Improved K-medoids Clustering Algorithm under Semantic Web

Ji Wentian, Guo Qingju
Department of Software Engineering
Hainan College of Software Technology
Qionghai, China
13976398691@163.com

Zhong Sheng*
College of Information Science & Technology
Hainan University
Haikou, China
shzhong@hainu.edu.cn

Zhou En
Department of Software Engineering
Hainan College of Software Technology
Qionghai, China
skywarps@163.com

***Abstract*—K-medoids clustering algorithm is highly efficient in classifying cluster categories. Based on algorithm analysis and selection improvement of centre point K, this paper sets up a web model of ontology data set object. It tries to demonstrate through experiment evaluation that the improved algorithm can greatly enhance the accuracy of clustering results under semantic web.**

Keywords-component; Ontology; Semantic Web; K-mediods algorithm

I. ONTOLOGY AND SEMANTIC WEB

Ontology is a philosophical concept which refers to the abstract nature of objective reality. Its definition in computer industry was originated in 1991 and from then on it becomes more clear and specific from vague and abstract generalization. The definition given by Fensel in 2000 is: ontology is the formalized description of the sharing of some important concepts of a field. Although there are many different definitions of ontology, they have similar connotations for they all believe that ontology is the foundation and specification for different subjects of a certain field or sphere to exchange. The goal of ontology is to obtain knowledge from related fields, provide common understanding and identify the concepts and their properties which are commonly recognized in these fields. Moreover, it also tries to provide a clear definition of the relationships between these concepts from formalized mode of different levels. Ontology is a conceptual model frame which can describe related problems in certain fields and is an important means in solving field knowledge sharing and its applications. Currently, matured tools and descriptive language have already been developed for the establishment of ontology.

Semantic web which will be the upgrade of internet page structure is the evolvement of revolutionary internet technologies. It enables information inside internet page to have clear definition and semantic meaning. Besides, it also provides good environment for users of this semanteme to bring out better service and cooperation between agents (or machine) and human. As early as in 1998, Tim Berners-Lee proposed the concept of Semantic Web informally and at the same time he gave out the seven-layer structure. The realization of semantic web relies on three key technologies: XML, RDF and Ontology. Its architecture follows the seven layer structure proposed by Tim Berners-Lee.

II. K-MEDOIDS ALGORITHM

Clustering is a data partition method and requires the maximum similarity of data under the same group and the maximum dissimilarity of data belonging to different groups. Its wide applications can be found in almost every field. There are mainly five kinds of clusters, namely, partition-based clustering, density -based clustering, model-based clustering, layer-based clustering and grid-based clustering, each of which includes many specific algorithms with different advantages and disadvantages. K-medoids algorithm studied by this paper is partition-based clustering algorithm which is characterized by wide usage, high efficiency and easy realization. K-medoids algorithm i.e. center point k algorithm is mainly intended to overcome the shortcomings of K-means algorithm, especially the sensibility of outliers (also called noise point).

A. Algorithm Principles

In K-means algorithm, the outliers in the centre point of a cluster will disrupt the computing process of the centre point of a class cluster, making the resulting centre point deviate from actual centre point. As a result of cluster “deformation” and deviation many samples which don’t belong to class cluster CI will be included in to make the clustering results incorrect. In K-medoids algorithm, first select k clustering centre points randomly from n data objects before computing the distance of other data objects to each clustering centre, then choose the one which is closest to clustering centre to set up an initial partition, and then use the iteration methods to change the clustering

Published by Atlantis Press, Paris, France.

centre continuously until the most suitable fixed partition is found.

In K-Medoids algorithm, the standard of improving clustering quality to make it more compact is used in selecting particles from clustering samples after each iteration. This algorithm adopts absolute error standard in defining the compactness of a class cluster.

$$E = \sum_{1}^{K} \sum_{p \in c_j} |P - o_j| \quad (1)$$

P is the sample point in space While o_j is the particle of class cluster c_j.

K-medoids algorithm considers that if the selection of some sample points as centre point can reduce the absolute error of original particle then we can use the sample point to replace original particle and select the sample point with minimum absolute error as the new centre point when recomputing the center point of class cluster iteratively.

Sample point $A \rightarrow E_1 = 10$

Sample point $B \rightarrow E_2 = 11$

Sample point $C \rightarrow E_3 = 12$

Original particle $D \rightarrow E_4 = 13$,

we select A as the new center point of class cluster.

K-medoids also adopts Euclidean distance to decide to which class cluster a sample point belongs. The terminal condition is that all particle center points change no more, which is considered to be the end of clustering.

B. *Algorithm Characteristic*

The advantage of this algorithm lies in the improvement of the "noise" sensibility of K-means, the overcoming of the selection randomness of clustering centers and the realization of global optimization instead of local optimization. Its disadvantages are the same with K-means. Besides, due to the adoption of new particle computing rules, algorithm time complexity increases also: $O(k(n-k)^2)$.

III. IMPROVED K-MEDOIDS ALGORITHM

K-medoids has high accuracy of pattern matching and is not sensitive to dirty and abnormal data. Besides, it requires enormous amount of iteration and a lot of time, reducing the efficiency of clustering; therefore, it can only be used to handle small amount of data. With the development of information and internet, data on web database increases rapidly not only in amount but also in complexity. In view of this, we should not only pay attention to the efficiency of information acquisition, but also to the accuracy of it. Under this condition, we should establish an improved K-medoids clustering algorithm which can meet the needs of complex web data set (such as ontology-base data set).

A. *The strategy of K-medoids algorithm*

In K-medoids algorithm, the selection of clustering centre will directly decide the accuracy and efficiency of clustering results; therefore, the improvement of K-medoids algorithm centers on the selection of k clustering centre point. The algorithm flow is described as follows:

1) add k pieces of data which differ from each other greatly but are very similar to the data in the data set into the existing data.

2) use these k pieces of data as the centre point of clustering algorithm.

3) divide all other data into various clusters, and compute the variance between data in each cluster and centre point. Replace them with the minimum value to get a new centre point.

4) repeat the procedure of 3) until the cluster centre changes no more to get clustering results before the algorithm ends.

B. *The analysis of improved K-medoids algorithm*

K-Medoids is a partition clustering algorithm which needs to select k clustering centers from data objects and establish an initial partition nearest to clustering centre for other data before iterating and moving clustering centers continuously until an optimum partition is reached. Due to the randomness of K value and initial selection of clustering centers, the efficiency and accuracy of it is very low. The improved K-medoids algorithm which adds k value under constraint conditions as clustering centers and only needs one iteration to get clustering results can not only solve the randomness of clustering center selection but also can improve its efficiency toward complicated data to achieve global optimization.

IV. EXPERIMENT RESULTS AND EVALUATION

A. *Experiment Results*

The data objects this paper studies come from Movielens database which is a data set of Movielens. RDF documents, the so-called "semanteme". It is not a language but a model to express web data. The basic data model of RDF includes three kinds of objects: resource which refers to data on websites, property which is used to describe the characteristics and relationship of resources and statement which is used to indicate a certain property or property value. In experiment, a web semantic model is first established and semantic web text mining adopted to define audience's age, sex, interested films and their times of movie watching during a fixed period as a n dimension vector $X = f * t$ according to user demand. Audience is treated as vector $f = (f_1, f_2, \ldots, f_n)$ while each audience's characteristic t is treated as the corresponding component. The relationship between film characteristics and audience characteristics is defined as characteristic item with weights (the weighted average of rating in RDF data), enabling vector matrix to possess plentiful relationships and thus to provide reliable data for clustering analysis.

Through pretreatment of 500 pieces of RDF data, data are divided into six groups according to film types each of which has 100~200 pieces of markings given by users and are taken as the usage data set for improved ontology- based semantic web K-means algorithm.

B. *Experiment Evaluation*

The method which is widely used in statistical classification and clustering algorithm to evaluate result quality is adopted to assess the experiment. The rate of accuracy refers to precision ratio of the number of correct text clustering and the number of text in clustering results while recall rate refers to ratio of the number of correct text in clustering results and the number of all relevant texts in all texts. Table 1 is a comparison of the precision ratio and recall ratio of improved K-medoids algorithm and unimproved K-medoids algorithm

Table 1 Experiment Results

	K-Medoids algorithm		**Improved K-Medoids algorithm**	
Classification	**Rate of Accuracy**	**Recall**	**Rate of Accuracy**	**Recall**
Story film	76.43	**78.41**	85.20	**88.57**
Science fiction film	77.36	81.23	87.12	90.40
Comedy	80.63	82.20	91.31	90.34
Cartoon	72.23	70.65	83.23	82.15
War Movie	78.20	83.12	88.30	91.33
Suspense	70.45	72.20	82.21	84.12

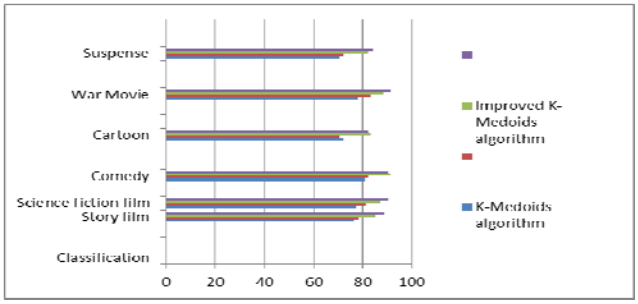

Figure 1 Experiment Results

From Table 1 and Figure 1 we can see the precision ratio and recall ratio of the improved K-medoids algorithm in story film, science fiction film, and comedy and war film are all higher than that of traditional K-medoids algorithm. The reason is that data selection for these four types is more concentrated while the selection of other objects is looser resulting in slightly lower clustering effect. From picture 4-1 we can see both precision ratio and recall ratio of the improved algorithm have been enhanced, which corresponds with the theory result.

C. *Problems Existed*

This paper proposed the method of adding k value under constraint conditions as clustering centers to improve clustering effect, but didn't go deep into the research of k value, which leads to restricted application of this algorithm. Moreover, its stability toward large data objects still needs inspection.

V. CONCLUSION

This paper first described partition clustering algorithm K-medoids and analyzed its characteristics and deficiencies under large data environment. It then proposed methods to improve algorithm and carried out experiment verification by applying the improved algorithm to ontology data. Through experiment and evaluation criteria of precision ratio and recall ratio, it demonstrated that precision ratio and efficiency of the improved algorithm rise up considerably.

ACKNOWLEDGMENT

Projects: Hainan province natural science fund, the numbers: 610228 and 611121.

Corresponding author: Zhong Sheng:

REFERENCES

[1] Li ping Jing. An entropy weighting K-Means algorithm for subspace clustering of high-dimensional sparse data[J].IEEE Transactions on Knowledge and Data Engineering,2007,19(8):1026-1041.

[2] Huang S, Chen Z. Multi-type features co-selection for Web document clustering[J].IEEE Transactions on Knowledge and Data Engineering,2006,18(4):448-458.

[3] Wu X D,Kumar V,Quinlan J Retal. Top 10 algorithms in data mining[J].Knowledge and In-formation Systems,2008,14(1):1-37.

[4] Basu S. Semi-supervised Clustering Probabilistic Models, Algorithms and Experiments[D].USA: the Faculty of the Graduate School of The University of Texas at Austin,2005.

[5] SU MC,CHOUCH.A modified version of the K-Means algorithm with a distance based on cluster symmetry[J].IEEE Trans on Pattern Analysis and Machine Intelligence,2001,23(6):670-690.

[6] GRUBER T. Toward principles for the design of ontologies used for knowledge sharing[J].International Journal of Human Computer Studies,1995,43(5):907-28.

[7] Xu Yifeng Chen Chunming. ONTOLOGY-BASED WEB MINING CLASSIFICATION METHOD AND ITS APPLICATION [J]. Computer Applications and Software,2009,26(3):208-209.

[8] FU Xiao;LUO Bin;CHEN Shi-Fu. Research of Semantic Web Mining [J]. Computer Science,2005 Vol.32 NO.3.

[9] Stefan Decker. The Semantic Web : The Roles of XML and RDF. IEEE Internet Computing (2000) Volume:4, Issue:5, Publisher: IEEE, Pages: 63-74.