The Analysis of the Ontology-based K-Means Clustering Algorithm

Guo Qingju Ji Wentian
Department of information management
Hainan College of Software Technology
Qionghai, China
31690582@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*—In view of research findings made from home and abroad on clustering algorithm and the traditional partition clustering method K-means algorithm, this paper first analyses the advantages and disadvantages of this algorithm and then combines it with ontology- based data set to establish a semantic web model. It then tries to improve the existing clustering algorithm in various constraint conditions to demonstrate that the improved algorithm has better efficiency and accuracy under semantic web.**

Keywords-component; Ontology; Semantic Web; Data; K-means Algorithm

I. INTRODUCTION

With the rapid development of internet, web has become a primary tool for people to obtain information. The invention of semantic web, a kind of descriptive language which can realize public framework and extend web functions, makes it possible to share vast data in web cyberspace maximally. In the processes of establishing and realizing semantic web, ontology-based semantic web clustering algorithm is a hot issue worth researching for it enables machine to have the capacity of understanding and cooperating both with people and other machines. It is a strong technology support for the exploration of unknown world and thus is of great realistic significance.

In recent years, a lot of researches have been made on the research findings of clustering algorithm. In view of the traditional clustering classifying method of K-means algorithm, this paper first analyses the advantages and disadvantages of this algorithm and then combines it with ontology- based data set to establish a semantic web model. It then tries to improve the existing clustering algorithm in various constraint conditions to demonstrate that the improved algorithm has better efficiency and correctness under semantic web.

II. TRADITIONAL K-MEANS ALGORITHM

Clustering has first been widely used in fields like math, statistics, biology and economics. In recent years, its application has been further extended to computer studies. The process of classifying a group of physical objects or abstract objects into similar object classes is called clustering analysis, in which a cluster is a set of data objects. All objects under the same cluster are similar to each other, but differ from each other if they belong to different clusters. Clustering can not only be applied to data segmentation but also to outlier detection. Clustering analysis which is the major task of data mining can be used as a pretreatment for other algorithms (such as classification algorithm and qualitative inductive algorithm).

A. Algorithm Principles

The traditional clustering classification method K-means algorithm proposed by Mac Queen in 1967 is an algorithm technology of relative large influence among clustering algorithms so far. K-means algorithm is characterized by fastness of clustering and easiness of realization. It is a typical distance-based clustering algorithm, adopting distance as the evaluating indicator of i similarity, i.e. the bigger the distance between two objects, the more similar they are. The final goal is to obtain a cluster which is tight and independent.

The principle of K-means is first to select k number points as the original clustering center randomly ,then calculate the distance between various samples and clustering center K and then adjust the samples. After that, calculate the average value of data objects in the newly formed cluster to get a new K value. If there are no differences to K value for two consecutive times, the adjustment and clustering algorithm come to an end.

There are many versions of realization methods for this algorithm, so here we will not repeat.

Published by Atlantis Press, Paris, France.

B. The advantages and disadvantages of this algorithm

The advantages of this algorithm include the examination of the correctness of the classification of all samples in all iterations. If there is incorrectness, adjustment is required. The next iteration will take place only after the adjustment of all samples and the modification of clustering centers. If all samples have been classified correctly, then the clustering center will remain the same and no adjustment is needed, which indicates the end of the algorithm. Another advantage of this algorithm is that it can handle large scale data set to produce tight resulting clusters separating from each other obviously, thus it is of strong flexibility and high efficiency. The complexity of this algorithm is $O(nkt)$, in which n refers to the number of data objects while t refers to the number of iteration.

The disadvantages of this algorithm include a preset K value and centre points which will often have a big influence on clustering algorithm. Besides, this algorithm can only be applied to data clustering of numeric type and usually ends with a partially optimum value. It is very sensitive to “noise” and the outlier data, some of which will affect the results of the whole average value, causing a great departure of average value.

III. IMPROVED K-MEANS ALGORITHM

Generally K-means algorithm must have a preset k value and a classification of data set which are hard to realize in practical situation, resulting in a big departure of clustering outcome. It is the major disadvantage and shortcoming of this algorithm.

As to the determination of k value, a great many algorithms have been proposed with the purpose of getting a better result.

A. Ideas of improving K-means algorithm

1) The paper selects the value of k with K-means initial distribution selection method based on the principle of limit value, and uses the resulting k value as input of K-means algorithm before text clustering.

Similarity matrix is first constructed before the analysis of it so as to start the selection of the original clustering point and identification of clustering value k automatically. It is proved by experiments that the so obtained k value is closer to true value and needs less human interference.

2) The improved algorithm computes cosine similarity Sim between each data object and clustering center.

3) Due to the random selection of original centre point, the clustering results of traditional K-means algorithm changes every time and thus is very unstable. In view of this, improved algorithm adds the following three conditions of convergence.

a) whether clustering center $Z_j(I)$ *becomes unchanged;*

b) whether category maximum average similarity $|sim(I) = sim(I-1)|$ *changes*

c) whether the maximum iteration is reached.

B. The establishment of model and realization of algorithm

1) The establishment of semantic web data model

The data objects this paper studies come from Movielens database which is a noncommercial practical website for research. Launched by GroupLens project team from Computer Science College and Engineering College of Minnesota University of America, this website recommends movies to its users. Possessing 1000000 marking data given by 6040 users regarding 3900 films, Movielens data set is widely used as experiment material. It is expressed in RFD form, the so-called “semanteme” which 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; therefore, a RDF description is actually a triplet:

(object[resource] ， attribute[property] ， value[resource or litera1])

For example, one piece of data from the ontology data set can be expressed in the following RDF form:

```
<movie:Movie rdf:about="http://imdb.com/title/tt0116790
">
<review:hasReview>
    <review:Review rdf:ID="R196-242">
    <rdfs:comment>A review of MovieLens item 242 - -
by user 196 at time 881250949
    </rdfs:comment>
    <review:reviewer rdf:resource="#196" />
    <review:rating>3</review:rating>
    <review:maxRating>5</review:maxRating>
    <review:minRating>1</review:minRating>
    <dc:date>881250949</dc:date>
</review:Review>
</review:hasReview>
<rdfs:label>242</rdfs:label>
<rdfs:comment>MovieLens item 242 - </rdfs:comment>
</movie:Movie>
```

During this experiment, the key procedure is the establishment of a web semantic model on Movielens database and the use of semantic web text mining. According to user demand, audience’s age, sex, interested films and their times of movie watching during a fixed period is defined as a n dimension vector $X = f * t$, in which, audience is treated as a vector $f = (f_1, f_2, ..., f_n)$ 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.

2) Pretreatment of data and the realization of algorithm

According to needs, audience’s age, interested films, their times of movie watching during a fixed period and their ratings toward films is defined as a n dimension vector matrix $X = f * t$ (i.e. the audience-characteristic matrix). Use singular value decomposition SVD to break down X^T into $X^T = T_0 S_0 D_0^T = (D_0 S_0^T D_0^T)^T$

and X into $X = (X^T)^T = (T_0 S_0 D_0^T)^T = ((D_0 S_0^T D_0^T)^T)^T = T_0 S_0^T D$

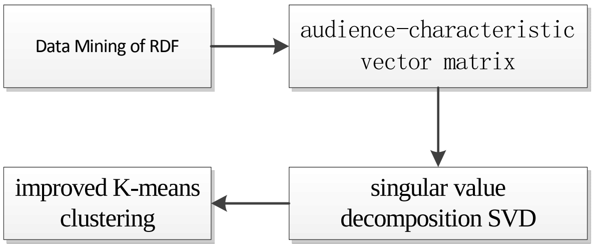
Data Mining of RDF
audience-characteristic vector matrix
singular value decomposition SVD
improved K-means clustering

Figure 1 The clustering process of improved K-means algorithm on ontology-based semantic web.

After dimensionality reduction of vector space we will get a new matrix $D_0 S_0$. Then adopt cluster algorithm to get the similarity of sim of audience, and then carry out clustering by using the improved K-means algorithm. Figure 1 shows the clustering process of this algorithm.

IV. EXPERIMENT ANALYSIS AND EVALUATION

A. *The generation of vector matrix files*

Through pretreatment of 500 pieces of RDF data, the age of audience, their sex, film preferences and film watching times during a fixed period is defined as n dimensional vector $X = f * t$, in which, audience is treated as vector f while the characteristics of each audience (suppose it to be 20)is treated as a corresponding component of it. In this way, we can get a 500×20 vector matrix.

In experiments, data from six types of films each of which has 100~200 pieces of markings given by users is taken as the usage data set of improved ontology- based semantic web K-means algorithm.

B. *Experiment Evaluation*

Evaluating results is the final procedure of clustering. In this paper, precision evaluation method is adopted although there exist many evaluation methods currently and the term precision is defined as:

$$precision\ (r, s) = n(r, s) / n_r$$

In which, $n(r, s)$ refers to cluster r after clustering and the number of users of predefined category s while nr is the number of users of cluster r.

In experiment, the original algorithm is first adopted and is classified according to users' age. The data precision obtained through unimproved K-means algorithm and improved K-means algorithm is as follow:

Table 1 The comparison of data precision between unimproved K-means algorithm and improved K-means algorithm

	Clustering Time （S）	**dimensionality reduction decomposition time （S）**	**Error number of users**	**Precison Rate**
Unimproved Algorithm	29	8.34	65/200	67.5%
Improved Algorithm	23	8.12	34/200	83%

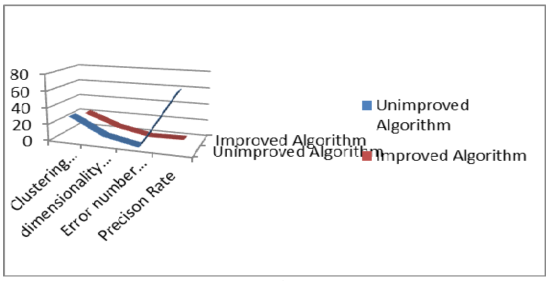

Figure 2 The precision distribution of unimproved and improved K-means algorithm

Table 1 and Figure 2 shows that the semantic web model and improved K-means algorithm can greatly enhance the efficiency and precision of clustering. In the execution process of algorithm, the limit value principle is used to get the k value and then semantic analysis of user characteristic relevance is then carried out to obtain a vector matrix. Finally, singular value dimensionality reduction and decomposition are used in the improved algorithm to get a more stable and correct clustering results. From this experiment, we can see the stability and authenticity of ontology data after clustering. But at the same time, we should notice that with the increase of data, eigenvector matrix expands rapidly to result in more operand, which corresponds completely with the conclusion of realization theory.

V. CONCLUSION

This paper proposes a "limit value"- based principle to select the starting point of text clustering and an improve K-means clustering algorithm through singular value dimensionality reduction decomposition. Experiments show that this method can improve stability and accuracy of clustering results and at the same can overcome the shortcomings of traditional manual selection of starting point. Through the comparison of clustering results of ontology data set Movielens obtained by unimproved K-means algorithm and improved K-means algorithm and the precision evaluation, we can see that the clustering quality of improved semantic web -based K-means clustering algorithm is obviously better than that of the original algorithm, which can provide a reliable analysis basis for the follow-up data application.

ACKNOWLEDGMENT

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

Corresponding author: Zhong Sheng

REFERENCES

[1] 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.

[2] 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.

[3] Spragins J． Learning without a teacher[J]． IEEE Transactions of Information Theory,2005,23(6)：223-230．

[4] FENSEL D,LASSILA O,VAN HARMELEN. The semantic Web and its languages[J].IEEE INTELLIGENT SYSTEMS AND THEIR APPLICATIONS,2000,15(67-73).

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

[6] JI Wen-tian .Fuzzy Cluster Analysis of Large Data under Semantic Web[J] Computer Knowledge and Technology,2011(27).

[7] 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.

[8] DOMINGUE J. Tadzebao AND Web Onto: Discussing, Browsing, AND Editing Ontologies[M].1998.

[9] Bamshad MobasherI，Honghua Dai。Tao Luo et al． Integrating Web Usage and Content Mining for More Effective Personalization [C]． In：Proc of the l st Int． Conf． on Electronic Commerce and Web Technologies． 2000：165-176.