signal and information processing, and received PHD from Northwestern Polytechnic University , majored signal and information processing, research on complicated factors modeling;

He is a professor in Electronic Business department in Jiangxi University of Finance and Economics now and his research interests include image process and text clustering.