Application of Shuffled Frog-Leaping Algorithm in Web's Text Cluster Technology

Yun Fang[1] and Jianxing Yu[2]

[1] Computer Science and Technology College
Taiyuan University of Technology
Taiyuan, Shanxi, China
fyun63@126.com
[2] Computer Science and Technology college
Taiyuan University of Technology
Taiyuan, Shanxi, China
yuchi0529@qq.com

Abstract. With the rapid development of Internet, more and more massive information, search engine technology developed rapidly, but the search engine's search results don't not meet the search requirements, The k-means clustering algorithm are introduced to gather web documents class, in order to improve the clustering performance, the introduction of leapfrog algorithm selection of k value aiming to improve the accuracy of search results and to increase the search engine returns results associated with the query topic.

Keywords: Clustering algorithm; Text clustering algorithm; Leapfrog algorithm.

1 Introduction

The rapid development of Internet makes the number of sites, more and more unstructured increasing web document information, Internet users to be able to collect valuable information, become the major research focus in the search engine technology, the proposed clustering technique is Find the exact search engine provides a very important role, allowing users to quickly find the needed information effectively has become an urgent problem. In this paper, we are clustering technology to the text on the handling of search engine results.

Clustering in data mining web document also plays an important role, document clustering can reveal the internal structure of the document collection, the discovery of new information, document clustering document collection is divided into several clusters, the cluster content of the document requested Similarity as large as possible, while the cluster similarity between the documents as small as possible.

In this paper, a document clustering algorithm[1] is widely used K-means based segmentation algorithm, K-means clustering algorithm is the clustering of a dynamic clustering method, K-means algorithm based on the predetermined value K, to be together Class samples are divided into K classes, so that all the samples cluster in the domain of the square of the distance to the cluster center and the smallest. However, the clustering algorithm selected by the number of cluster centers of K, especially for

G. Zhiguo et al. (Eds.): WISM 2011, CCIS 238, pp. 363–368, 2011.

the class number of the sample set, K value of the options to be specified and a random person, in order to obtain better clustering results, Usually test the different K values. K are chosen in view of the uncertainty, this K value selected by Leapfrog algorithm[2] to avoid the use of random numbers as the initial cluster centers resulting from the same K value of the clustering effect of instability.

2 Cluster

2.1 Text Clustering Ideas and Steps

The main idea of clustering[3] is through a specific algorithm, according to the similarity of text data and difference, the text is divided into several categories, in the same category as similar as possible differences in different categories. The main process: First, the clustering of the text to pre-processing, segmentation of words and the exclusion of irrelevant words, and the frequency of statistical terms, that each of the text processing, and generates the text of the feature vector space, Then for each particular feature vector space decomposition of its extraction, as far as possible to represent the text extracted feature vector, and calculate the similarity between the text and then choose the clustering model to test and evaluate the clustering quality, and finally The results displayed.

2.2 K-means Clustering Algorithm

Clustering algorithm for text clustering algorithm K-means algorithm is a clustering algorithm based on segmentation, segmentation-based clustering algorithm can be simply described as; to construct a set of objects form a partition of K clusters, makes the evaluation function is optimal. Because the number of K clusters will not change. The algorithm to K clusters start to the end of K clusters, each iteration, each instance of the original, if not retained in the cluster, the other is to be assigned to a cluster, this process is repeated until it meets a Stop criteria.

X_j (j = 1,2,3, ... N) is divided into K classes Gk (k = 1,2,3 .... k), all samples of each class form a group, find the center of each cluster, and Allocation vector A [1], A [2], ..., A [n], makes the non-similarity (or distance) index of the value function (or objective function) to a minimum.

The value is the set of cluster centers of all the mean vectors:

Here, this calculation is repeated basis, until the evaluation function below a threshold or 2 adjacent close iteration. This calculation is repeated basis, until the evaluation function below a threshold, or 2 adjacent. The difference is lower than a threshold, given the following K-means clustering algorithm pseudo code. The difference is lower than a threshold, given the following K-means clustering algorithm pseudo code.

K-means clustering algorithm K-means clustering algorithm

1: procedure KMeansCluster (X1, ..., XN, K)
2: A [1], A [2], ..., A [n] initial cluster⇓initial cluster assignment
3: repeat
4: Change false

5: For i = 1 to N do 5: For i = 1 to N do
6: function (xi, ck)
7: if A [i] is not equal k then
8: until change is equal to false return A [1], A [2], ..., A [n]
9: end procedure

K-means clustering algorithm is an important aspect of the choice of initial cluster centers, in addition to generating k-cluster, but also generates the center of each cluster, cluster performance and choose the initial cluster centers, and that most choose to be the first K sample set cluster samples as initial cluster centers. K-means clustering algorithm is an important aspect of the choice of initial cluster centers, in addition to generating k-cluster, but also generates the center of each cluster, cluster performance and choose the initial cluster centers, and that most choose to be the first K sample set cluster samples as initial cluster centers. but this selection is not scientific and increase a great deal of time and space overhead, the introduction of leapfrog algorithm for this iteration K, aims to improve the clustering performance. But this choice is not scientific and increase a great deal of time and space overhead, the introduction of leapfrog algorithm for this iteration K, aims to improve the clustering performance.

3 Experimental Result Analysis

Purpose of this object to web document clustering is a form of unstructured data, in order to apply clustering algorithm, the document must be expressed to form a structured data form. We use the vector space model (VSM) to represent the document in the form of VSM We use the vector space model (VSM) to represent the document in the form of VSM.

3.1 Data Preprocessing and Feature Extraction

Purpose of this object to web document clustering is a form of unstructured data, in order to apply clustering algorithm, the document must be expressed to form a structured data form. We use the vector space model (VSM) to represent the document in the form of VSM We use the vector space model (VSM) to represent the document in the form of VSM.

V (di) = (Ti1, Wi1; Ti2, Wi2; ..., Tin, Win) Where T for the entry, Wij to Tij in the document in the weights di, TF is term frequency, TF greater, indicating the term the greater weight in the article. Where Tij for the entry entry, TF is term frequency, TF greater, indicating the term the greater weight in the article. IDF is the document frequency, the classification of the document to be centralized, The document contains a number of S words. Where W (t, di) for the word t in the text di in weight, t (t, di) for the word t in the text di of word frequency, N is the text of the total number, ni is the training of the text focus appears t the text of the number of denominator is a normalization factor. Where W (t, di) for the word t in the text di in the weight, t (t, di) for the word t in the text di in the word frequency, N is the text of the total number, ni is the training of the text focus appears the number of text t, the denominator is a normalization factor. log (N / nt +0.01) actually represents the IDF. log (N / n t +0.01) actually represents the IDF.

XML documents can be expressed by the title, keyword, abstract and so on. XML documents can be expressed by the title, keyword, abstract and so on. So we modified the weight of the important position of the weight of larger vocabulary. To this end we correct weight, an important position to make the words bigger weight. We are introducing a random number that represents an important position in terms of the degree of value Where w (t, di) is the weight, w (t, di) to calculate the original TF-IDF weight, It for the term t in the document title, keyword, abstract, URL, and other important positions on the number of times. Extracted from the document that best represents the content of the document vocabulary, to improve the running efficiency and reduce the impact of noise words. Extracted from the document that best represents the content of the document vocabulary, to improve the running efficiency and reduce the impact of noise words.

Text similarity calculation is an important part of preprocessing on the text similarity measure, we use the Euclidean distance calculation method used to calculate the similarity between vectors, Where xi, and xj, respectively, the text refers to the characteristics of the two vectors, xi, and xj are respectively characteristic xik and xjk k-dimensional elements in the first, |xi| is the dimension of feature vector xi. Where xi, and xj denote the two texts are the eigenvectors, xi and xj are, respectively, characteristics of xik and xjk k-dimensional elements in the first, xi is the dimension of feature vector number.

3.2 Leapfrog Algorithm K-Means Clustering Algorithm

Conditional (or observed) variable belong to the category of evidence cliques, while the cliques which contain only target (or label) variables belong to the category of compatibility cliques. Obviously, in the example, the cliques in the flat conditional Markov networks indicated by the thin lines belong to the category of evidence cliques, while the cliques specified by the relational clique templates indicated by the bold lines belong to the category of compatibility cliques.

The initial position of cluster[4] centers will affect the performance of K-means algorithm, the initial location of cluster centers so the choice is based on iterative leapfrog algorithm selected K values, so that the location of the initial cluster centers to be scattered clustering of several groups of artificial selection of the document. The first is the formation of the initial population, followed by frog design groups: According to F frogs (solutions), randomly generated frog populations; each frog for a specific calculate the target function value; in descending order based on the objective function value F S frogs were divided into subgroups. for each sub-group of frogs are found in one of the best individual and the worst individual to identify the best individual groups; for each subgroup, in descending order according to the objective function value of the individual, re-distribution and mixing operations; termination conditions are met, the end of iteration, the optimal objective function value of the output information, or turn to the original sequence. The first is the formation of the initial population, followed by frog design groups: According to F frogs (solutions), randomly generated frog populations; each frog for a specific calculate the objective function value; descending order according to the objective function F frogs into S sub-groups according to; for each sub-group of frogs are found in one of the best individual and the worst individual to identify the best individual groups; for each

subgroup, in descending order according to the objective function value of the individual, re-distribution and mixing operations; termination conditions are met, the end of iteration, the optimal objective function value of the output information, or turn to the original sequence.

Leapfrog algorithm parameters: maximum step size Dmax = 20, the total number of frogs F = 2000, every subgroup of evolution algebra N = 1200, the number of m = 8 subgroup iteration Iter = 1000, the number of sub-group of the frog n = 250, for the parameter settings are currently no guiding principle, most of them are obtained by experimental test, K values between 1 and maxnum, leapfrog algorithm iteration the value of K must satisfy this condition. Leapfrog algorithm parameters: maximum step size Dmax = 20, the total number of frogs F = 2000, every subgroup of evolution algebra N = 1200, the number of subgroups m = 8 iterations Iter = 1000, sub-group n = number of frogs 250, set the parameters there is no guiding principles, most of them are obtained by experimental test, K values between 1 and maxnum, leapfrog algorithm iteration the value of K must satisfy this condition.

3.3 Leapfrog Algorithm K-Means Clustering Algorithm

The design of fitness function evaluation of clustering results is very important. A good clustering should be a large class of distance; class within the distance is small; the number of samples between classes evenly. A good clustering[5][6] should be a large class of distance; class within the distance is small; the number of samples between classes evenly. Between the samples which NumDifference that the number of different types of statistics, K-means algorithm performance depends on the value of the initial cluster centers selected good or bad, when the initial cluster centers were randomly selected, for the same value as the initial center of K different fitness function value will lead to the same parameters in different results. Between the samples which NumDifference that the number of different types of statistics, K-means algorithm performance depends on the value of the initial cluster centers selected good or bad, when the initial cluster centers were randomly selected, for the same value as the initial center of K different fitness function value will lead to the same parameters in different results. leapfrog algorithm is applied to the selection of K value is more reasonable to reduce the time and improve efficiency. Leapfrog algorithm is applied to the selection of K value is more reasonable to reduce the time and improve efficiency.

4 Conclusion

Experimental results show that adding leapfrog algorithm with the original K-means algorithm K-means algorithm is basically the same running speed, the average accuracy than in the original algorithm has generally improved, especially in the right when the specified number of clusters K, the average accuracy improved by nearly 29%, indicating that the K means clustering algorithm using leapfrog algorithm, can improve the clustering performance[7].

References

1. Jiang, M.F., Tseng, S.S.: Two phase Clustering process for outliersdetection. Pattern Recognition Letters 22, 691–700 (2001)
2. Rahimi-Vahed, A., Dangchi, M., Rafiei, H., Salimi, E.: A novel hybrid multi-objective shuffled frog-leaping algorithm for a bi-criteria permutation flow shop scheduling problem, May 5 (2008)
3. Hearst, M.A.: Clustering versus faceted categories for information exploration. Communications of the ACM 49(4), 59–61 (2006)
4. Jiang, Y.-Q., Zhang, Y., Zhou, Y.: K-means algorithm for optimizing the number of clusters based on particle swarm optimization. Control Theory and Applications 1175-1179 (October 2009)
5. Li, T.: Document clustering via Adaptive Subspace Iteration. In: Proceedings of the 12thth ACM Intrnational Conference on Multimedia (2006)
6. Makoto, I., Takenobu, T.: Hierarchical Bayesian clustering for automatic text classification
7. Bruce Croft, W., Metzler, D., Strohman, T.: Search Engines Information Retrieval in Practice. Pearson Education, London (2010)