Effective K-Means Document Clustering using Dictionary Defined Lexical Analyzer (DDLA)

R.Ranga Raj
Head of the Department, Computer Science
Hindusthan College of Arts and Science
Coimbatore

M.Punithavalli, PhD.
Director
Department of Computer Applications
Sri Ramakrishna Engineering College
Coimbatore

ABSTRACT

Due to tremendous increase in number of documents, clustering of such document is difficult one. Document Clustering is the process of grouping related documents from the large collection of database. The mining of such related documents from the database which are unlabelled is a challenging one. To overcome this process, clustering is used to filter the unlabelled documents from the large collection of database.

In this paper, a new concept is introduced for the document clustering by using k-means Enhanced Approach algorithm [1] with the Dictionary Defined Lexical Analyzer (DDLA). Basically K-Mean algorithm clusters the numeric values efficiently. But with the inclusion of DDLA the characters, words and sentences can also be clustered. Based on the weights, documents are clustered [7] by using bisecting k-means algorithm [1, 2] and topic detection method. The discovery of meaningful labels for the document is based on semantic similarity [8]. The efficient clustering of unlabeled documents with enhanced K-Mean algorithm and DDLA is one of the techniques which make clustering in an easiest way.

General Terms

Effective Clustering Using DDLA

Keywords

Clustering, K-Means Enhanced Approach Algorithm, Lexical Analyzer, Defined Dictionary DDLA.

1. INTRODUCTION

Now a day’s retrieval of the information from a collection of a document is important one. Searching for the related documents from World Wide Web is a becoming a challenge. The function of today's search engine is to perform string matching, so the documents filtered may not be so relevant according to user's need. In order to get an efficient routing, the computer organizes the document body into a meaningful cluster chain of commands with the help of good clustering [2] approach. It helps us to overcome the deficiencies of traditional information filtering methods. It is a more specific technique for unsupervised document organization [9].

The unsupervised learning is the process of clustering unlabeled document. Documents which are labeled or having set of topics can be clustered in an easiest way. It is necessary for the cluster not only to indicate the main concept of the cluster but also to make difference between clusters.

The accuracy is improved by incorporating semantic features. In this paper certain modifications are made in the existing semantic features [8]. It uses Enhanced K-Mean algorithm with DDLA which allows, forming cluster on the basis of its meaning. Also it captures each term of a character, word and sentences as well as document other than frequency. Based on similarity of a document [7] with a predefined dictionary each labeled term is weighted .The terms that possess maximum weights are considered as top terms and are added to the term vector and also synonyms/hyponyms of each word are added. These concepts are used in document clustering and topic discovery. Cosine similarity is similarity measure used.

2. EXISTING WORK

Mostly, Vector Space Model [5] is used to illustrate data for the classification of text and clustering. VSM point outs each document as a feature vector which contains term weights. The TF-IDF weight (term frequency-inverse document frequency) is a weight used to calculate the importance of a word in a document in a body. As the appearance of words increases its importance increases but is offset by the frequency of the word in the corpus. Similarities between the documents are based on the feature vector.

Common ones include the cosine measure and the Jaccard measure. A survey of document clustering algorithms [2] with topic discovery is presented. If the topic is represented in clustering keywords then it is used to compare the proposed and existing topic detection. There is a need to consider a set of keywords for the clustered documents. For indexing documents, the problem of finding a good set of keywords is similar to that of determining term weights. Terms that have high TFIDF Terms are used as cluster keywords. Clustering is done by bisecting k-means algorithm [2] using cosine similarity as the similarity measure. After that the centers of the defined cluster are taken as the representations of the topic.

Induced bisecting k-means clustering algorithm used above is based on the standard bisecting k-means algorithm [3]. A simplified version of the method is as follows. Two elements of largest distances are chosen as the seeds of two clusters. Then assign all other terms closest to any one of the cluster seed. Then compute the center of cluster seeds. Illustration of items that naturally allows defining a center which typically is not an item proper but a weighted sum of items is needed. The new centers serve as new seeds for finding two clusters. The process is repeated till the two centers are converged up to some predefined precision. If the diameter of a cluster is larger than a specified threshold value, the whole procedure is applied recursively to that cluster. The algorithm therefore finds a binary tree of clusters.

It is obtained when the points no longer switch clusters (or alternatively centroid are no longer changed).For most practical purposes, it proves to be fast enough to generate good clustering solution.

3. K-MEANS ALGORITHM

The effective clustering method with reduced complexity is achieved by k-means algorithm. The clustering of unlabeled documents from the large database can do easily by using k-means with the enhanced approach.

3.1 Enhanced Approach

In the enhanced clustering method discussed in this paper [1] both the phases of the original k-means algorithm are modified to improve the accuracy and efficiency. In the enhanced method data points are assigned to the clusters.

Algorithm 1: Assigning data point to cluster

```
Input:
    D={d1,d2,….,dn}//set of n datapoints
    C={c1,c2,…..,cn}//set of k clusters
Output:
    A set of k clusters
Steps:
        1. Compute the distance of each data- point di (1<=i<=n) to all the centroids cj (1<=j<=k) as d(di, cj);
        2. For each data-point di, find the closest centroid cj and assign di to cluster j.
        3. Set ClusterId[i]=j; // j:Id of the closest cluster
        4. Set Nearest_Dist[i]= d(di, cj);
        5. For each cluster j (1<=j<=k), recalculate the centroids;
        6. Repeat
        7. for each data-point di,
          7.1 Compute its distance from the centroid of the present nearest cluster;
          7.2 If this distance is less than or equal to the present nearest distance, the data-point stays in the cluster;
Else
        7.2.1 For every centroid cj (1<=j<=k) Compute the distance d (di, cj);
End for;
            7.2.2 Assign the data-point di to the cluster with the nearest centroid cj
            7.2.3 Set ClusterId[i]=j;
            7.2.4 Set Nearest_Dist[i]= d(di, cj);
End for;
        8. For each cluster j (1<=j<=k), recalculate the centroids;
```

In the above algorithm, Euclidean distance is used for determining the closeness of each data point. The distance between one vector X = (x1, x2 ...xn) and another vector Y = (y1, y2 ….yn) is obtained as

$$d\,(X,Y)=\sqrt{(x1-y1)^2+(x2-y2)^2+\ldots\ldots+(xn-yn)^2}$$

The initial centroids of the clusters are given as input to the second phase, for assigning data-points to appropriate clusters. The steps involved in this phase are outlined as Algorithm 1.

4. LEXICAL ANALYZER

A program or function which performs lexical analysis is called a lexical analyzer, lexer, or scanner. Lexical analysis is the process of converting a sequence of characters into sequence of tokens [10, 11].

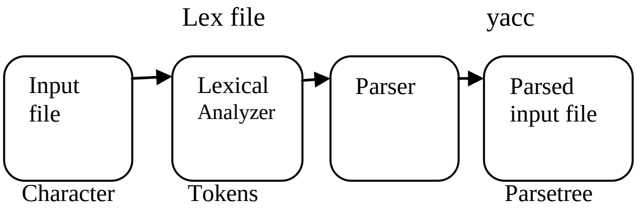


Figure 1: Process of Lexical Analyzer

4.1 Parser

It analyses a text made of sequence of tokens to determine its grammatical structure because human sentences are not easily parsed by programs, as there is substantial ambiguity in the structure of human language, whose usage is to convey meaning (or semantics) amongst a potentially unlimited range of possibilities. It is an interpreter/compiler used to check correct syntax and builds a data structure in input tokens.

For example, a calculator program would look at an input such as "12*(3+4) ^2" and split it into the tokens 12, *, (, 3, +, 4,), ^, 2, each of which is a meaningful symbol in the context of an arithmetic expression. The lexer would contain rules to tell it that the characters *, +, ^, (and) mark the start of a new token, so meaningless tokens like "12*" or "(3" will not be generated [10].

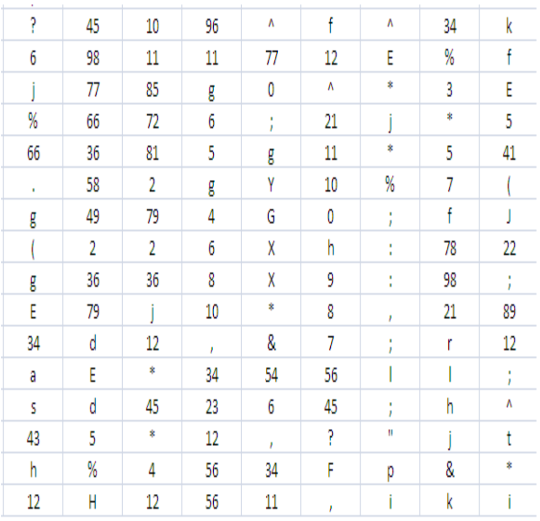
4.2 DDLA

Dictionary Defined Lexical Analyzer (DDLA) is the process of finding similarity of documents with the predefined dictionary. Generally Lexical Analysis is the process of converting sequence of characters into tokens [10]. The main objective is to cluster a character, word and sentence from the documents in the large database. The documents are related to different fields such as computer, medical, management, library etc.The predefined dictionary in DDLA which helps to find the matches between the character, word and sentence. Based on this it allots the weights to the document.

6. PROPOSED WORK

By using K-means Enhanced Approach with the Dictionary Defined Lexical Analyzer (DDLA), the sentence, word and character in the documents are simulated. Hence, the following (Table1) and (Table2) describe the result of clustering the character and word.

Table 1: Database for Character



?	45	10	96	^	f	^	34	k
6	98	11	11	77	12	E	%	f
j	77	85	g	0	^	*	3	E
%	66	72	6	;	21	j	*	5
66	36	81	5	g	11	*	5	41
.	58	2	g	Y	10	%	7	(
g	49	79	4	G	0	;	f	J
(	2	2	6	X	h	:	78	22
g	36	36	8	X	9	:	98	;
E	79	j	10	*	8	,	21	89
34	d	12	,	&	7	;	r	12
a	E	*	34	54	56	l	l	;
s	d	45	23	6	45	;	h	^
43	5	*	12	,	?	"	j	t
h	%	4	56	34	F	p	&	*
12	H	12	56	11	,	i	k	i

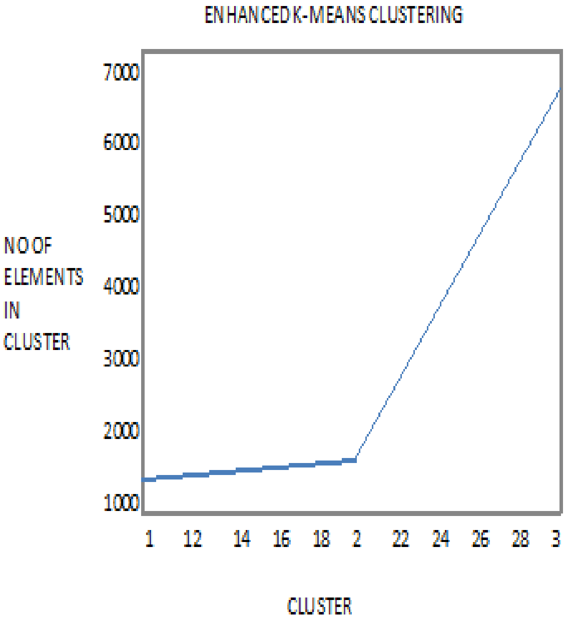
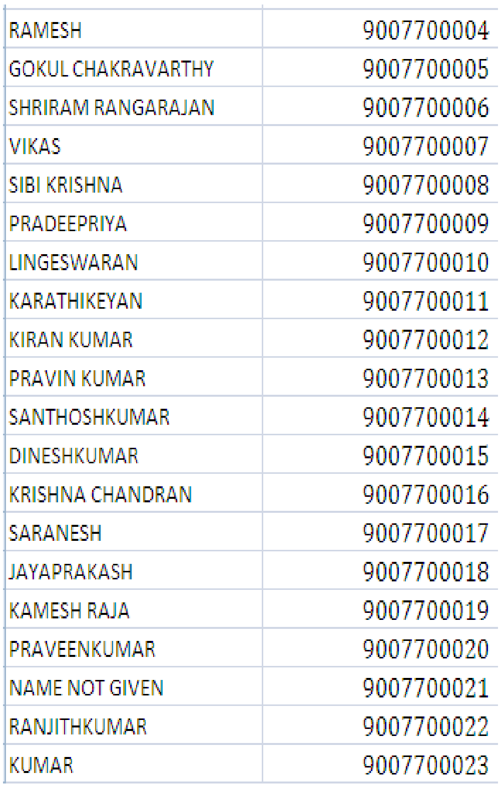


Figure 2: Number of Characters into three Clusters

The result of clustering of character is shown in (Table 1) which contains number of characters in the documents that are converted in to tokens and Figure 2 which shows the result of clustering two characters.

Table 2: Database for word



RAMESH	9007700004
GOKUL CHAKRAVARTHY	9007700005
SHRIRAM RANGARAJAN	9007700006
VIKAS	9007700007
SIBI KRISHNA	9007700008
PRADEEPRIYA	9007700009
LINGESWARAN	9007700010
KARATHIKEYAN	9007700011
KIRAN KUMAR	9007700012
PRAVIN KUMAR	9007700013
SANTHOSHKUMAR	9007700014
DINESHKUMAR	9007700015
KRISHNA CHANDRAN	9007700016
SARANESH	9007700017
JAYAPRAKASH	9007700018
KAMESH RAJA	9007700019
PRAVEENKUMAR	9007700020
NAME NOT GIVEN	9007700021
RANJITHKUMAR	9007700022
KUMAR	9007700023

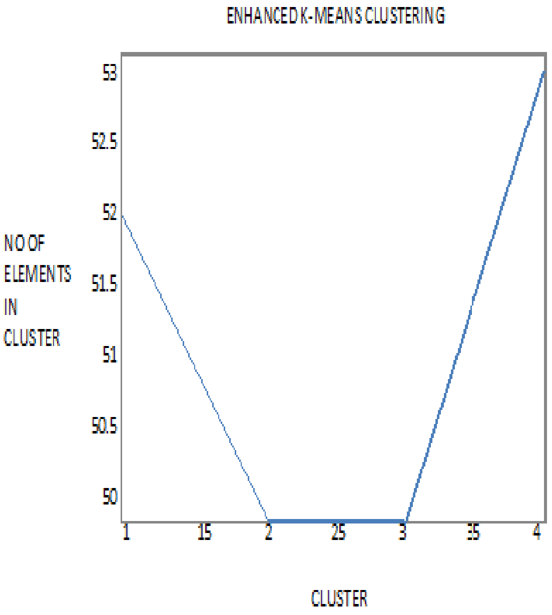


Figure 3: Number of Characters into four Clusters

Figure 3 which show the result of clustering 4 words from the document that is specified in (Table 2).

Table 3: Our Database Contains

Document Type	Actual No of Documents
Cacm	3204
Cisi	1460
Cran	1398
Med	1033

This dataset in (Table 3) is usually referred to as Classic3 dataset (CISI, CRAN and MED only), and sometimes referred to as Classic4 dataset. These are the documents we used for our Document Clustering.

In order to extract the concept of clustering sentence by K-Means Enhanced Approach Algorithm using Lexical Analyzer Defined Dictionary DDLA.

After clustering, the clustered numbers of documents in the Classsic4 dataset are shown as result in (Table4).

Table 4: After Clustering

Document Type	Clustered No of Documents
Cacm	2900
Cisi	1399
Cran	1234
Med	989

The below Figure 4 and 5 shows the number of Actual and Measured value of clusters which contains the same Documents.

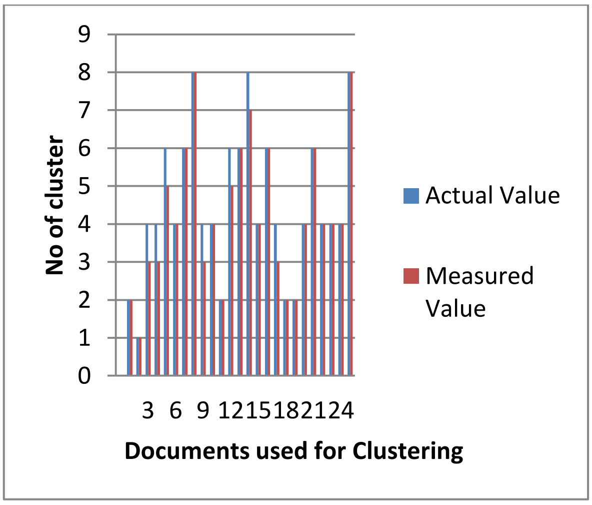


Figure 4: Clustering Results for Combined database

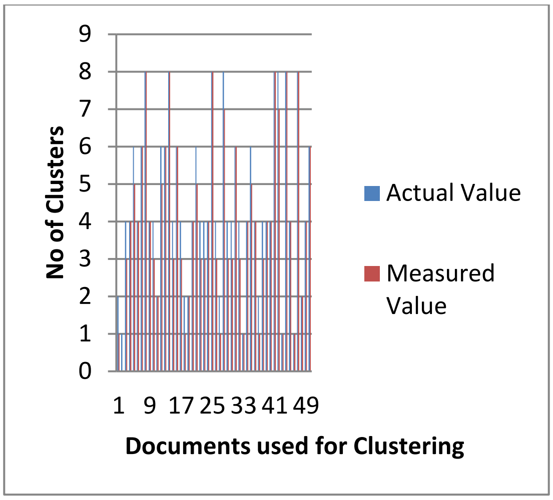


Figure 5: Clustering Results for Combined database

7. CONCLUSION

The Clustering of unlabeled documents from large set of database is one of the challenge .In this paper, the concept K-Means Enhanced Approach Algorithm with Dictionary Defined Lexical Analyzer (DDLA) focuses on clustering of word, character and documents are categorized and by creating separate composed of most appropriate used words related to that category. The function of the lexical is to read the character, word and document line by line. So, each and every word from the document is compared with the library functions. As a result, matching ratio is calculated. Thus we conclude that the clustering of unlabeled documents from large dataset will be efficient by using the above specified approach.

If the document has highest ratio while matching with libraries are given suitable weights and are treated as top terms by lexical analyzer [12]. According to the weights given, the documents are grouped in relevant categories.

REFERENCES

[1] "Improving the accuracy and efficiency of K-Mean Clustering Algorithm", by K.A. Abdul Nazeer, M.P. Sebastian. Proceeding of the world congress on Engineering 2009 vol I WCE 2009, July 1-3, 2009, London, U.K.

[2] Korean Text Extraction by "Local Color Quantization and K-means Clustering" In Natural Scene Anh-Nga Lai*, KeonHee Park, Manoj Kumar, GueeSang Lee*Department of Computer Science, Chonnam National University, 500-757 Gwangju, Korea ltanhnga@gmail.com, gslee@chonnam.ac.kr

[3] "Cluster Analysis for Gene Expression Data," Daxin Jiang, Chum Tong and Aidong Zhang, IEEE Transactions on Data and Knowledge Engineering, 16(11): 1370-1386, 2004.

[4] "Fast Document Clustering Based on Weighted Comparative Advantage"Jie Ji Intelligent System Lab The University of Aizu Aizuwakamatsu, Fukushima, Japan d8102102@u-aizu.ac.jp

[5] ”A Comparison of Document Clustering Techniques”,Michael Steinbach,George Karypis. Department of Computer Science University of Minnesota Technical Report #00-034 steinbac, karypis, kumar@cs.umn.edu Vipin Kumar

[6] ”Clustering Of Image Data Set Using K-Means and Fuzzy K-Means Algorithms” Vinod Kumar Dehariya I.T dept. S.A.T.I Vidisha (M.P), India Vidisha (M.P), India Vidisha (M.P), Indiavkdworld@yahoo.com.

[7] ”Document Clustering in Correlation Similarity Measure Space” Taiping Zhang; Yuan Yan Tang; Bin Fang; Yong Xiang Knowledge and Data Engineering, IEEE Transactions on Volume: 24 ,,2012

[8] “A Web Search Engine-Based Approach to Measure Semantic Similarity between Words” Bollegala, D.; Matsuo, Y.; Ishizuka, M. Knowledge and Data Engineering, IEEE Transactions on Volume: 23 ,,2011

[9] ”Spoken Document Retrieval With Unsupervised Query Modeling Techniques Chen”, B.; Kuan-Yu Chen; Pei-Ning Chen; Yi-Wen Chen Audio, Speech, and Language Processing, IEEE Transactions on Volume: 20 , Issue: 9 ,2012

[10] ”Data Extraction for Deep Web Using WordNet Jer Lang Hong Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on Volume: 41 , Issue: 6 ,2011

[11] ”Unsupervised Motif Acquisition in Speech via Seeded Discovery and Template Matching Combination” Muscariello, A.; Gravier, G.; Bimbot, F.
Audio, Speech, and Language Processing, IEEE Transactions on Volume: 20 , Issue: 7,2012

[12] Automatic Discovery of Personal Name Aliases from the Web Bollegala, D.; Matsuo, Y.; Ishizuka, M. Knowledge and Data Engineering, IEEE Transactions on Volume: 23 , Issue: 6 ,2011