International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.3, 2015

REVIEW ARTICLE

ISSN: 2321-7758

A SURVEY OF TEXT DOCUMENTS CLUSTERING APPROACHES

SHRUTI MADAN[1], RAKESH GARG [2] , SUMAN [3]

[1]Student, M.Tech., Computer Science, Hindu College of Engg. Sonipat

[2,3]Assistant Professor, Computer Science, Hindu College of Engg. Sonipat

Article Received:04/05/2015 | Article Revised on:11/05/2015 | Article Accepted on:13/05/2015

SHRUTI MADAN

ABSTRACT

Clustering text documents into different category groups is an important step in indexing, retrieval, management and mining of abundant text data on the Web or in corporate information systems. It is a technique of grouping a set of objects into subsets or clusters. The goal is to create the clusters that are coherent internally, but different from each other. In Text Document Clustering, grouping of text documents occurs based upon their content. Document clustering is a used in unsupervised document text data mining, organization, automatic topic extraction, and information retrieval. There are many fast and high-quality document clustering algorithms available which play an important role in effectively organizing the information. The documents that are clustered can be abstracts of research papers, web news articles, etc. In this paper we are going to discuss about various techniques of clustering that are used in text mining.

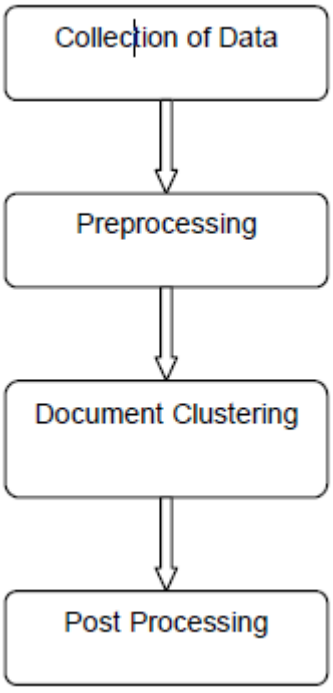
Keywords: Data mining, Text mining, Document Clustering, Optimization, K-Means, Fcm

INTRODUCTION

With the advancement of technologies in World Wide Web, huge amounts of rich and dynamic information's are available. With web search engines, a user can quickly browse and locate the documents [16]. Usually search engines returns many documents, a lot of which are relevant to the topic and some may contain irrelevant documents with poor quality. Cluster analysis or clustering plays an important role in organizing such massive amount of documents returned by search engines into meaningful clusters. Data mining is the process of discovering the pattern by searching large stores of data. Data mining uses some mathematical algorithms to divide the data and measure the probability of future events. For example, data mining software can help the companies to find customers having common interests. It is also called Knowledge Discovery in Data (KDD).The KDD process have following stages i.e. Selection, Preprocessing, Transformation, Data Mining, Interpretation. We can find the need of data mining in various fields like Retail Industry, Telecommunication, Biology, Medicare and Healthcare, Finance, Banking, Education, Science etc. Text mining is a type of data mining. It is also referred as text data mining. Text mining is the process of deriving high-quality of information from text. In Text Mining, patterns are extracted from natural language text rather than from the databases. Text mining helps an organization to derive valuable information from the content in the form a of text such as word

documents, emails and postings on the social networking websites like Facebook, LinkedIn.

Clustering: This is a data mining technique used to group a set of objects into clusters, with the purpose of low intra-cluster distances and high inter-cluster distances. Application of clustering algorithms include Wireless Sensor Network, Academics, Search Engine, identifying the cancerous data set, Text Mining, knowledge discovery etc. In business, it can help marketers to categorize their customers based upon their purchasing behavior and find their target customer group. In case of biology, it is used to extract plant and animal taxonomies categorize genes with identical functionality. It can be used to help cluster related documents on the Web. It is an unsupervised learning (unlike classification) where class labels are not provided in advance, and in sometimes clustering can be done in a semi-supervised way background knowledge is available.



Collection of Data includes the processes like crawling, indexing, filtering etc which are used to collect the documents that need to be clustered, index them to store and retrieve in a better way, and filter them to remove the extra data, for example, stop words.

Pre-processing consists of steps that take as input a plain text document and output a set of tokens (which can be single terms or n-grams) to be included in the vector model.

Document Clustering: Clustering of documents is used to group documents into relevant topics. The major difficulty in document clustering is its high dimension.

Post processing includes the major applications in which the document clustering is used, for example, the application that uses the results of clustering for recommending news articles to the users.

Literature Survey:

Michael Steinbach performed a work," A Comparison of Document Clustering Techniques" [1]. In this paper the results of some common document clustering techniques are shown. In this paper, the author compares the two main approaches to document clustering, i.e Hierarchical clustering and K-means. (For K-means, Author used a K-means algorithm and a variant of K-means i.e Bisecting-K-means.) Hierarchical clustering is often considered as the better quality clustering approach, but is limited because of its time complexity.

Alan F. Smeaton performed a work," An Architecture for Efficient Document Clustering and Retrieval on a Dynamic Collection of Newspaper Texts" [2]. The Author explained a technique for clustering a collection of documents such as a collection of online newspapers which uses a number of short-cuts to make the process computable for large collections. Furthermore, proposed design is extensible in that it caters for a dynamic collection of documents which would be periodically, can be nightly, updated, amended or have deletions. An archive of the the Irish Times newspaper's implementation is reported here.

Text Clustering via Particle Swarm Optimization [3].In this paper PSOVW (particle swarm optimizer for variable weighting) is used to handle the problem of text clustering, called Text Clustering via Particle Swarm Optimization (TCPSO). PSOVW has been exploited for evolving optimal feature weights for clusters and has demonstrated to improve the clustering quality of high-dimensional data.

Mihai Surdeanu performed a work," A Hybrid Unsupervised Approach for Document Clustering"[4] Author proposes a hybrid, unsupervised document clustering approach that combines a hierarchical clustering algorithm with Expectation Maximization.

Anna Huang performed a work," Similarity Measures for Text Document Clustering"[5] Clustering is a useful technique that organizes a large quantity of unordered text documents into a small number of meaningful and organized clusters. Proposed experiments utilize the standard K means algorithm and Author report results on seven text document datasets and five distance/similarity measures that have been most commonly used in text clustering.

Clustering Approaches

As stated by Han and Kamber [10] clustering algorithms can be categorized as follows:

Partitioning Algorithm: A partitioning method partitions a dataset of m objects into clusters (k<=m). The K-Means and K-Medoids methods are good known partitioning algorithms. The K-Means approach is a centroid technique where the similar cluster is evaluated with respect to the average value of the objects in a cluster (i.e. each cluster is act as the center of the cluster). Strength of the K-Means is that it is relatively good with a complexity O (tkm), where t denotes iteration, k denotes clusters, and m denotes objects. It terminates at a local optimal solution.

Hierarchical Algorithm: Unlike partitioning method where clusters are defined in advance, that is not required in hierarchical clustering methods. A tree-view of clusters is provides by these algorithms and are also called dendograms. These methods can be classified as follows:

1. Agglomerative (bottom up approach): Agglomerative approach initiated with each of the item in its own cluster, and then, in a bottom-up manner, combines the two closest groups repeatedly to form a new cluster.
2. Divisive (top down approach): Iteratively split the cluster. It starts with all items at a time in one cluster and sub-divides them into smaller pieces.
3. Some more useful clustering approach produced as a result of integration of hierarchical and distance-based algorithms are: Wei Xu [6], Ye-Hang Zhu [7] and Andreas Hotho [8] are a hierarchical clustering algorithm for categorical data.

Document Clustering

Clustering of documents is used to group documents into relevant topics. The major difficulty in document clustering is its high dimension. It requires efficient algorithms which can solve this high dimensional clustering [16]. A document clustering is a major topic in information retrieval area .Example includes search engines. The basic steps used in document clustering process are shown in figure given below.

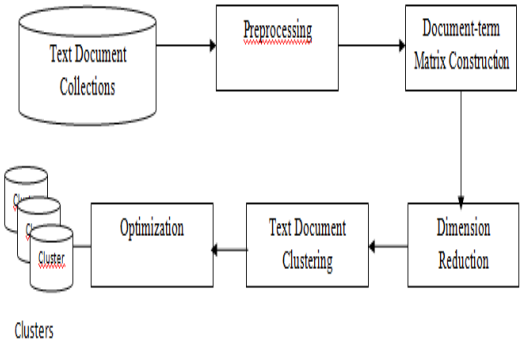


Fig.1 Flow diagram for representing basic Steps in text clustering

Document Clustering Procedure:

The document clustering process consists of the following steps [9]

Pre-processing: The documents to be clustered are in an unstructured format therefore before the actual clustering begins, some pre-processing steps need to be performed. The pre-processing includes Tokenization, Stemming of words in the document, and Stop word removal.

- *Tokenization* means breaking the text up into words, phrases, symbols, or other meaningful elements called tokens where each token refers to a word in the document.
- *Stemming* involves conversion of various forms of a word to the base word. E.g. 'computing' and 'computed' will be stemmed to the base word 'compute'. Similarly 'sarcastically' is stemmed to the word 'sarcasm'.
- *Stop word removal:* Stop words are the words present in documents which do not contribute in differentiating a collection of documents. Hence, they are removed from the documents. These are basically prepositions, articles, helping verbs, pronouns etc.

Techniques:

K- Means Approach: This technique is a popular traditional partitioning clustering technique [11]. It

is one of the easy approaches used for resolving known clustering issues. For document clustering problem, this approach allocates each of the document to one of the K number of clusters. An efficient cluster here will be a sphere where centroid is considered to be its centre of gravity. The aim of this approach is to reduce the approximate mean coverage of document data set corresponding to their cluster midpoint; when centre of the group can be used as the mean of the document set in a group. The centroid $\mu 1$ of the document set in a cluster ω is calculated as mentioned below:

$$\mu 1\ (\omega) = \frac{1}{|\omega|} \sum_{x \in \omega} x$$

K-means algorithm is fast, robust and easy to understand. Gives best result when data set are distinct or well separated from each other but it is difficult to predict the value of k and fails for non linear data set.

FCM Algorithm: FCM is a popular soft clustering approach that combines features of K Means and Fuzzy technique. [14].It resembles K means except it includes membership matrix that contains the degree of membership of a data points to all clusters. This approach divides data in *k* groups through coverage measurement between data (x_i) & the centroid of group (v_b) of the size of vector L (l=1...L). The coverage method for relevance count between document x_a and centroid v_b is usually taken as the Euclidean Distance function (d (x_i,v_j)). FCM minimizes the following function:

$$J_{FCM} = \sum_{b=1}^{k} \sum_{a=1}^{N} \mu_{ab}^{l}\, d(x_a, b_b)^2,$$

$$l \in (1, \infty), \quad \forall x \sum_{=1}^{k} \mu_{ab} = 1$$

Centroid of a group is the average of all positions calculated by their degree of belonging to group:

$$centrej = \frac{\sum_i \mu_{ij}^{m}\ i}{\sum_i \mu_{ij}^{m}}$$

Degree of belonging ≡ inverse when the coverage to the group centers:

$$\mu_{ij} = \frac{1}{d\left(center_j, i\right)}$$

m, A real parameter whose value is greater than 1 makes the coefficient normalized and fuzzified so that their sum is 1.

$$\mu_{ij} = \frac{1}{\sum_k \left(\frac{d(center_j, i)}{d(center_k, i)}\right)^{2/(m-1)}}$$

When l is near to 1, then centre of group nearest to the position is given more weight than any other and the approach behaves same as k means.

The plus point is that the data points belong to more than one cluster and membership is assigned to each cluster but it takes more time to compute.

Bi-Section-*k*-means: It is one of the fastest text clustering algorithms & deals with the large size of the textual data. [12] Bi-Section-*k*-means is a fast and high-quality clustering algorithm for the text documents. It is based on the *k*-means algorithm. It splits the largest cluster repeatedly using *k*-means until the desired number of cluster is obtained.

This algorithm selects and bisects each one of the leaf clusters iteratively until singleton clusters are reached and singleton cluster is meaningless and it takes a lot of time.

EM algorithm: It is an iterative algorithm. It is similar to k-means algorithm in which the centroid is recomputed until the desired coverage value is achieved [13] .This can be used to cluster the continuous data and estimate the destiny function.

The cost per iteration is low as compared to others. It is easily implemented and requires small storage space. The main drawback is that in some cases it is very slow to converge.

K-medoid algorithm: K-medoid algorithm and k-means algorithm are quite similar. The main difference is that in k-medoid algorithm, to represent the cluster, an object closet the center of the cluster is used but in k-means algorithm centroid is used to represent the cluster. [15]

TEXT MINING APPLICATIONS

The main Text Mining applications [17] are most often used in the following sectors:

- Publishing and media.
- Telecommunications, energy and other services industries.
- Information technology sector and Internet.
- Banks, insurance and financial markets.
- Political institutions, political analysts, public administration and legal documents.
- Pharmaceutical and research companies and healthcare.

CONCLUSION:

In this paper we discussed about the various text document clustering techniques and we have concluded that the existing K-Means clustering

algorithm for the Document clustering is much better than other algorithms. K-Means is fast, robust and easy to use. This algorithm produces the good quality clusters and better results. In Future, we can optimize the results after applying the clustering techniques. We can use the hybridized approach for our work so that even better clusters can be produced.

REFERENCES

[1] Michael Steinbach, George Karypis & Vipin Kumar" A Comparison of Document Clustering Techniques".

[2] Alan F. Smeaton " An Architecture for Efficient Document Clustering and Retrieval on a Dynamic Collection of Newspaper Texts", New Zealand, 2008

[3] Khaled Hammouda" Text Clustering via Particle Swarm Optimization"

[4] Mihai Surdeanu," A Hybrid Unsupervised Approach for Document Clustering".

[5] Anna Huang," Similarity Measures for Text Document Clustering".

[6] Wei Xu," Document Clustering Based On Non-negative Matrix Factorization"

[7] Ye-Hang Zhu," Document Clustering Method Based on Frequent Co-occurring Words".

[8] Andreas Hotho," Wordnet improves Text Document Clustering".Dingding Wang," Integrating Clustering and Multi-Document Summarization to Improve Document Understanding"

[10] Han & Kamber,"Spatial Clustering Methods in Data Mining: A survey".

[11] Gupta& S. Lehal, "A Survey of Text Mining Techniques and Applications"

[12] Hotho& Nurnberger, "A Brief Survey of Text Mining"

[13] Sumit Vashishtha," Efficient Retrieval of Text for Biomedical Domain using Expectation Maximization Algorithm"

[14] Sumit Goswami, "A Fuzzy Based Approach To Text Mining And Document Clustering".

[15] Fasheng Liu & Lu Xiong, "Survey on Text Clustering Algorithm".

[16] R.Jensi and Dr.G.Wiselin Jiji " A survey on optimization approaches to text document clustering" International Journal on Computational Sciences & Applications (IJCSA) Vol.3, No.6, December 2013.

[17] Vishal Gupta and Gurpreet S. Lehal "A Survey of Text Mining Techniques and Applications" Journal of Emerging Technologies in web intelligence, vol. 1, no. 1, august 2009