The Document Clustering based on the Visual Features Using Visual Features Using Crossover and Mutation

[1]M.Ashok Kumar, M.Tech, [2]V D Phani kumar,[3]Y.Sandeep

[1]Assistant Professor,[2]M.Tech Student,[3]B.Tech Student

[1]Dept of Information Technology,V R Siddhartha Engineering College,Kanuru, Krishna (DT).

[2]Dept of Information Technology,V R Siddhartha Engineering College,Kanuru, Krishna (DT).

[3] Dept of Information Technology,V R Siddhartha Engineering College,Kanuru, Krishna (DT).

Abstract: There are two vital problems cost conducting research within the fields of customized data services based on user model. One is the way to get and describe user personal information, i.e. building user model, the other is the way to organize the data resources, i.e. document clustering. It is tough to search out the specified data while not a correct clustering algorithmic rule. In the recent years many ideas have been proposed, but most of the ideas are limited to the books and some other useful information is contributed for the clustering the document which referred as the visual features. In this paper we propose a method to cluster the scientific document based on the visual features which we refer to use the VF-Clustering algorithm. There are five types of the visual features for the document. They are defined, including body, title, keyword, subtitle and abstract. The thought of crossover and mutation in genetic algorithm is used to adjust the value of *k* and cluster center in the k-means algorithm dynamically. The clustering accuracy and steadiness of subtitle are only less than that of body but the efficiency is much better than body because the subtitle size is much less than body size. Combining the keyword and subtitles shows the better accuracy than the each of individual. If the efficiency is an essential factor, clustering by combining subtitle and keyword can be an optimal choice.

Index Terms: k-Means, Document clustering, genetic algorithm, visual features.

I. INTRODUCTION

Now-a-days personalized information services play an important role in people's life. The two important problems that costs the researching in the fields. They are

a. how to get and describe user personal information
b. how to organize the information resources

Personal information is described exactly only if user behavior and the resource what they look for or search have been accurately analyzed. Depending on completeness and accuracy the user model effectiveness of a personalized service. The basic operation is organizing the information resources. Indexing or searching millions of documents and retrieving the desired information has become an increasing challenge and opportunity with the rapid growth of scientific documents as there are millions of scientific documents available on the Web. Clustering plays an important role in analysis of user interests in user model, which it requires; high-quality scientific document clustering plays a more and more important role in the real word applications such as personalized service and recommendation systems [6][8][9] Scientific document clustering is a technique which puts related papers into a same group. The documents within each group should exhibit a large degree of similarity while the similarity among different clusters should be minimized.

There are lots of algorithms about [1][5][10]clustering including partitioning methods [5], hierarchical methods [9],density-based methods, grid based methods and model-based methods so on. MacQueen first put forward the k-means [2][3][4] clustering algorithm . *K*-means method has shown to be effective in producing good clustering results for many practical applications. One major disadvantage is that the number of cluster k must be specified prior to application and another is the sensitivity to initialization. The disadvantages in the k-means not only influence the efficiency of the algorithm but also accuracy of the clustering.

There are many existing document representation approaches including Boolean Approach, Probabilistic Retrieval Model, Language Model and Vector Space Model (VSM). At present the most popular document representation is Vector Space Model (VSM). VSM is an algebraic model for representing text documents as vectors of identifiers. The documents are represented as the vectors as $d_i=(w_{i1},w_{i2},w_{i3},w_{i4},\ldots\ldots\ldots.w_{in},)$.Themain advantages of this representation are its conceptual simplicity and its efficiency of similarity computation.

The motivational aim of this paper is to develop technique which will guide the user to get desired information with proper clustering of scientific documents in web or information retrieval systems. We propose a high performance document clustering algorithm "VFClustering" based on document's visual features. We integrate several visual features to represent documents and even we use thought of crossover and mutation in genetic algorithm to improve the k-means algorithm. To adjust the value of k and cluster center dynamically we merge and add cluster centers during the process of clustering.

II. DOCUMENT CLUSTERING KEY STEPS

Document Segmentation

As it is necessary to segment document into words before document feature extraction. We use lexicon based Word segmentation tools of the ICTCLAS. Its lexicon version is too low so that we add a large amount of new words into this lexicon and remove stop words from the result set of words segmentation.

Document Representation and Feature-Words Selection

Vector Space Model (VSM) is widely used in document clustering in which each n-dimensional vector represents a document. VSM can be represented as

$$d_i=((t_{i1},w_{i1}),(t_{i2},w_{i2}),\ldots\ldots\ldots.(t_{ik},w_{ik}),\ldots\ldots\ldots(t_{in},w_{in}))$$

where d_i= i^{th} Document
t_{ik}= k^{th} Keyword of i^{th} document
w_{ik}=w_{ik} weight of the k^{th} Keyword of i^{th} document

This paper adopts classical TF-IDF as the clustering keywords weight calculation method because it has an advantage in considering words occurrence frequency not only in a document but also in the whole date set. In this paper the size of each document is also taken into account and the parameter weight is defined.

$$w_{ij}=\frac{tf\left(i,t\_{ik}\right)*\log\left(\frac{N}{m}+0.01\right)*\frac{\sum_{x=1}^{x=n}size(x)}{N}}{size(i)}$$

Where *size* (*i*)= number of effective characters of the *i-th* document

$\frac{\sum_{x=1}^{x=n}size(x)}{N}$ = shows the average size of all the document in date set

Similarity Measurement

A document can be represented by a point in n-dimensional space after the document representation using VSM. While the similarity measurement between different documents was represented by the distance between corresponding points. The distance between the two points is in n-dimensional is the more similar the documents represented by the two points. There are many different methods to calculate the distance such as Mahalanobis distance and Euclidean distance. Documents' similarity here is presented by cosine similarity which is defined as

$$\cos(i,j)=\frac{d_i.d_j}{||d_i||.||d_j||}$$

III. ALGORITHM BASED ON VISUAL FEATURES

The main characteristics of document clustering algorithm based on visual features as follows:

- Five kinds of visual features are defined according to the analysis of content and structure of scientific document, including abstract (A), body (B), subtitle (S), title (T), and keyword (K).
- In view of the two drawbacks of k-means algorithm, the thought of crossover and mutation in genetic algorithm is used to improve the k-means algorithm.

Document Presentation Based on Visual Features

The mentioned model VSM represents document in two ways:

a. We can segment words and select clustering keywords according to words' frequency by mainly analyzing the body of the document (or) put clustering keywords selected in the first time into selection from the whole document and according to the clustering keywords' position.

b. Only title and abstract are analyzed to retrieve clustering keywords and do further clustering the result obtained in this way is not accurate enough.

A document representation based on visual features is defined with a full consideration of the importance of each visual feature in the whole document. We segment words on the basis of every visual feature independently and retrieve clustering keywords from each part with features extraction method introduced above. According to the importance of every visual feature, it shall be adjusted for the clustering keywords' weight (w_{ij}') of comprehensive document representation.

$$w'_{ij} = \frac{p * B(w_{ij}) + q * S(w_{ij}) + r * A(w_{ij}) + s * K(w_{ij}) + t * T(w_{ij}))}{p + q + r + s + t}$$

K-means Algorithm Optimization Based on Crossover and Mutation

This algorithm dynamically adjusts the values of k as well as cluster center by means of mergence and addition to take advantage of the idea of crossover and mutation in genetic algorithm during the process of clustering. Optimized clustering algorithm process is as follows:

Input: The initial number of cluster center k

Output: The clustering clusters formed finally

Step 1: Initialize cluster centers. It is necessary to check whether the newly selected cluster center is the existed one. Calculate the similarity between the current centers and compare the similarities with theλ.

Step 2 Calculate the similarity between each data and each cluster center. Then compare the biggest similarity with a given threshold λ. If the similarity is bigger, then the data shall be put into a cluster with its similarity biggest.

Step 3 Recalculate the center of each cluster which is definedas the arithmetic average value of all data in this cluster.

Step 4 Calculate the similarity for every pair of new clustercenters obtained in step 3. Thought of crossover ingenetic algorithm is used in here. Two clusters have to bemerged if the similarity between them is bigger thanλ.

Step 5 Execute steps 2, step 3 and step 4 once more, and finish this process if cluster center reaches a table value or maximize iteration times. Otherwise return to step 2 and continue to execute this process.

Fig.1: DOCUMENT CLUSTERING ALGORITHM BASED ON VISUAL FEATURES

IV. Evaluation of Clustering Accuracy

A. Accuracy

Evaluation of document clustering results rate and recall rate which reflect two different aspects of quality clustering must be taken into account together. We use the most commonly Evaluation recall rate and *F*1 test value to evaluate the effect of the document clustering. Artificial labeled theme *Ti* in data set corresponds to a clustering result set c_{ij} in clustering result. It is denoted as the

$$R(T_i, C_{ij}) = \frac{|T_i \cap C_{ij}|}{T_i}$$

$$P(T_i, C_{ij}) = \frac{|T_i \cap C_{ij}|}{C_{ij}}$$

$$F1 = \frac{2 * R(T_i, C_{ij}) * P(T_i, C_{ij})}{R(T_i, C_{ij}) + P(T_i, C_{ij})}$$

B. Experiment and Result Analysis

Text data sets are from 20 articles including 20 Documents(D), We pre-treatment the data set and separately extract five visual features of each document to a save to the database table.

The first step of the experiment: make word segment for five visual features independently and remove stop words and extract clustering keywords; then make a clustering for each visual feature that represents documents independently. The k-means shows the basic clustering algorithm and make the body representing the documents; all others adopt the improved algorithm. Table 1 shows the experimental results.

		k-means (%)	B(%)	A(%)	S(%)	K(%)	T(%)
D	R	69.89	69.89	68.12	69.89	52.89	51.12
	P	79.90	79.90	49.57	78.90	84.90	46.57
	F1	77.42	77.42	54.23	77.42	67.42	52.23

TABLE 1. RESULTS OF CLUSTERING BY FIVEVISUAL FEATURES

The results of the second step of experiment as follows:From the whole analysis of the two results in TABLE I and TABLE II it's obviously draw that the clustering result of the comprehensive visual features is better than any single visual feature in representing documents. Although the clustering results of visual features that consist of subtitle and keyword are slightly better than the visual feature body representing documents independently. The effective number of characters of subtitle and keyword is less than the body's, so it greatly enhances the efficiency of feature words selection when making words segment. The integrated independent visual feature includes body in which each one has the best clustering results to express text and its clustering results is almost the same as the one that integrate five visual features to represent a document.

		S,K (%)	B,S(%)	B,S,K(%)	B,S,K,A(%)	B,S,K,A,T(%)
D	R	80.85	85.11	95.74	95.74	95.74
	P	90.48	93.02	91.84	90.00	93.75
	F1	94.74	88.89	93.75	92.78	94.74

TABLE II. RESULTS OF CLUSTERING BY DIFFERENT COMBINATION

V. CONCLUSION

We implements a method to cluster the scientific documents based on visual features and through the deep analysis of these clustering results we find some useful information The basic 5 features represent the different aspects as body representing documents independently to cluster have the best accuracy and steadiness and subtitle is next, the clustering effect of abstract, keyword and title are not very good. The accuracy of clustering by combining subtitle and keyword is better than each of them individually. Clustering by combining subtitle and keyword can be an optimal choice if the efficiency is an essential factor. Clustering combining body, keyword and subtitle is a better choice if the higher accuracy is demanded .We use the thought of crossover and mutation in genetic algorithm to improve the k-means algorithm and heighten the efficiency greatly by adjusting the values of *k* and cluster center dynamically in the process of clustering.

IV. REFERENCE

[1] S. Guha, R. Rastogi, and K. Shim, "*An efficient clustering algorithm for large databases*," ACM SIGMOD international conference on Management of data, Volume 27 Issue 2, June 1998.

[2] A. Likasa, and N. Vlassisb, "Verbeekb. The global k-means clustering algorithm. Pattern Recognition," 2003, pageno. 451 – 461.

[3] J. A. Hartigan, and M. A. Wong, "A K-Means Clustering Algorithm," Journal of the Royal Statistical Society, Series C (Applied Statistics), Vol. 28, No. 1,1979, pageno.100-108.

[4] K. Wagsta, C. Cardie, S. Rogers, and S. Schroedl, “Constrained Kmeans Clustering with Background Knowledge,” Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pageno. 577-584.

[5] R. Dutta, I. Ghosh, A. Kundu, and D. Mukhopadhyay, “An Advanced Partitioning Approach of Web Page Clustering utilizing Content & Link Structure,” Journal of Convergence Information Technology Volume 4, Number 3, 2009.

[6] J. L. Neto, A. D. Santos, and C. A. A. Kaestner, “Alex A. Freitas,Document Clustering and Text Summarization,” InformationProcessing and Management, 2000.

[7] J.M. Pena , J.A. Lozano, and P. Larranaga, “An empirical comparisonof four initialization methods for the K-Means algorithm,” PatternRecognition Letters, 1999, pageno.1027-1040.

[8] L. Yanjun, M. Chung , and D. Holt, “Text document clustering basedon frequent word meaning sequences,” Data & KnowledgeEngineering, 2008, pageno. 381–404.

[9] J. F. Navarro, C. S. Frenk, and S. D. M. White, “A universal density profile from hierarchical clustering,” The astrophysical jouranl, 1997, pageno.490-493.

[10] A. K. Jain, and M. N. Murty, “Data Clustering: A Review,” ACM Computing Surveys (CSUR), 1999, pageno.264–323.

[11] N. Grira, Crucianu, and M. Boujemaa, “Unsupervised and semisupervised clustering: a brief survey,” 7th ACM SIGMMinternational workshop on Multimedia information retrieval, 2005, pageno.9-16.