Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 6, June 2017, pg.158 – 167

AUTOMATIC TEXT SUMMARIZATION

Ruchika Aggarwal, Latika Gupta

Aryabhatta College, University of Delhi, New Delhi, India

{Ruchika.aggarwal1989, Latikagup}@gmail.com

Abstract- Content summarization is an old challenge however the modern research look into courses occupies towards rising patterns in biomedicine, item audit, instruction areas, messages and web journals. This is because of the way that there is data over-burden in these zones, particularly on the World Wide Web. Automated summarization is an imperative zone in NLP (Natural Language Processing) research. It comprises of consequently making a summary of at least one or more texts. The motivation behind extractive report summarization is to consequently choose various demonstrative sentences, entries, or passages from the original document. Text summarization approaches in light of SSHLDA, Vector Space Model and Modified K-Means and Cluster have, to an extent, prevailing with regards to making a powerful summarization of a document. Both extractive and abstractive techniques have been inquired about. Most summarization procedures depend on extractive strategies. Abstractive strategy is like summaries made by people. Abstractive summarization starting at now requires overwhelming machinery for language generation and is hard to reproduce into the domain particular territories.

Index Terms- Text Summarization, Clustering, SSHLDA, Support Vector Machine, Modified K-Means.

I. INTRODUCTION:

1.1 What is Clustering?

Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabelled data.

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense

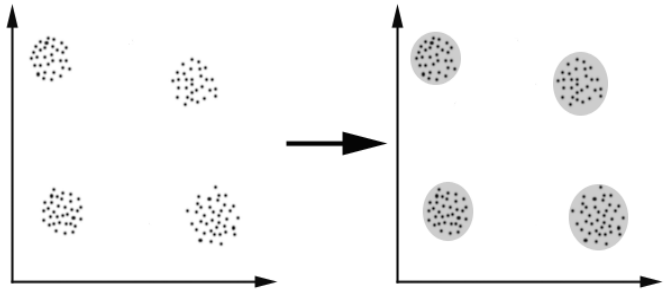
areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is often necessary to modify data pre-processing and model parameters until the result achieves the desired properties.

Cluster analysis was originated in anthropology by Driver and Kroeber in 1932 and introduced to psychology by Zubin in 1938 and Robert Tryon in 1939[1][2] and famously used by Cattell beginning in 1943[3] for trait theory classification in personality psychology.

OR
A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way".

A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

We can show this with a simple graphical example:



In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is *distance*: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called *distance-based clustering.*
Another kind of clustering is *conceptual clustering*: two or more objects belong to the same cluster if this one defines a concept *common* to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

A "clustering" is essentially a set of such clusters, usually containing all objects in the data set. Additionally, it may specify the relationship of the clusters to each other, for example, a hierarchy of clusters embedded in each other. Clusterings can be roughly distinguished as:

Hard clustering: each object belongs to a cluster or not

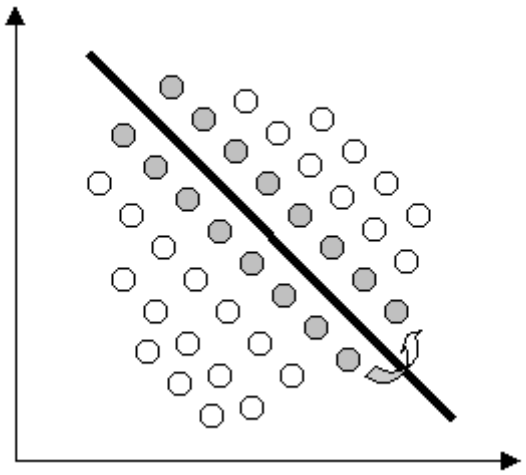
Soft clustering (also: fuzzy clustering)**:** each object belongs to each cluster to a certain degree (for example, a likelihood of belonging to the cluster)

1.2 CLUSTERING ALGORITHMS

Classification
clustering algorithms may be classified as listed below:

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering
- Probabilistic Clustering

In the first case data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster. A simple example of that is shown in the figure below, where the separation of points is achieved by a straight line on a bi-dimensional plane.
On the contrary the second type, the overlapping clustering, uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value.



Instead, a hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted.

Finally, the last kind of clustering use a completely probabilistic approach.

Clustering algorithms can be categorized based on their cluster model, as listed above. The following overview will only list the most prominent examples of clustering algorithms, as there are possibly over 100 published clustering algorithms. Not all provide models for their clusters and can thus not easily be categorized.

Four of the most used clustering algorithms:

- K-means
- Fuzzy C-means
- Hierarchical clustering
- Mixture of Gaussians

1.3 K-MEANS CLUSTERING

The Algorithm

K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycentre of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 ,$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centres.

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

2. Assign each object to the group that has the closest centroid.

3. When all objects have been assigned, recalculate the positions of the K centroids.

4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centres. The k-means algorithm can be run multiple times to reduce this effect. K-means is a simple algorithm that has been adapted to many problem domains.

1.4 MODIFIED K-MEANS:

This paper presents a data clustering approach using modified K-Means algorithm based on the improvement of the sensitivity of initial center of clusters. This algorithm partitions the whole space into different segments and calculates the frequency of data point in each segment. The segment which shows maximum frequency of data point will have the maximum probability to contain the centroid of cluster. The number of cluster's centroid (k) will be provided by the user in the same manner like the traditional K-mean algorithm and the number of division will be k*k (`k' vertically as well as `k' horizontally). If the highest frequency of data point is same in different segments and the upper bound of segment crosses the threshold `k' then merging of different segments

become mandatory and then take the highest k segment for calculating the initial centroid of clusters. In this paper we also define a threshold distance for each cluster's centroid to compare the distance between data point and cluster's centroid with this threshold distance through which we can minimize the computational effort during calculation of distance between data point and cluster's centroid. It is shown that how the modified k-mean algorithm will decrease the complexity & the effort of numerical calculation, maintaining the easiness of implementing the k-mean algorithm. It assigns the data point to their appropriate class or cluster more effectively.

We have presented a modified k-means algorithm which eliminates the problem of generation of empty clusters (with some exceptions). Here, the basic structure of the original k-means is preserved along with all its necessary characteristics. A new center vector computation strategy enables us to redefine the clustering process and to reach our goal. The modified algorithm is found to work very satisfactorily, with some conditional exceptions which are very rare in practice.

1.4.1 MODIFIED APPROACH K-MEAN ALGORITHM

The K-mean algorithm is a popular clustering algorithm and has its application in data mining, image segmentation, bioinformatics and many other fields. This algorithm works well with small datasets. In this paper we proposed an algorithm that works well with large datasets. Modified k-mean algorithm avoids getting into locally optimal solution in some degree, and reduces the adoption of cluster -error criterion.

Algorithm: Modified approach (S, k), S={x1,x2,…,xn }

Input: The number of clusters k1 (k1> k) and a dataset containing n objects(Xij+).

Output: A set of k clusters (Cij) that minimize the Cluster - error criterion.

Algorithm

1. Compute the distance between each data point and all other data- points in the set D

2. Find the closest pair of data points from the set D and form a data-point set Am (1<= p <= k+1) which contains these two data- points, Delete these two data points from the set D

3. Find the data point in D that is closest to the data point set Ap, Add it to Ap and delete it from D

4. Repeat step 4 until the number of data points in Am reaches (n/k)

5. If p<k+1, then p = p+1, find another pair of data points from D between which the distance is the shortest, form another data-point set Ap and delete them from D, Go to step 4.

1.5 SEMI-SUPERVISED HIERARCHICAL LATENT DIRICHLET ALLOCATION

Topic models, such as latent Dirichlet allocation (LDA), are useful NLP tools for the statistical analysis of document collections and other discrete data. Furthermore, hierarchical topic modeling is able to obtain the relations between topics — parent-child and sibling relations. Unsupervised hierarchical topic modeling is able to detect automatically new topics in the data space, such as hierarchical Latent Dirichlet Allocation (hLDA). hLDA makes use of nested Dirichlet Process to automatically obtain a L-level hierarchy of topics. Modern Web documents, however, are not merely collections of words. They are usually documents with hierarchical labels – such as Web pages and their placement in hierarchical directories. Unsupervised hierarchical topic modelling cannot make use of any information from hierarchical labels, thus supervised hierarchical topic models, such as hierarchical Labeled Latent Dirichlet Allocation (hLLDA), are proposed to tackle this problem. hLLDA uses hierarchical labels to automatically build corresponding topic for each label, but it cannot find new latent topics in the data space, only depending on hierarchy of labels.

Supervised hierarchical topic modelling and unsupervised hierarchical topic modeling are usually used to obtain hierarchical topics, such as hLLDA and hLDA. Supervised hierarchical topic modeling makes heavy use of the information from observed hierarchical labels, but cannot explore new topics; while unsupervised hierarchical topic modeling is able to detect automatically new topics in the data space, but does not make use of any information from hierarchical labels. In this paper, we propose a semi-supervised hierarchical topic model which aims to explore new topics automatically in the data space while incorporating the information from observed hierarchical labels into the modeling process, called **Semi-Supervised Hierarchical Latent Dirichlet Allocation (SSHLDA)**. We demonstrate the effectiveness of the proposed model on large, real-world datasets in the question answering and website category domains on two tasks: the topic modeling of documents, and the use of the generated topics for document clustering. Our results show that our joint, semi-hierarchical model outperforms the state-of-the-art supervised and un- supervised hierarchical algorithms. The contributions of this paper are threefold: (1) We propose a joint, generative semi-supervised hierarchical topic model, i.e. Semi-Supervised Hierarchical Latent Dirichlet Allocation (SSHLDA), to overcome the defects of hLDA and hLLDA while combining the their merits. SSHLDA is able to not only explore new latent topics in the data space, but also makes use of the information from the hierarchy of observed labels.

1.6 VECTOR SPACE MODEL

The vector-space models for information retrieval are just one subclass of retrieval techniques that have been studied in recent years. The taxonomy provided in labels the class of techniques that resemble vector-space models ``formal, feature-based, individual, partial match'' retrieval techniques since they typically rely on an underlying, formal mathematical model for retrieval, model the documents as sets of terms that can be individually weighted and manipulated, perform queries by comparing the representation of the query to the representation of each document in the space, and can retrieve documents that don't necessarily contain one of the search terms. Although the vector-space techniques share common characteristics with other techniques in the information retrieval hierarchy, they all share a core set of similarities that justify their own class.

Vector-space models that don't attempt to collapse the dimensions of the space treat each term independently, essentially mimicking an inverted index. However, vector-space models are more flexible than inverted indices since each term can be individually weighted, allowing that term to become more or less important within a document or the entire document collection as a whole. Also, by applying different similarity measures to compare queries to terms and documents, properties of the document collection can be emphasized or deemphasized. For example, the dot product (or, inner product) similarity measure finds the Euclidean distance between the query and a term or document in the space. The cosine similarity measure, on the other hand, by computing the angle between the query and a term or document rather than the distance, deemphasizes the lengths of the vectors. In some cases, the directions of the vectors are a more reliable indication of the semantic similarities of the objects than the distance between the objects in the term-document space.

Vector-space models were developed to eliminate many of the problems associated with exact, lexical matching techniques. In particular, since words often have multiple meanings (polysemy), it is difficult for a lexical matching technique to differentiate between two documents that share a given word, but use it differently, without understanding the context in which the word was used. Also, since there are many ways to describe a given concept (synonomy), related documents may not use the same terminology to describe their shared concepts. A query using the terminology of one document will not retrieve the other related documents. In the worst case, a query using terminology different than that used by related documents in the collection may not retrieve any documents using lexical matching, even though the collection contains related documents.

Vector-space models, by placing terms, documents, and queries in a term-document space and computing similarities between the queries and the terms or documents, allow the results of a query to be ranked according to the similarity measure used. Unlike lexical matching techniques that provide no ranking or a very crude ranking scheme (for example, ranking one document before another document because it contains more occurrences of the search terms), the vector-space models, by basing their rankings on the Euclidean distance or the angle measure between the query and terms or documents in the space, are able to automatically guide the

user to documents that might be more conceptually similar and of greater use than other documents. Also, by representing terms and documents in the same space, vector-space models often provide an elegant method of implementing relevance feedback. Relevance feedback, by allowing documents as well as terms to form the query, and using the terms in those documents to supplement the query, increases the length and precision of the query, helping the user to more accurately specify what he or she desires from the search.

1.7 PRE-PROCESSING

Before the original text could be used for the GA, it is needed to adapt the entry of the original text to the format of the GA. In this step, the original text is separated in sentences. Also, the text is pre-processed with the well-known Porter Stemmer [18] in order to find related words. Since the proposed method is based on the frequency of the words as a measure of its relevance (section 4.3), this does not take into account the frequency of stop words because it is higher than meaningful words.

II. RELATED WORK:

Shubankar, K. [01], In this paper we introduce a novel and efficient approach to detect topics in a large corpus of research papers. With rapidly growing size of academic literature, the problem of topic detection has become a very challenging task. We present a unique approach that uses closed frequent keyword-set to form topics. Our approach also provides a natural method to cluster the research papers into hierarchical, overlapping clusters using topic as similarity measure. To rank the research papers in the topic cluster, we devise a modified Page Rank algorithm that assigns an authoritative score to each research paper by considering the sub-graph in which the research paper appears. We test our algorithms on the DBLP dataset and experimentally show that our algorithms are fast, effective and scalable.

Yonghui Wu, Yuxin Ding, Xiaolong Wang, Jun Xu [02], Topic model is an increasing useful tool to analyze the semantic level meanings and capture the topical features. However, there is few research about the comparative study of the topic models. In this paper, we describe our comparative study of three topic models in the extrinsic application of topic clustering. The topic model distance is defined on the converged parameters of topic models, which is used in the topic clustering. Then, the topic models are compared using the clustering result of the corresponding topic distance matrix. A series of comparative experiments are carried on a corpus containing 5033 web news from 30 topics using the cosine distance as the base-line. Web page collections with different number of topics and documents are used in experiments. The experiment results show that topic clustering using topic distance achieves a better precision and recall in the data set containing related topics. The topic clustering using topic distance benefits from the topic features captured by topic models. The complex topic model does provide further help than the simple topic model in topic clustering.

Wongkot Sriurai [03], Most text categorization algorithms represent a document collection as a Bag of Words (BOW).The BOW representation is unable to recognize synonyms from a given term set and unable to recognize semantic relationships between terms. In this paper, we apply the topic-model approach to cluster the words into a set of topics. Words assigned into the same topic are semantically related. Our main goal is to compare between the feature processing techniques of BOW and the topic model. We also apply and compare between two feature selection techniques: Information Gain (IG) and Chi Squared (CHI). Three text categorization algorithms: Naïve Bayes (NB), Support Vector Machines (SVM) and Decision tree, are used for evaluation. The experimental results showed that the topic-model approach for representing the documents yielded the best performance based on F1 measure equal to 79% under the SVM algorithm with the IG feature selection technique.

Susumu Harada, Shashikant Khandelwal [04], We implemented a topic extraction system that takes as input a collection of postings from a Usenet newsgroup and outputs a list of prominent topics that characterize the contents of the newsgroup, and for each topic, gives the set of threads that discuss the topic along with their relevance measure with respect to the topic. We use two methods, chi-square feature extraction and Latent Semantic Analysis, to extract the topic terms.

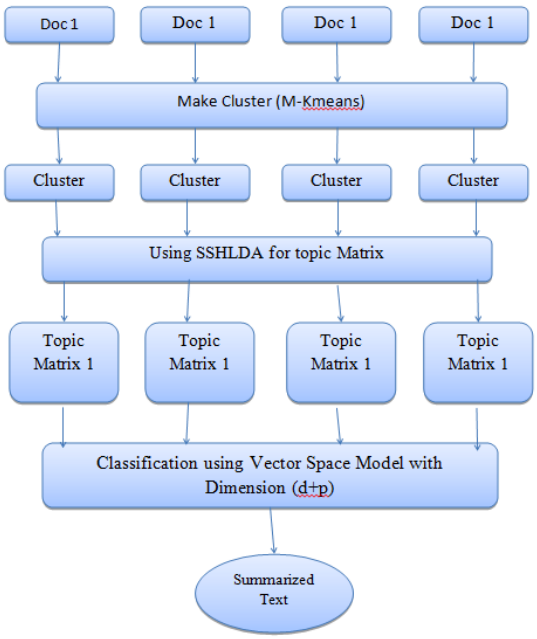
Mohamad Alkhouja [05], Latent Dirichlet Allocation (LDA) is a popular topic modelling technique for exploring document collections. Because of the increasing prevalence of large datasets, there is a need to improve the scalability of inference for LDA. In this paper, we introduce a novel and flexible large scale topic modeling package in MapReduce (Mr. LDA). As opposed to other techniques which use Gibbs sampling, our proposed framework uses variational inference, which easily fits into a distributed environment. More importantly, this variational implementation, unlike highly tuned and specialized implementations based on Gibbs sampling, is easily extensible. We demonstrate two extensions of the models possible with this scalable framework: informed priors to guide topic discovery and extracting topics from a multilingual corpus. We compare the scalability of Mr. LDA against Mahout, an existing large scale topic modeling package. Mr. LDA out-performs Mahout both in execution speed and held-out likelihood.

Yonghui Wu, Yuxin Ding, Xiaolong Wang [06], Topic model is an increasing useful tool to analyze the semantic level meanings and capture the topical features. However, there is few research about the comparative study of the topic models. In this paper, we describe our comparative study of three topic models in the extrinsic application of topic clustering. The topic model distance is defined on the converged parameters of topic models, which is used in the topic clustering. Then, the topic models are compared using the clustering result of the corresponding topic distance matrix. A series of comparative experiments are carried on a corpus containing 5033 web news from 30 topics using the cosine distance as the base-line. Web page collections with different number of topics and documents are used in experiments. The experiment results show that topic clustering using topic distance achieves a better precision and recall in the data set containing related topics. The topic clustering using topic distance benefits from the topic features captured by topic models. The complex topic model does provide further help than the simple topic model in topic clustering.

III. PROPOSED METHODOLOGY:

In our research work we use three main techniques to propose method for text summarization SSHLDA and Modified k means Clustering and vector space model**.** Previously Supervised hierarchical topic modeling makes heavy use of the information from observed hierarchical labels, but cannot explore new topics; while unsupervised hierarchical topic modeling is able to detect automatically new topics in the data space, but does not make use of any information from hierarchical labels. In this synopsis , we will use a semi-supervised hierarchical topic model which aims to explore new topics automatically in the data space while incorporating the information from observed hierarchical labels into the modeling process, called *Semi- Supervised Hierarchical Latent Dirichlet Allocation (SSHLDA).* SSHLDA can automatically explore latent topic in data space, and extend the existing hierarchy of observed topics. SSHLDA makes use of not only observed topics, but also latent topics.

To make cluster of data we use modified k means clustering algorithm because one of the major problems of the k-means algorithm is that it may produce empty clusters depending on initial center vectors. For static execution of the k-means, this problem is considered insignificant and can be solved by executing the algorithm for a number of times. In situations, where the k-means is used as an integral part of some higher amount of data, this empty cluster problem may produce anomalous behavior of the system and may lead to significant performance degradation. This proposed work uses a modified version of the k-means algorithm that efficiently eliminates this empty cluster problem. We will prove that the proposed algorithm is semantically equivalent to the original k-means and there is no performance degradation due to incorporated modification.



IV. Result Analysis / Implementation

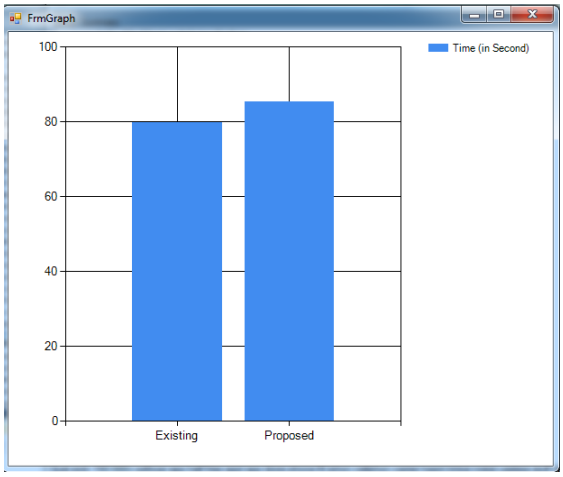


Fig – Comparison Graph between Proposed and Existing methodology

V. CONCLUSION:

We use three main techniques to propose method for text summarization Modified K-Means Clustering, SSHLDA and Vector Space Model. We have introduced a semi-supervised hierarchical topic models, i.e. SSHLDA, which aims to solve the drawbacks of hLDA and hLLDA while combine their merits. Specially, SSHLDA incorporates the information of labels into generative process of topic modeling while exploring latent topics in data space. In addition, we have also proved that hLDA and hLLDA are special cases of SSHLDA. We have conducted experiments on the Legal Case Dataset, and assessed the performance in terms of Recall and Precision measure. The experimental results show that the prediction ability of Hybrid Proposed Technique is the best, and Proposed Technique can also achieve significant improvement over the baselines on Recall and Precision measure.

REFERENCES

[1]. Shubankar, K., A.P.Singh, Pudi V., "**A frequent keyword-set based algorithm for topic modeling and clustering of research papers",** Data Mining and Optimization (DMO), 2011 3rd Conference on Date of Conference: 28-29 June 2011, ISSN : 2155-6938, E-ISBN : 978-1-61284-212-7, Print ISBN: 978-1-61284-211-0, INSPEC Accession Number: 12171919.

[2]. Yonghui Wu, Yuxin Ding, Xiaolong Wang, Jun Xu, **" A comparative study of topic models for topic clustering of Chinese web news",** Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on (Volume:5), Date of Conference: 9-11 July 2010, Page(s): 236 – 240, Print ISBN: 978-1-4244-5537-9, INSPEC Accession Number: 11520534, Conference Location : Chengdu, DOI: 10.1109/ICCSIT.2010.5564723, Publisher: IEEE.

[3]. Wongkot Sriurai, **"Improving Text Categorization By Using A Topic Model",** Advanced Computing: An International Journal (ACIJ), Vol.2, No.6, November 2011, DOI: 10.5121/acij.2011.2603.

[4]. Susumu Harada, Shashikant Khandelwal,"**Automatic Topic Extraction and Classification of Usenet Threads".**

[5]. Ke Zhai, Jordan Boyd-Graber, Nima Asadi, Mohamad Alkhouja, **"Mr. LDA: A Flexible Large Scale Topic Modeling Package using Variational Inference in MapReduce",** WWW 2012 – Session: Information Extraction April 16–20, 2012, Lyon, France.

[6]. Yonghui Wu, Yuxin Ding, Xiaolong Wang, Jun Xu, **"A comparative study of topic models for topic clustering of Chinese web news",** CONFERENCE PAPER · AUGUST 2010 DOI: 10.1109/ICCSIT.2010.5564723 · Source: IEEE Xplore.

[7]. N K Nagwani,**"Summarizing large text collection using topic modeling and clustering based on MapReduce framework",** Nagwani Journal of Big Data (2015) 2:6 DOI 10.1186/s40537-015-0020-5.

[8]. DAVID M. BLEI, JOHN D. LAFFERTY, "**TOPIC MODELS".**

