Survey on Text Clustering Algorithm

-Research Present Situation of Text Clustering Algorithm

Fasheng Liu
Jiangxi University of Science and Technology
Ganzhou Jiangxi China
fashengliu@hotmail.com

Lu Xiong
Jiangxi University of Science and Technology
Ganzhou Jiangxi China
luzi0822@163.com

***Abstract*—With the popularity of Internet and large-scale improvement in the level of enterprise information, the explosive growth of resources, the research of text mining, information filtering and information search appear the unprecedented prospect. So, the cluster technology is becoming the core of text information mining technologies. Clustering is an important form of data mining. This paper introduces common text clustering algorithms, analyses and compares some aspects of clustering algorithms which contains the applicable scope, the initial parameters, termination conditions and noise sensitivity. Algorithms contain hierarchical clustering, partitioned clustering, density-based algorithm and self-organizing maps algorithm.**

Keywords- text clustering algorithm, hierarchical clustering, k-means algorithm, cluster text

I. INTRODUCTION

Text clustering is the process which is carried on the organization or division to the text set under the condition of un-learning. Its basic idea is to divide the similar text into the same class. Using text clustering technology, you can find a large-scale of text set classification system, and provides a broad view for the text set. It is applied at information extraction and Web data mining. Clustering algorithm can be divided into the following categories: hierarchical clustering, partitioned clustering, density-based algorithm, self-organizing maps algorithm. At the same time, the text clustering problem has its particularity. On one hand, the text vector is a high-dimensional vector, usually thousands or even ten thousands; On the other hand, the text vector is usually sparse vector, so it is difficult for the choice of cluster center. As an unsupervised machine learning method, because of not need to train the process and manual label document at category in advance, clustering has certain flexibility and high automation handling ability. It is become a important mean which pays attention for more and more researchers.

II. TEXT FEATURES REPRESENTATION

Because the text information has a limited structure or no structure, meanwhile, the document content is used by humans of natural language, thus the fundamental question of text clustering is how to text content can be expressed as a mathematical analysis and processing of form. Professor Salton proposed Vector Space Model VSM (essential) is widely applied and is one of better effect method in recent years. Its main idea is which document space is seen as vector space that is composed by a group of orthogonal term Vector, and each document is represented as a normalized feature vector $V(d) = (t_1, w_1(d); ...t_i, w_i(d); ...t_n, w_n(d);)$, Which t_i is entry items, $w_i(d)$ is t_i in d in weight, thus, article is expressed as a vector in high dimensional space.

Fig 1 gives a kind of data division which is generated by a decision tree algorithm. The initial clustering result contains 3

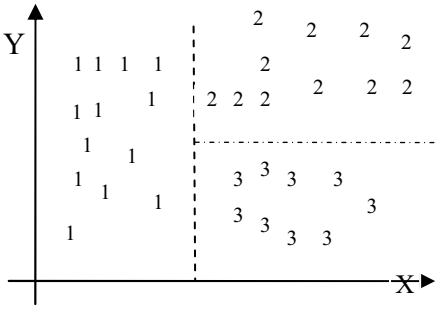

Fig.1. the description of clustering rule

clusters. The data points in cluster 1 represent by 1, the data points in cluster 2 represent by 2, the data points in cluster 3 represent by 3.

III. THE COMPARISON STANDARD OF TEXT CLUSTERING ALGORITHM

The purpose of text clustering is a large-scale text data sets which can be grouped into several categories, and made between the text information in the same class which has high similarity, rather than the difference of text between the different types. It is easy for people using text information. Therefore, comparison of the text clustering algorithm should be based on the following six criteria:

(1)It has a higher scalability. Clustering algorithm is not only used in the sample data sets, but also a large scale of the reality text data sets which should have a good effect.

(2)It can handle high dimensional data. The text data sets expressed by VSM typically have thousands or even web dimensional, so algorithm for text clustering is able to handle high dimensional data.

(3)It can discover the arbitrary shape of clusters. Since the development of cross disciplinary and comprehensive disciplinary, the boundaries between classes become increasingly blurred, the shape of class is not limited to spherical or other convex, it requires the text clustering algorithm can find the arbitrary shape of class.

(4)The dependence of input parameters and domain knowledge is low. Many algorithms need to give some parameters first, but in the absence of prior knowledge, these

The work is supported by the Science and Technology Research Fund of Jiangxi Provincial Education Department of China (No.GJJ08283, No GJJ11463)

978-1-4244-9698-3/11/$26.00 ©2011 IEEE

parameters are difficult to determine, and the clustering results are very sensitive to these parameters, so try to avoid it.

(5)The input sequence of data is not sensitive. The text expressed by VSM using the glossary as a feature item unit, and using glossarial word frequency value as feature item value, so that the input sequence of the text data has no effect on the results of the final clustering.

(6)It has good ability to deal with noise data. The vast majority of databases contain isolated points, the unknown data and so on, if the algorithm is sensitive to such data, it will reduce the quality of clustering results.

IV. COMMON TEXT CLUSTERING

A. Hierarchical clustering

Text clustering is a typical problem of unsupervised machine learning. Hierarchical clustering algorithm by combining the appropriate similarity measure similarity such as cosine similarity, Dice coefficient, Jaccard similarity coefficient, has become the mainstream technology on the document clustering. Hierarchical clustering is commonly text clustering method, which can generate hierarchical nested class. Hierarchical clustering method takes category as hierarchical, in other words, with the change of category hierarchical, object also corresponding change.

The result of hierarchical clustering forms a single category tree. Each class node contains several child nodes, brother node is division of its parent nodes (Fig.2). The bottom of the tree has 5 clusters, in the last layer ,cluster 2 contain data point 5 and data point 6, cluster 4 contain data point 8 and data point 9. With the bottom-up tree traversal, the number of clusters is less and less.

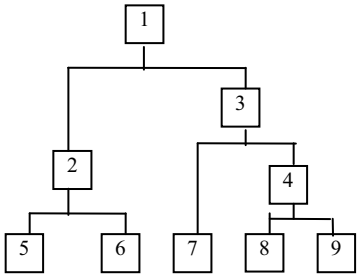

Fig.2. a sample of hierarchical clustering

Therefore, this method allows classifing data at different granularity. In accordance with generation methods of the category tree, hierarchical clustering method can be divided into two categories, one kind is integration method (bottom-up method), and the other kind is to split methods (top-down method).Hierarchical clustering accuracy is relatively high, but when each class merges, it needs to compare all classes' similarity in the global and selecting the most similar of two classes, so it's relatively slow. The defect of hierarchical clustering is that once a step (merge or split) completed, it can not be revoked, so it can not correct the wrong decision. Hierarchical clustering methods are generally divided into bottom-up hierarchical clustering method and top-down hierarchical clustering method.

- Bottom-up hierarchical clustering method

Bottom-up (merge) hierarchical clustering method starts from a single object, first takes an object as a separate category, and then repeatedly merges two or more appropriate categories. Until meeting stop conditions, hierarchical clustering doesn't loop (usually the number of parameters is K).Bottom-up hierarchical clustering process can be seen as a process of constructing the tree, which contains the class hierarchy information, and the similarity among all the classes. For a given document sets $D = \{d_1, d_2, \ldots d_n\}$,the specific process is as following hierarchical clustering:

(1)Each document d_i in Dwill be considered as a single member of the classes $C_i = \{d_i\}$, these classes form a cluster $C = \{C_1, C_2, \ldots C_n\}$.

(2)Calculate in C for each similarity between classes (C_i, C_j), denote by $sim(C_i, C_j)$.

(3)Select the largest class of similarity, C_i and C_j will be merged into a new class C_k, so as to constitute a cluster of set $C = \{C_1, C_2, \ldots C_{n-1}\}$.

(4)Repeat the above steps until C has only one class.

Hierarchical clustering has the following advantages: applied to any shape; applied to any form of similarity or distance form; the inherent flexibility of clustering granularity. And the disadvantage is which termination condition is imprecise ,and it needs to determine the human experience; once the clustering result form, generally no longer rebuild to improve the result , the error categories produced by error don't have ability of repair.

- top-down hierarchical clustering method

Top-down (splitting) hierarchical clustering method starts from the object's complete works, and gradually be divided into more categories. The typical approach is to construct a minimum spanning tree on similar graphes, and then at each step choosing a side which in the smallest similarity of the spanning tree (or in the farthest distance of the spanning tree) and removing it. If it deletes one side, it can create a new category. When the smallest similarity achieves some threshold value the cluster may stop. In general, the amount of computation of top-down method is greater than the bottom-up method, and the applications of top-down method is inferior widespread than the latter.

When using top-down (splitting) hierarchical clustering method, at first all objects are one kind, denoted by $C_1^{(0)} = C$. Wishing to be divided into k classes, there are two methods: one is the optimal partitioning, so take $C_1^{(0)}$ once divided into k classes; Another way is divided into two, so take $C_1^{(0)}$ divide into two categories: $C_1^{(1)}$ and $C_2^{(1)}$ .Then $C_1^{(1)}$ and $C_2^{(1)}$ respectively divided into two categories, it goes on until the class is divided into k up.

Bottom-up and top-down hierarchical clustering method have the advantage of simple, can flexibility deal with the multi-granularity clustering problem, can use various forms of similarity measure or distance metric, can handle a variety of property types. The disadvantage is difficult to determine the termination conditions of the algorithm, and choose to merge or split the points. And such decision and very critical, once a

group of objects will be merged or splited, next step is carrying on the newly generated cluster, process which has done can not be revoked, between the objects clustering can not be exchanged between the objects. If decides improper in some step merge or spilt, it possibly will lead to the cluster result quality which is low, and the expandable of such clustering method will be unsatisfactory.

B. Partitioned clustering

The task of partitioned clustering is which the data set is divided into k disjoint point sets, so that each sub-set point as far as possible homogeneity. Homogeneity is implemented as follows: select the appropriate score function and make the distance of every point and centroid which it belongs to each cluster minimize. The key of based-on division clustering is evaluate function, other parts of the algorithm is not much different from the general algorithm.

Partitioned clustering is applicable for carring on the cluster to the small scale's database to discover the spherical bunch (each cluster class regards as one cluster). Typical methods include classification based on k-means, PAM, CLARA and so on.

- k-means algorithm

The goal of k-means algorithm is based on the input parameters k, the data set is divided into k clusters. Algorithm use iterative update Method, in each round, based on k point of reference points were grouped around k clusters, each cluster centroid will be used as a reference point next round of iteration. Iteration makes the selected reference point closer to the true cluster centroid, so the clustering effect is getting better. [4]

k-means algorithm is as follows:

Supposes the set of the data point D is $\{x_1, x_2, \ldots x_n\}$, $x_i = (x_{i1}, x_{i2}, \ldots x_{ir})$ is vector in real space $X \subseteq \Re$, and r expresses by the attribute number of the data(the dimension of the data space).

Algorithm k-means (k, D)

(1)Choose k data points as the initial centroid (cluster center).

(2)Repeat for each data point when $x \in D$.

(3)Compute the distance from x to each centroid and assign x to the closest centriod.

(4)Re-compute the centroid using the current cluster memberships until the stopping criterion is met.

k-means method has the advantage as follows: has good geometry and statistical significance in the numeric attribute; be less sensitive to the order; has good effect on the convex cluster; can run in parallel; can carry on the cluster under the random norm. But its disadvantage is: need the user to give the number of cluster in advance; be unable to processes categorical attribute data; be sensitive to the isolated point; Cannot discover the non-sphere cluster or clusters which the difference of size is greatly; fall into local optimal solution frequently, but is unable to obtain the globally optimal solution; be vulnerable to the interference of abnormal points; be lack of scalability; clustering results are sometimes unbalanced.

- k-medoid algorithm

The process of k-medoid algorithm and k-means algorithm is similar, the only difference is that k-medoid algorithm uses a object closet the center of the cluster to represent the cluster, but k-means algorithm uses centroid to represent the cluster. The k-means algorithm is sensitive to the noise and the isolated point data, because a great value confront a profound influence on the calculation of centroid. The k-medoid algorithm by using the center point instead of centroid can effectively eliminate such effects.

k-medoid algorithm process is as follows: First, randomly select k objects as initial representative points of k clusters, according to the distance of remaining objects and the representative point object, remaining objects are assigned to the nearest cluster. Then, the repeatedly use of non-representative point take the place of the representative point, check whether the quality of clustering is improved. If so, then reserve this replacement; else abandon it, repeat the above of process until the situation not changed. The quality of clustering can be used to estimate a cost function, the function measure average dissimilarity between objects and representative points.

C. Density-based algorithm

Density-based algorithm suggest that, category is connected domain to expand in any direction by the same density. Therefore, density-based algorithm can discover the arbitrary shape, simultaneously the algorithm has the natural resisting effect on the noise. This algorithm is mainly to consider the density and connectivity and boundary areas of data space. For non-convex, irregular shape, k-means algorithm is often difficult to processing, however, density-based algorithm is well able to deal with such problems. According to different methods density-based clustering methods can be divided into connectivity density-based methods and based on density distribution function methods. The most typical algorithm of the former is density-Based Spatial Clustering of Application with Noise, it is a density clustering algorithm based on the high-density connectivity area; The most typical algorithm of the latter is density-based clustering.

- DENCLUE algorithm

The main idea of DENCLUE algorithm is certain influence relationship which can be described as a mathematical function between each data point and other adjacent data points. So a mathematical function is called influence function. It makes a data point quantify a value in their areas. The density of an object at the data space is modeled as the influence function of all data objects at the object space. It determines the density attraction points to divide and merge for the cluster and form the clustering results. The density attraction point is the maximum local point of the global density function. With other algorithms, denclue algorithm has the following advantages:

have a solid mathematical foundation, summary some clustering algorithm including partitioned clustering, hierarchical clustering etc; have a good clustering result for a large number of noisy data in data sets; algorithm provides a simple mathematical description for high-dimensional data set clusters of arbitrary shape; based on unit organization data make the algorithm to efficiently handle with a large high dimensional data.

D. -Organizing Maps algorithm

As a high-dimensional of clustering and visualization unsupervised learning algorithm, self-organizing maps algorithm is simulates the characteristic of the human brain to the signal processing developed an artificial neural network. This model proposed by Finnish Helsinki professor Teuvo Kohonen in 1981, now it is become applies widely method of self-organizing neural network. Neural network will be described as a prototype vector for each cluster, as the cluster prototype, prototype vector do not necessarily correspond to the data instance and objects specific. According to some distance measure, the new object is assigned gives the bunch which this object most similar prototype vector represents.

SOM algorithm can be described as follows[5]:

(1) Random initial connection weights, set the maximum training times for the K, the training counter k = 0.

(2)Randomly selected input mode, calculate the Euclidean distance of all input unit.

(3)Select Get node.

(4)The connection weights of winning node and its domain node makes the adjustment.

Performance	hierarchical clustering	partitioned clustering	density-based algorithm	self-organizing maps algorithm
attribute value	no requirement	numeric atteribute	numeric atteribute	numeric atteribute
shape	arbitrary	convex	arbitrary	?
measure	any	distance of normal space	density function	euclidean distance
granularity	flexible	K and initial point	threshold	Parameters
results optimization	No optimization	rebuild an optimization	optimization	optimization
initial condition	no	yes	yes	yes
termination condition	Not precise	precise	precise	precise
adapt to dynamic data	no	yes	yes	yes
noise	No influence	influence	not much influence	?

Table 1 Performance Comparison of Clustering Algorithms

(5)Counter is incremented, if k<K run in the step 2,else end train.

(6) Output result.

V. CONCLUSION

As an important form of knowledge discovery, clustering received more and more attention. This paper analyses the existing common clustering algorithm and summarizes the characteristics of each algorithm. Now with natural language understanding and research, Clustering is not only widespread applied to the natural language understanding, but also the application of natural language understanding has been improved and the corresponding increase. So it generates some new target-oriented clustering algorithm, such as based on ontology clustering algorithm, based on semantic clustering algorithm. With the development of network technology, on-line information processing in particular the treatment of various documents will become increasingly important; clustering technology will play an important role; new, highly efficient and targeted clustering algorithms will be developed and applied.

REFERENCES

[1] Salton G, Wong A, Yang C.A vector space model for automatic indexing[J] .Communications of the ACM, 1975, 18(11) : 613- 620.

[2] Zhao Y and Karypis G. Hierarchical Clustering Algorithms for Document Datasets[A]. *Data Mining and Knowledge Discovery* [C].10(2),pp.141-168,2005

[3] Karypis G, Zhao Y. Evalution of hierarchical clustering algorithms for document datasets[A]. In: *Proc of the International Conference on Information and knowledge Management*[C]. New York,2002. 512-524

[4] Bing. Liu Web Data Mining[M] Tsinghua University Press,2009

[5] T. Kohonen. Self-Organizing Maps. *Series in Information Sciences*, 30, Springer, Heidelberg, Second Edition. 1995

[6] Kim H J, Lee S G. A semi-supersvised document clustering technique for information and organization[A].In: *Proc of the Ninth International Conference on Information and Knowledge Management*[C].McLean, Virginia, 2002.159-168

[7] Wache H, Vogele T, Visser U, Ontology-Based Integration of Information-A Survey of Existing Approaches[C]. In: Proc of the IJCAE01 Workshop: Ontologies and Information Sharing. Seattle, WA, 2001:108-117

[8] Leuski A.Evaluating document clustering for interactive information retrieval[D].Massachusetts:University of Massachusetts, 2006.

[9] Xiaojun Wang, Jianwu Yang, Xiaoou Chen. an Improved K-means Document Clustering Algorithm[J] Computer Engineering, 2003, 29(2): 102-104

[10] Hongbin Gao, Haizhen Yang, Xiaobin Zhang. an Improved Document Clustering Algorithm[J] Computer Applications, 2008,27(9):30-32