Volume 2, Issue 5, May 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Classification of Document Clustering Approaches

Swatantra kumar sahu* *M.G.C.G Vishawavidyalaya Satna* India
Neeraj Sahu *Singhania University Rajasthan* India
G.S.Thakur *MANIT Bhopal* India

***Abstract*— This paper presents classification of Document Clustering Approaches. Document Clustering Approaches are Based on classification of Large data Sets . In this paper User search Information with few keywords by Internet. Document Clustering play major role for Information searching process by user. The Experimental Results show the proposed approach out performs.**

***Keywords*— Classification, Clustering, Data Sets.**

I. INTRODUCTION

Document Clustering is an important issue in text mining. Clustering has been widely applicable in different areas of science, technology, social science, biology, economics, medicine and stock market. Clustering problem appears in other different field like pattern recognition, statistical data analysis, bio-informatics, etc.There exist classification of Document Clustering Approaches in the literature[2]. classification of Document Clustering Approaches are mainly divided into Three categories shown in Fig.1

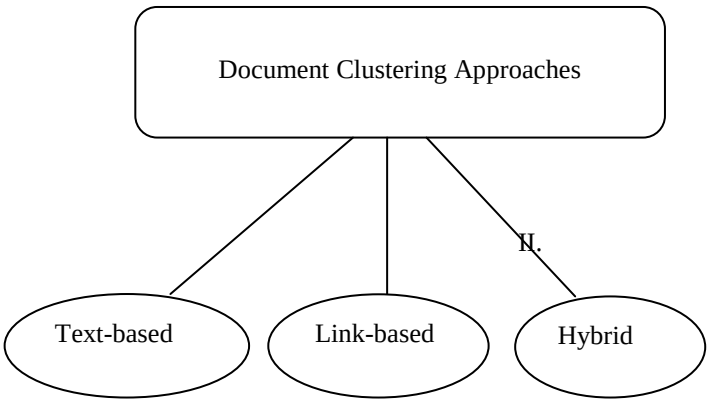

Fig.1 Structure of Document Clustering Approaches classification

In last recent years lot of research work has been done on Document Clustering. Some contributions are as follows:

In 2002 Beil et al. worked to improve the cluster accuracy using frequent item based technique and find overlapping clusters and meaning full cluster label.

In 2010 Chun-Ling Chen,Frank S.C. Tseng,Tyne Liang,Fuzzy-based Multi-label Document Clustering (FMDC) algorithm concentrated on clustering accuracy and used frequent item based clustering concept and find overlapping cluster, semantic discovery and meaningful cluster label.The above mentioned work suffers from lack of efficiency and accuracy. The high complexity and low accuracy are still issues and challenges in the clustering. This motivates the study of Document Clustering[2].

The paper is organized as follows. Section-I described the introduction and review of literatures. In Section-II, Classification of Document Clustering Approaches are described. In Section-III, Methodology of document clustering is described. In Section-IV, Experimental results are described. In Section-V, Evaluation measurement is described. Finally, we concluded and proposed some future directions in Conclusion Section.

DOCUMENT CLUSTERING APPROACHES CLASSIFICATION

1. Supervised: In Supervised classification method, a set of predefines classes are given.
2. Unsupervised: In Unsupervised classification methods, a set of predefine classes are not given. This is also known clustering.

Approaches classification :

A)Text based:Text based is depend on content of document.
B)Link based: Link based is depend on link structure of the pages.
C)hybrid:Hybrid is depend on content and link of document.

Text based: Text based classification of Document Clustering Approaches are mainly divided into five categories shown in Fig.2:

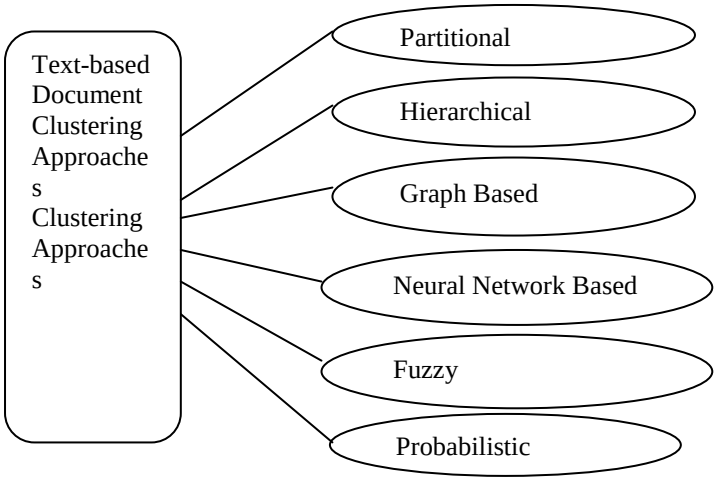

Fig.2 Text based Structure of Document Clustering Approaches classification [1]

III. METHODOLOGY

In the Document clustering different the steps are used. The steps are as follows:

A. Document Collection:

In this phase we collect relevant documents like e-mail, news, web pages etc. from various heterogeneous sources. These text documents are stored in a variety of formats depending on the nature of the data. The datasets are downloaded from UCI KDD Archive [4]. This is an online repository of large datasets and has wide variety of data types.

B. Text Pre-processing:

Text pre-processing means transform documents into a suitable representation for the clustering task. The text documents have different stop words, punctuation marks, special character and digits and other characters. Algorithm 1 is removed HTML Tages,Stop words from the text Documents. After removing stop words, word stemming is performed .Word stemming is the process of suffix removal to general word stems. A stem is a natural group of words with similar meaning. In text-pre-processing we performed the following task:

Removal of HTML tags and special character
Removal stop words
Word stemming

The algorithm 1 is proposed for Text pre-processing. The proposed algorithm removes special characters and stop words from the document[2].

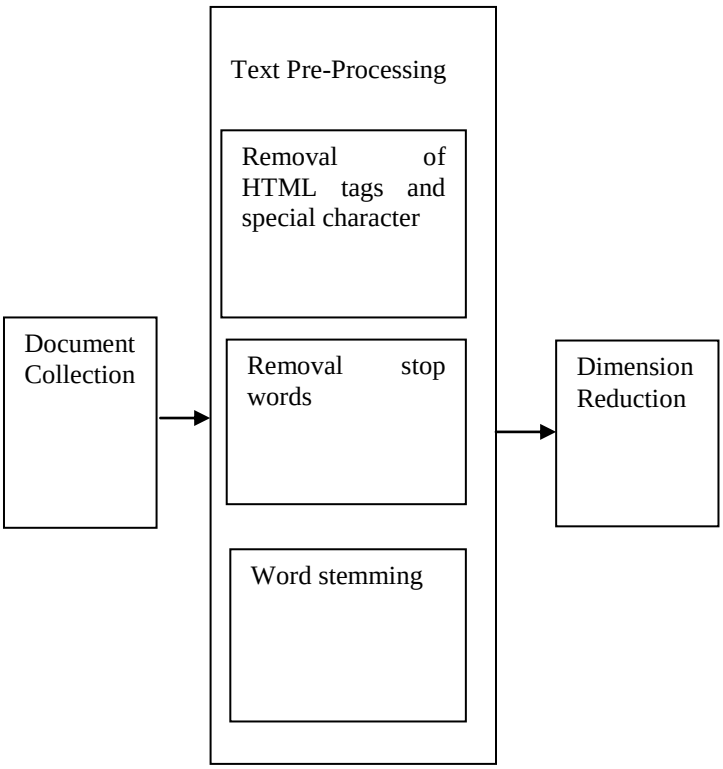

Fig.3 Steps of Methodology[6]

C. Preprocessing :

Preprocessing consists of steps that take as input a plain text document and output a set of tokens to be included in the vector model. These steps typically consist of

1.Filtering :

The process of removing special characters and punctuation that are not thought to hold any discriminative power under the vector model. This is more critical in the case of formatted documents,such as web pages, where formatting tags can either be discarded or identified and their constituent terms attributed diferent weights [3].

2.Tokenization :

Splits sentences into individual tokens, typically words. More sophisticated methods, drawn from the field of NLP, parse the grammatical structure of the text to pick significant terms or chunks (sequences of words), such as noun phrases [3].

3. Stemming :

The process of reducing words to their base form, or stem. For example, thewords "connected,""connection", "connections" are all reduced to the stem "connect." Porter's algorithm is the de facto standard stemming algorithm[3].

4.Stopword removal :

A stopword is defined as a term which is not thought to convey any meaning as a dimension in the vector space (i.e. without context). A typical method to remove stopwords is to compare each term with acompilation of known stopwords. Another approach is to first apply a part-of-speechtagger and then reject all tokens that are not nouns, verbs, or adjectives[3].

5. Pruning: Removes words that appear with very low frequency throughout the corpus. The underlying assumption is that these words, even if they had any discriminating power, would form too small clusters to be useful. A pre-specified threshold is typically used, e.g. a small fraction of the number of words in the corpus. Sometimes words which occur too frequently (e.g. in 40% or more of the documents) are also removed[3].

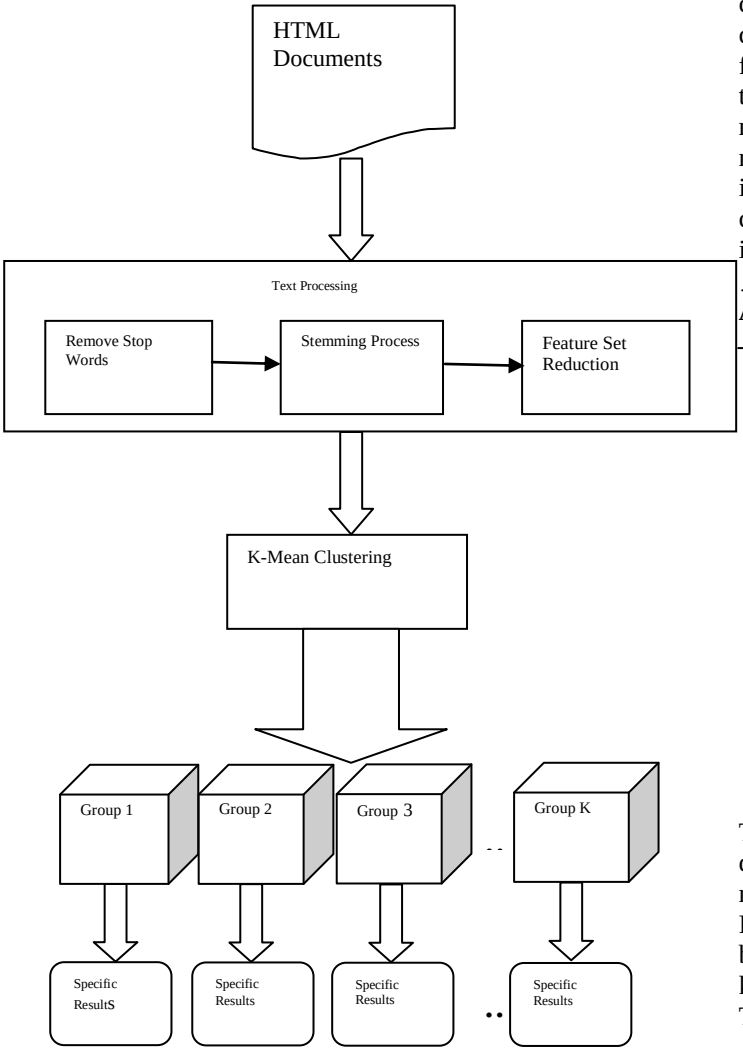

Fig. 4 Framework for HTML Document clustering

Algorithm1: This algorithm obtain to remove stop words & special characters

Input:A document Data Base D and List of Stop words L
D={d1,d2,d3,….,dk} ;where 1<=k<=i
tij is the jth term in ith document
Output: All valid stem text term in D
for (all di in D) do
for (1 to j) do
Extract tij from di
If(tij in list L)
Remove tij from di
End for
End for

D. Dimension reduction:

High dimension is the greatest challenge of document clustering, so dimension reduction became major issue of clustering. This module performs two functions- indexing and feature selection. In indexing method we assign the value to the terms in the documents. After indexing, feature selection method is applied. Feature selection is the process of removing indiscriminate terms from the documents to improve the document clustering accuracy and reduce the computational complexity. The algorithm 2 is proposed for indexing and Feature selection.

Algorithm 2: Basic algorithm obtain for feature selection

Input :A document DataSet D and y minimum threshold value,N is the counter
D={d1,d2,d3,….,dk} ;where 1<=k<=i
tij is the jth term in ith document
Output: Documents Dataset D after feature selection
for (all di in D) do
for (1 to j) do
Count total occurrence of tij in document di
Assign the total occurrence of tij in N
If (N<y)
Remove tji from the document di
End for
End for

The algorithm 2 removes all low frequency terms from the documents. This improves clustering effectiveness and reduces the computational complexity.
For the K-mean algorithm we have to decide value of K when beginning algorithm starts, it is noticed that different value of k will cause different levels of accuracy of the grouping[5]. The basic steps of K-mean clustering algorithm are:

Algorithm3:
Input: Number of cluster K.
Preprocessed Dataset.
Start
{
Take k samples from total number of N randomly as the centroid of each cluster.
Now Calculate the D of the remaining N-k sample to each centroid, and assign them to the cluster with the nearest centroid.
}
After each assignment, again calculate the centroid of the attainment cluster.
Now go to step 2 until find no new assignment.
Stop

IV . EXPERIMENTAL RESULTS

In this paper the unstructured datasets are used. The datasets are downloaded from UCI KDD Archive [4]. This is an online repository of large datasets
with wide variety of data types. This repository has twenty newsgroups dataset for text analysis. This data set consists of 20000 messages taken from Usenet newsgroup. The subset of twenty newsgroups is mini newsgroup. We have done our experiments on 20 newsgroup datasets. Each category contains 1000 documents, so there are 20000 documents for experiments.
The five categories Computer Hardware,Computer Graphics,Medical,Sports and Automobile are used in first experiment.
We performed our experiments on five newsgroups- Computer graphics, Computer hardware, Automobile, Sports and Medical. In this research the 80% dataset are used as training dataset and 20% dataset are used as test dataset.

Table I: Clustering Results

Text based(T)	75.792
Link based(L)	76.892
Hybrid(H)	79.2
Partition(P)	79.92
Hierarchical(HRAR)	79.982
Graph based(GB)	84.0592
Neural Network(NN)	84.192
Fuzzy(F)	84.92
Probabilistic(PRO)	85.29

The Table I shows the percentage of document clustering results generated from the Document Clustering Approaches T to PRO shown in table I. The results in Table I shows that the clustering accuracy is increased from T to PRO.The graphical representation of the results are shown in Fig.5 to Fig.8. We performed our experiments on five newsgroups- Computer graphics, Computer hardware, Automobile, Sports and Medical. In this research the 80% dataset are used as training dataset and 20% dataset are used as test dataset.

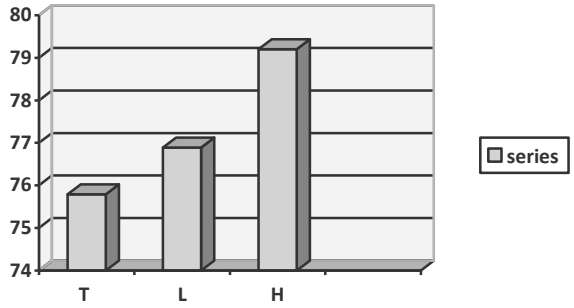

Fig.5: Clustering Results of T to H

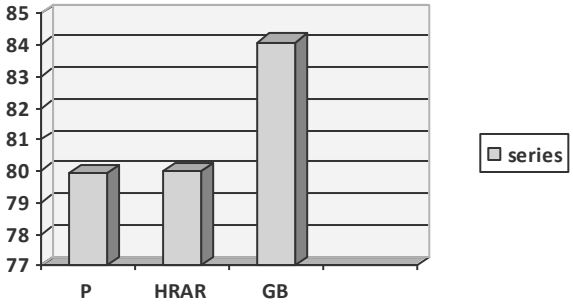

Fig 6: Clustering Results P of to GB

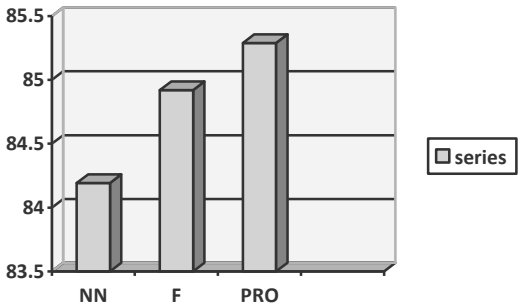

Fig.7 Clustering Results of NN to PRO

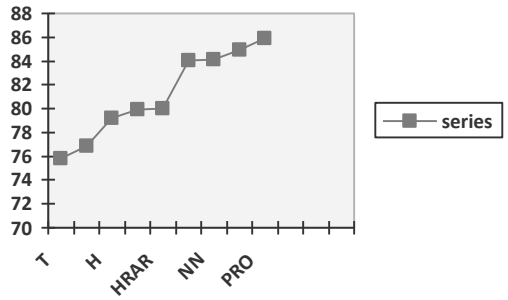

Fig.8: Clustering Results

Table II: Comparison of the clustering algorithms[1]

Name	Input	Output	Type of clusters	Overlap
Single linkage	Similarity matrix	Assign documents to clusters, dendrogram	Few, long, ellipsoidal loosely bound, chaining effect	Crisp clusters
Group Average	Similarity matrix	Assign documents to clusters, dendrogram	Intermediate in tightness between single and complete linkage	Crisp clusters
Complete linkage	Similarity matrix	Assign documents to clusters, dendrogram	Small, tightly bound	Crisp clusters
Ward's Method	Similarity matrix	Assign documents to clusters, dendrogram	Homogeneous clusters, symmetric hierarchy	Crisp clusters
Centroid/ Median HAC	Similarity matrix	Assign documents to clusters	----------	Crisp clusters
K-means	K, iter Feature vector matrix	Assign documents to clusters, refinement of initial clusters	Arbitrary sizes	Crisp clusters

V . CONCLUSIONS

This paper analyzed Document Clustering. The experimental results of Document Clustering Approaches on dataset shows . By this analysis we can easily understand the various conditions responsible for the various Clustering used . This analysis also shows that this method works efficiently, for large text data.

ACKNOWLEDGMENT

This work is supported by research grant from MPCST, Bhopal M.P., India under Grants in Aid Scheme 2011-12 Endt.No. 2427/CST/R&D/2011dated 22/09/2011.

REFERENCES

[1] [1] N.Oikonomakou,M. Vazirgiannis A Review of web Document clustering approaches.
[2] [2] Neeraj Sahu & G.S.Thakur "Hesitant Distance Similarity Measures for Document Clustering" IEEE 2011
[3] [3] Nicholas O.Andrews and Edward A.Fox Recent developments in document clustering 2007.
[4] [4] www.kdd.ics.uci.edu
[5] [5] Neeraj sahu, R. S. Thakur,G. S. Thakur, D. S. Rajput"Analysis of Social Networking sites using K-Mean clustering algorithm" 2012
[6] [6] Neeraj sahu, R. S. Thakur,G. S. Thakur, D. S. Rajput"Clustering Based Classification and Analysis of Data"2012

