International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 4 Issue 9|| September 2015 || PP.50-55

Document Clustering using GA enabled k-Means Algorithm

Nibedita Sahoo[1], Pappu Sharada[2], Manmath Namth Dash[3]
[1,3]*Associate Professor, Department of Computer Science Engineering, Gandhi Institute For Technology (GIFT), Bhubaneswar*
[2] *Assistant Professor, Department of Computer Science Engineering, Gandhi Engineering College, Bhubaneswar*

Abstract: *There are two important problems worth conducting research in the fields of personalized information services based on user model. One is how to get and describe user personal information, i.e. building user model, the other is how to organize the information resources, i.e. document clustering. It is difficult to find out the desired information without a proper clustering algorithm. Several new ideas have been proposed in recent years. But most of them only took into account the text information, but some other useful information may have more contributions for documents clustering, such as the text size, font and other appearance characteristics, so called visual features.In this paper we introduce a new technique called Closed Document Clustering Method (CDCM) by using advanced clustering metrics. This method enhances the previous method of cluster the scientific documents based on visual features, so called VF -Clustering algorithm. Five kinds of visual features of documents are defined, including body, abstract, subtitle, keyword and title. 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. Experimental result supports our approach as better concept.The main aim ofthis paper is to eliminate the redundant documents and set priority to each document in the cluster. In the five visual features, 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. The accuracy of clustering by combining subtitle and keyword is better than each of them individually, but is a little less than that by combining subtitle, keyword and body. If the efficiency is an essential factor, clustering by combining subtitle and keyword can be an optimal choice. The proposed system outperforms than the previous system.*
Keywords: *Document Clustering; k-Means; Visual Features; Genetic Algorithm*

I. Introduction

In recent years, personalized information services play animportant role in people's life. There are two important problems worth researching in the fields. One is how to get and describe user personal information, i.e. building user model, the other is how to organize the information resources, i.e. document clustering. Personal information is described exactly only if user behavior and the resource what they look for or search have been accurately analyzed. The effectiveness of a personalized service depends on completeness and accuracy of user model. The basic operation is organizing the information resources. In this paper we focus on document clustering. At present, as millions of scientific documents available on the Web. 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. Clustering plays an important role in analysis of user interests in user model. So high-quality scientific document clustering plays a more and more important role in the real word applications such as personalized service and recommendation systems. Clustering is a classical method in data mining research. Scientific document clustering [6, 8-9] 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.

In general, there are lots of algorithms about clustering [1,5,10,13], including partitioning methods [5], (k-means, k- medoids etc), hierarchical methods [16], (BIRCH, CURE, etc), density-based methods (DBSCAN, OPTICS, etc), gridbased methods (STING, CLIQUE, etc) and model-based methods, etc. In 1967, MacQueen first put forward the k-means [2-4,7], clustering algorithm. The k-means method has shown to be effective in producing good clustering results for many practical applications. However it suffers from some major drawbacks that make it inappropriate for some 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 two drawbacks of kmeans not only affect the efficiency of the algorithm but also influence clustering accuracy. There are many existing ddocument representation approaches [11], including Boolen Approach, Vector Space Model (VSM), Probabilistic Retrieval Model and Language Model. At present the most popular document representation is Vector Space Model (VSM).The most important goal of this paper is to develop a technique which will guide the user to get desired information with proper clustering of scientific documents in web or information retrieval systems.

In this paper we propose a high performance document clustering algorithm (called CDCM VF Clustering) based on document's visual features, and the document properties including Doctype, body, abstract, subtitle, keyword and title We integrate several visual features to represent documents. We also use the thought of crossover and mutation in genetic algorithm [14-15], to improve the k-means algorithm. We merge and add cluster centers during the process of clustering to adjust the value of k and cluster center dynamically. Experimental result shows that our approach is better in terms of clustering performance of the scientific documents.

II. Related Work

Information is better utilized when it is processed to be easier to find, better organized, or summarized for eas-ier digestion. Areas dealing with such problems are at the cross-roads of information retrieval, machine learning (e.g. classification and clustering), and statistical analysis. Text and web mining problems in particular use methodologies often spanning those areas. Document clustering is an area that deals with the unsupervised grouping of text documents into meaningful groups, usually representing topics in the document collection. It is one way to organize information with-out requiring prior knowledge about the classification of documents, and could be used as a base for document categorization by forming an initial classification.Document clustering has many applications, such as clustering of search engine results to present organized and understandable results to the user (e.g. Vivisimo1), clustering documents in a collection (e.g. digital li-braries), automated (or semi-automated) creation of document taxonomies (e.g. Yahoo and Open Directory styles), and efficient information retrieval by focusing on relevant subsets (clusters) rather than whole collections.Perhaps the most popular application of document clus-tering is the Google News2 service, which uses document clustering techniques to group news articles from mul-tiple news sources to provide a combined overview of news around the Web. Traditionally, document clustering has been studied as a centralized process; i.e. all data is assumed to be present at a central site, then a single process applies A more wider view of how clustering can be applied in distributed environments is outlined in Table 1.clustering to the data .

Table 1: Types of data and clustering process distribution Centralized Data - Centralized Clustering (CD-CC) This is the standard approach where the clustering process and data both reside on the same machine. Distributed Data - Centralized Clustering (DD-CC) Data might be dispersed across different machines, a typical case in the Web domain, while the clustering process runs on a single machine. Centralized Data - Distributed Clustering (CD-DC) Data is stored in one location, with clustering processes running on different machines accessing the same data; a typical case of parallel processing. Distributed Data - Distributed Clustering (DD-DC) The highest level of distribution, where both the data and the clustering processes are distributed.

A. Key Steps of Document Clustering 1. Document Segmentation

As it is necessary to segment document into words before document feature extraction, in our research, we use lexicon based Word segmentation tools of the ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System). However, 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.

2. Document Representation and Feature-Words Selection

As we know, Vector Space Model (VSM) is widely used in document clustering, in which each n-dimensional vector represents a document. In this paper, VSM can be represented as (1) [17]. 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. Furthermore, in this paper the size of each document is also taken into account, and the parameter weight is defined by (2) [17].

3. Similarity Measurement

After the document representation using VSM, a document can be represented by a point in n-dimensional space, while the similarity measurement between different documents was represented by the distance between corresponding points. The closer the distance between the two points is in n-dimensional, the more similar the documents represented by the two points is, and vice versa. To calculate the distance, there are many different methods, such as Mahalanobis distance and Euclidean distance, etc. The more similar two documents is, the more similar coefficient close to 1, conversely, the similar coefficient is close to 0. In this paper, documents' similarity here is presented by cosine similarity which is defined by (3) [17].

III. Proposed Model

A. Closed Document Clustering Method based on Visual Features

The main characteristics of document clustering algorithm based on visual features, so called on visual features, so called VF-Clustering. We introduce Closed Document Clustering Method (CDCM) by using advanced clustering metrics.

In this paper we introduce a new technique called Document division similarity metrics document clustering approach by using the advanced search techniques to eliminate the redundant duplicate document and this techniques sets the priority to each document while clustering. We compare the documents with various document formats using the previous Visual Feature method .This method works close to cluster the documents.

1) Six kinds of visual features are defined according to the analysis of content and structure of scientific document, including document type (T) body (B), abstract (A), subtitle (S), keyword (K) and title

(T). And the importance of these features to scientific document clustering will be compared through experiments.

2) 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. Adjust the values of k and cluster center dynamically by merging and adding cluster centers in the process of clustering. The implementation of clustering algorithm introduces below.

1. Document Presentation Based on Visual Features

As the most widely used document presentation method, the mentioned model VSM represents document in two ways. In one way 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, their weight shall be adjusted if they occurrences in title or abstract. In the other way, only title and abstract are analyzed to retrieve clustering keywords and do further clustering, though effective, the result obtained in this way is not accurate enough.In this paper, a document representation based on visual features is defined with a full consideration of the importance of each visual feature in the whole document. Therefore, we segment words on the basis of every visual feature independently and retrieve clustering keywords from each part with features extraction method introduced above. And according to the importance of every visual feature, it shall be adjusted for the clustering keywords' weight of comprehensive document representation, with be obtained by (4)[a] and comprehensive document presentation shown in (5)[a] where B(Wij)means the weight of clustering keyword i in body part, and the values of a, b, c, d, e must be either part equal to 0 or all no less than 1. In our experiment we set the values of a and b equal to 2, others equal to 1. K-means Algorithm Optimization Based on Crossover and

Mutation

Take advantage of the idea of crossover and mutation in genetic algorithm during the process of clustering, this algorithm dynamically adjusts the values of k as well as cluster center by means of mergence and addition, to achieve kmeans algorithm optimization. Optimized clustering algorithm process is as follows[a].

Input: The initial number of cluster center k, Similarity Threshold. In our experiment we set the value of k equal to 4.

Output: The clustering clusters formed finally (the number of clusters not necessarily equals k).

Step 1. Initialize cluster centers. First of all, it is necessary to check whether the newly selected cluster center is the existed one. If it is, the cluster center can be reproduced. Or else calculate the similarity between the current centers and selected one and compare this similarity with metric. If the Smiilarity is bigger, reselect a document as a new center and go back to execute step 1 once more until the number of cluster center equal to k.

Step 2. Calculate the similarity between each data and each cluster center, and then compare the biggest similarity with a given threshold On one hand, if the similarity is bigger, the data shall be put into a cluster with its similarity biggest.On the other hand, the thought of mutation in the genetic algorithm is used in here, the data should be added into cluster enter as a new one which can cause cluster center number change.

Step 3. Recalculate the center of each cluster which is defined as the arithmetic average value of all data in this cluster. For example, it is assumed that there are 3 documents in the first cluster, which are[a]

Step 4. Calculate the similarity for every pair of new cluster centers obtained in step 3. The thought of crossover in genetic algorithm is used in here. Two clusters have to be merged if the similarity between them is bigger than _. For example, there are 2 cluster centers: center1 and center 2 and the two merged into one cluster center, that is[a]

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

IV. Test Environment

A. Clustering Results

Text data sets are from 195 articles of scientific and technical document including 47 articles of Clustering Algorithm (CA), 58 articles of Data Mining (DM) 43 articles of Cloud Computing (CC) and 47 articles of Genetic Algorithm (GA). We pre-treatment the data set, we separately extract five visual features of each document to a save to the database table. The overall performance of CDCM is better than the previous method. The first step of the experiment: firstly, make word segment For six visual features independently, remove stop words and extract clustering keywords; then, make a clustering for each visual feature that represents documents independently. The experimental result is shown in Table 2.Where the k-means shows the basic clustering algorithm and make the body representing the documents, all others adopt the improved algorithm.

Table 2: Results of Clustering by Five Visual Features

		k-means(%)	B (%)	A (%)	S (%)	K (%)	T (%)
CA	R	76.60	76.60	74.47	76.60	55.32	53.19
	P	83.72	83.72	53.03	85.71	92.86	48.93
	F1	80.00	80.00	61.95	80.90	69.33	50.97
DM	R	93.10	93.10	91.38	93.10	86.25	87.93
	P	90.00	90.00	82.81	88.52	84.85	82.26
	F1	91.53	91.53	86.89	90.76	85.54	85.00
CC	R	93.02	93.02	90.69	90.70	95.35	93.02
	P	90.69	90.69	88.38	86.04	97.62	78.43
	F1	91.84	91.84	89.50	88.30	96.47	85.11
GA	R	93.62	93.62	40.43	89.36	100.00	91.79
	P	84.62	84.62	95.00	86.27	79.66	58.11
	F1	88.89	88.89	56.72	87.79	88.68	71.07

Through the analysis of the first step of the experimental results, we could conclude as follows:

1. Because the value of the k is equal to 4, so the basic algorithm and the improved algorithm to clustering have the same results when make the body representing document independently.

But the clustering running time are reduced when use the improved algorithm.

2. The clustering performance by visual features body and subtitle are best in representing documents independently, and good steady is exhibited in these types of data sets. What's more, the visual feature body has slightly better clustering results than subtitle.
3. The visual feature keyword is better than abstract and title in clustering effect, moreover, abstract and title are poor in the stability of the clustering result by representing document independently. Among these three visual features, clustering has a good effect in a new subject or a subject with fewer applications. However, it has a relatively poor effect in subject with extensive applications.
4. The visual features title has poor clustering results in subject with extensive applications. Under the first step of the experimental results, we make an analysis of clustering results obtained through different visual features representing the document independently. We make different combinations of visual features to represent the document and clustering.

The result is shown in Table 3.

Table 3:

		S, K (%)	B, S (%)	B, S, K (%)	B, S, K, A (%)	B, S, A, K, T (%)
CA	R	80.85	85.11	95.74	95.74	95.74
	P	90.48	93.02	91.84	90.00	93.75
	F1	85.39	88.89	93.75	92.78	94.74
DM	*R*	94.83	98.28	93.10	93.10	94.83
	P	96.49	90.48	100.00	96.43	98.21
	F1	95.65	94.21	96.43	94.74	96.49
CC	R	97.67	100.00	100.00	97.67	97.67
	P	97.67	100.00	100.00	100.00	100.00
	F1	97.67	100.00	100.00	98.82	98.82
GA	R	95.74	93.62	100.00	95.74	100.00
	P	84.91	95.65	95.92	95.74	95.92
	F1	90.00	94.62	97.92	95.74	97.92

The overall performance of CDCM versus VF methods with different combinational settings higher shown in fig. 1.

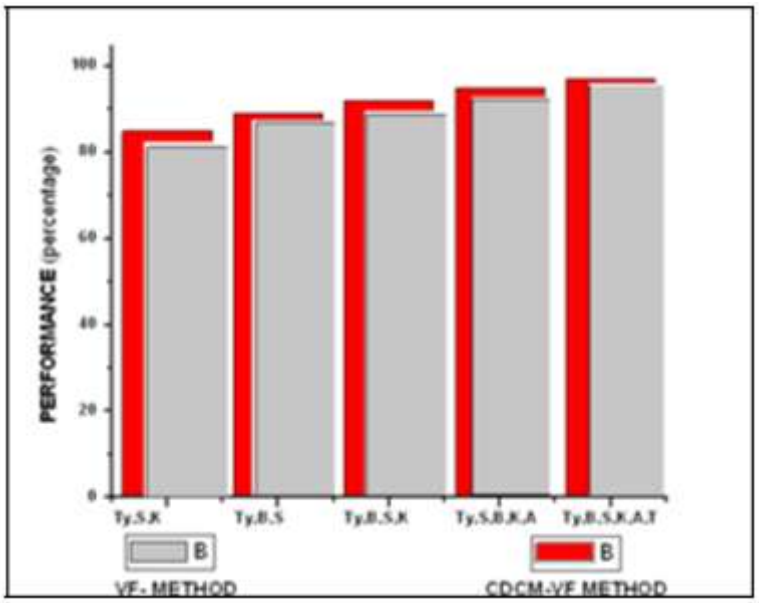

Fig. 1:

V. Conclusion

This paper implements a method to cluster the scientific documents based on visual features (CDCM-VF-Clustering). And through the deep analysis of these clustering results we find some useful information as follows:

1. In the six visual features, body representing documents Independently to cluster have the best accuracy and steadiness, and subtitle is next. However, the clustering effect of abstract, keyword and title are not very good, especially in the widely applied field of knowledge clustering.
2. The accuracy of clustering by combining subtitle and keyword is better than each of them individually. Moreover, operation time can be saved greatly for the less effective characters in the two parts. If the efficiency is an essential factor, clustering by combining subtitle and keyword can be an optimal choice.
3. If the higher accuracy is demanded, clustering combining body, subtitle and keyword is a better choice. This paper also uses the thought of crossover and mutation in genetic algorithm to improve the k-means algorithm and heightens the efficiency greatly by adjusting the values of k and cluster center dynamically in the process of clustering.

References

[1]. S. Guha, R. Rastogi, K. Shim,"An efficient clustering algorithm for large databases", ACM SIGMOD international conference on Management of data, Vol. 27 Issue 2, June 1998.
[2]. A. Likasa, N. Vlassisb,"Verbeekb. The global k-means clustering algorithm. Pattern Recognition", 2003, pp. 451– 461.
[3]. J. A. Hartigan, M. A. Wong,"A K-Means Clustering Algorithm," Journal of the Royal Statistical Society, Series C (Applied Statistics),Vol. 28, No. 1,1979, pp.100-108.
[4]. K. Wagsta, C. Cardie, S. Rogers, S. Schroedl,"Constrained Kmeans Clustering with Background Knowledge", Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp.577-584.
[5]. R. Dutta, I. Ghosh, A. Kundu, D. Mukhopadhyay,"An Advanced Partitioning Approach of Web Page Clustering utilizing Content & Link Structure", Journal of Convergence Information Technology Vol. 4, No. 3, 2009.
[6]. J. L. Neto, A. D. Santos, C. A. A. Kaestner,"Alex A. Freitas, Document Clustering and Text Summarization", Information Processing and Management, 2000.
[7]. J.M. Pena , J.A. Lozano, P. Larranaga,"An empirical comparison of four initialization methods for the K-Means algorithm", Pattern Recognition Letters, 1999, pp. 1027-1040.
[8]. L. Yanjun, M. Chung, D. Holt,"Text document clustering based on frequent word meaning sequences", Data & Knowledge Engineering, 2008, pp. 381–404.
[9]. E Rasmussen, P. Hall, E. Cliffs,"Clustering algorithms", Information Retrieval, 1992, pp. 419-442.
[10]. A. K. Jain, M. N. Murty,"Data Clustering: A Review", ACM Computing Surveys (CSUR), 1999, pp.264–323.

[11]. W. B. Cavnar,"Using An N-Gram-Based Document Representation With A Vector Processing Retrieval Model", Proc. of TREC-3 (Third Text REtrieval Conference), Gaithersburg, 1994.
[12]. G. Salton, C. Buckley,"Term-weighting approaches in automatic text retrieval", Information Processing and Management 24, 513-523. 1988. Reprinted in: Sparck Jones, K. and Willet, P. Eds. Readings in Information Retrieval, 1997, pp. 323-328.
[13]. N. Grira, Crucianu, M. Boujemaa,"Unsupervised and semisupervised clustering: a brief survey", 7th ACM SIGMM international workshop on Multimedia information retrieval, 2005, pp. 9-16.
[14]. U. Maulik, S. Bandyopadhyay,"Genetic algorithm-based clustering technique, Pattern Recognition", 2000, pp. 1455-1465.
[15]. K. Krishna, M. Narasimha Murty,"Genetic K-Means Algorithm, Item Identi" 1999, pp.1083-4419.
[16]. J. F. Navarro, C. S. Frenk, S. D. M. White,"A universal density profile from hierarchical clustering", The astrophysical jouranl, 1997, pp. 490-493.
[17]. "IEEE International Conferences on Internet of Things, and Cyber, Physical and Social Computing", 2011.