Web Documents Clustering Approach Base on Improvise Fuzzy Clustering using Cosine Similarity and Name Entity Recognition Method

[1]Kalyani R Pole, [2]Vishakha R Mote

[1]M.E. Student CSE, [2]Assistant Professor [CSE/IT] Dept Computer Science Engineering,
[1,2]P. E. S College of Engineering Dr. Babasaheb Amedkar University, Aurangabad, India

Abstract—Recent advances in computers and technology have resulted in a growing body of documents. The need is to classify the set of documents according to type. Placing related documents together is convenient for making decisions. Researchers conducting interdisciplinary research acquire repositories on different topics. The classification of the repositories according to the theme is a real need to analyze the research work. The experiments are tested on different sets of real and artificial data, such as NEWS 20, Reuters, emails, research on different topics. The term frequency inverse document frequency algorithm is used together with the fuzzy hierarchical algorithm and K-means. Initially, the experiment is being carried out in small data sets and cluster analysis was performed. The best algorithm applies to the extended data set. Together with the different groups of related documents, the resulting coefficient and the trend of measure F are presented to show the behavior of the algorithm for each data set. Our model combines two components: a mixing component used to discover latent groups in the collection of documents and a theme model component used to mine multigrain issues, including cluster-specific local issues and global topics shared between clusters. We use the variational inference to approximate the posterior part of the hidden variables and learn the parameters of the model. The experiments in two data sets demonstrate the effectiveness of our model.

IndexTerms— collaborative filtering and information filtering; Web content; linguistic topological space; Name Entity Recognition, Natural Language Processing

---

I. INTRODUCTION

Data mining (DM)[1][2][3] analyzes observational (often large) data sets to find unsuspected relationships and summarizes the data in novel ways that are understandable and useful to the owner of the data. The field of data mining and knowledge discovery is emerging as a new and fundamental area of research with important applications for science, engineering, medicine, business and education. DM attempts to process classify and formulate basic induction processes that help extract significant information and knowledge from unstructured data. In today's knowledge-based economy, DM is an essential tool for achieving greater productivity, decreased uncertainty, delighted customers, mitigated risk, maximized returns, refined processes and optimally allocated resources. DM can be used in financial applications, such as business, banking and marketing to obtain significant advantages in today's competitive global market.

Grouping is an important technique for extracting data to extract useful information from several large data sets [7]. Grouping is a process of grouping a set of objects into groups so that the objects are quite similar in the same cluster but very different compared to the objects in other clusters. Several types of grouping methods have been proposed and developed or can be defined as a mathematical technique designed to reveal the classification structures in the data collected in real-world phenomena. Grouping methods organize a clustered data set so that the data points in a cluster are similar and the data points in other clusters are different.

The grouping of documents and the modeling of themes are highly correlated and can benefit each other. On the one hand, subject models can discover the latent semantics embedded in the corpus of the document and semantic information can be much more useful for identifying groups of documents than the characteristics of the raw term. In classical document grouping approaches, documents are usually represented with a word bag model (BOW) that is purely based on raw terms and is insufficient to capture all semantics. Theme models can put words with similar semantics in the same group called subject, where synonymous words are treated as the same. In the subject models, the corpus of the document is projected in a thematic space that reduces the noise of the similarity measure and the grouping structure of the corpus can be identified more effectively. On the other hand, the grouping of documents can facilitate the modeling of topics. Specifically, the grouping of documents allows us to extract specific local topics for each group of documents and global issues shared between clusters. In a collection, documents usually belong to several groups.

The models of standard themes (Blei et al., 2003, Hofmann, 2001) lack the mechanism to model the grouping behavior between documents, so they can only extract a single set of flat themes where local issues and global issues they mix and cannot be distinguished. Naively, we can perform these two tasks separately. So that the modeling of subjects facilitates the creation of

clusters, we can first use theme models to project documents in a thematic space, then perform grouping algorithms such as K-means in the thematic space to obtain clusters. To make clusters promote topic modeling, we can first obtain clusters using standard clustering algorithms, then create topic models to extract cluster-specific local topics and cluster-independent global topics by incorporating cluster tags into the cluster design. Cluster model. However, this ingenuous strategy ignores the fact that the grouping of documents and the modeling of themes are highly correlated and follow an egg-chicken relationship. Better grouping results produce better subject models and better subject models in turn contribute to better grouping results. Doing them separately does not cause them to promote each other to achieve the best overall performance. In this document, we compare a generative model that integrates the grouping of documents and the modeling of themes based on the grouping of Kmeans and the improvised group Fuzzy. Given a corpus, we assume that there are several latent groups and each document belongs to a latent group. Each group has a set of local themes that capture the specific semantics of the documents in this group and a previous Dirichlet that expresses preferences about local issues.

Propose system use the Reuters-21578 data set is a test collection for evaluation of automatic text categorization techniques. Although it is widely used in many research studies, few have reported the details of how it is used. In Reuter data set around 21578 documents are present it is freely available on net which is a xml document use for clustering techniques.

In addition, we assume that there is a set of global themes shared by all groups to capture the common semantics of the entire collection and a common Dirichlet that governs the sampling of proportional vectors in global themes for all documents. Each document is a mixture of local problems and global problems. The words of a document can be generated from a global theme or a local theme of the group to which the document belongs. In our model, the latent variables of the cluster membership, the distribution of documents and issues and the problems are deduced together. The grouping and the modeling combine perfectly and promote each other.

There are two main types in the application of k-means algorithms in the cluster analysis: "hard" non-fuzzy or fuzzy. In the first type, the number of conglomerates K must be determined in advance as an input to these algorithms. In a real data set, k is generally unknown. In practice, different K-values are tested, and cluster validation techniques are used to measure the cluster results and determine the best value of K. The K-means algorithm is a classical technique, and there are many descriptions and variations available. In addition, it is popular because it is conceptually simple and computationally fast and efficient from the point of view of memory. However, several limitations in the k-means algorithm make extraction difficult. Observations of K-mean groups in k groups, where k is provided as an input parameter. K-means clustering starts with a single cluster with its center as the average of the data. This cluster is divided into two and the means of the new clusters are trained iteratively. These groups are divided again, and the process continues until the specified number of clusters is obtained.

In grouping algorithms, points are grouped by some notion of "closeness" or "similarity". In k-means, the default measure of closeness is the Euclidean distance. The idea of fuzzy grouping was first introduced as an alternative to traditional cluster analysis by applying membership values to points among Ruspini's groups. The c-means fuzzy grouping approach is also known as Fuzzy k-means. It is analogous to traditional cluster analysis. Fuzzy c-means developed by Bezdek in 1981 adapted the theory of fuzzy sets that assigns a data object (observation) to more than one group. The essential difference between the diffuse grouping of c-means and the grouping of standard k-means is the partition of objects in each group. Instead of the hard partitioning of the standard k-means grouping, where the objects belong to a single cluster, the diffuse c-means group considers that each object is a member of each cluster, with a variable degree of membership.

II. Related Work

The grouping of documents [1][7][17] is a widely studied problem . with many applications such as document organization, navigation, summary, classification. See (Aggarwal and Zhai, 2012) for a general description. Popular grouping methods such as K-means and spectral grouping [2][3] in the general grouping literature are widely used to group documents. Specific to the domain of the text, a popular paradigm of grouping methods is based on matrix factorization, which includes latent semantic indexing (LSI) [4][5], the factorization of non-negative matrices [8] and Factorization of concepts [10]. The basic idea of methods based on factoring is to transform the documents of the original temporal space into a latent space. The transformation can reduce the dimensionality of the data, reduce the noise of the measure of similarity and magnify the semantic effects in the underlying data [11], which are beneficial for the grouping. Researchers have applied thematic models to the cluster documents. [12] [13] investigated the grouping performance of PLSA and LDA. They use LDA and PLSA to model the corpus and each topic is treated as a cluster. The documents are grouped by examining the topic proportion vector θ. A document is assigned to the cluster x if x = arg maxj θj.

Subject models [3] are probabilistic [3, 4] generative models that are initially created to model texts and identify the latent semantics underlying the collection of documents. The subject models postpone the collection of documents that show multiple latent semantic themes where each subject is represented as a multinomial distribution over a given vocabulary and each document is a mixture of hidden themes. In the domain of vision, thematic models [7] [17] are also widely used for image modeling. Several models have been designed to jointly model data and its category labels or cluster labels. [7] proposed a Bayesian hierarchical model to jointly model the images and their categories. Each category has an LDA model with specific Dirichlet categories and themes. In your problem, category labels are observed. In this document, we are interested in the unsupervised cluster where the cluster label is unknown. Wallach [14] proposed a cluster-based theme model (CTM) that introduces latent variables in LDA to model groups and each group has a specific Dirichlet group that governs sampling of the distribution of documents and topics.

Each document is associated with a group indicator and its theme proportion vector is generated from the specific Dirichlet prior to that group. (Zhu et al., 2010) proposed a similar model used for the classification of scenes in artificial vision. They associate each group with a normal previous logistic instead of a previous Dirichlet. However, in both models, all groups share a single set of topics. They lack the mechanism to identify specific local problems of each cluster and global problems shared by all the clusters. Another problem is that topics intrinsically belonging to group A can be used to generate documents in group B, which is problematic. For example, when modeling scientific articles, it is not reasonable to use a "computer architecture" theme in a computer science group to generate an economics document. The models proposed in [14][17] cannot prohibit this problem since the topics are shared between the groups. Eventually, the inferred topics will be less coherent and will not be sufficiently discriminative to differentiate the clusters.

The idea of using detailed themes belonging to several sets instead of flat themes from a single set to model documents is exploited in[2][12] . In [6] author represents each document as a combination of a background distribution on common words, a mix distribution on general topics and a distribution on words that are considered specific to that document. [12] Proposed a multi-grain theme model for online review modeling. They use local themes to capture evaluable aspects and use global themes to capture the properties of the reviewed articles. [2] Proposed a multiple vision theme model for the analysis of ideological perspectives. Each ideology has a set of specific ideological themes and a specific distribution of ideology over words. All documents share a set of issues independent of ideology. In your problem, you see the ideology label for each document.

The extraction of data and the discovery of knowledge is a family of computational methods that aim to collect and analyze data related to the function of a system of interest to obtain a better understanding of the system. Data mining analyzes massive observation data sets to find unsuspected relationships and summarizes the data in novel ways that are understandable and useful to the owner of the data and refers to extracting or "extracting" knowledge from large amounts of data. Data mining has its origins in various disciplines, of which the two most important are statistics and machine learning.

Grouping is a process of grouping a set of physical or abstract objects into a set of classes, called groupings, according to some function of similarity. The cluster is a collection of objects that are similar to each other within the cluster and dissimilar to objects in other clusters. There are different types of grouping paradigms, such as the representative, hierarchical and density-based grouping based on graphs and spectra, according to the data and the desired characteristics of the cluster.

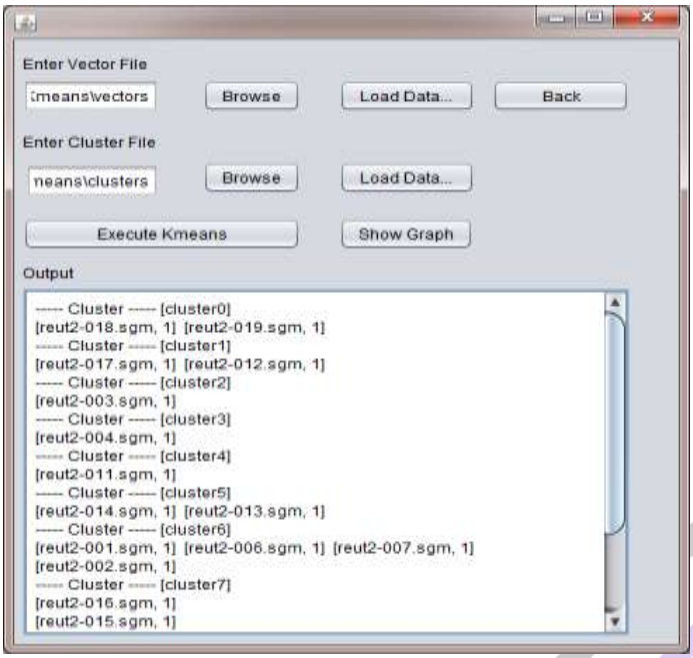
III. EXISTING K-MEANS CLUSTERING ALGORITHM

The k-means is one of the simplest unsupervised learning algorithms for grouping problems. The algorithm aims to form k groups of n objects, resulting in intra-clusters. The k-means algorithm is a simple and iterative grouping algorithm that divides a given data set into a series of clusters specified by the user, k. One of the main advantages of this algorithm is that it is simple to implement and execute, relatively fast, easy to adapt and common in practice. The k-means is an efficient centroid-based algorithm that has been widely used in several key areas, such as microarray data sets, large data sets, etc. Two terms, a group and a distance must be defined:

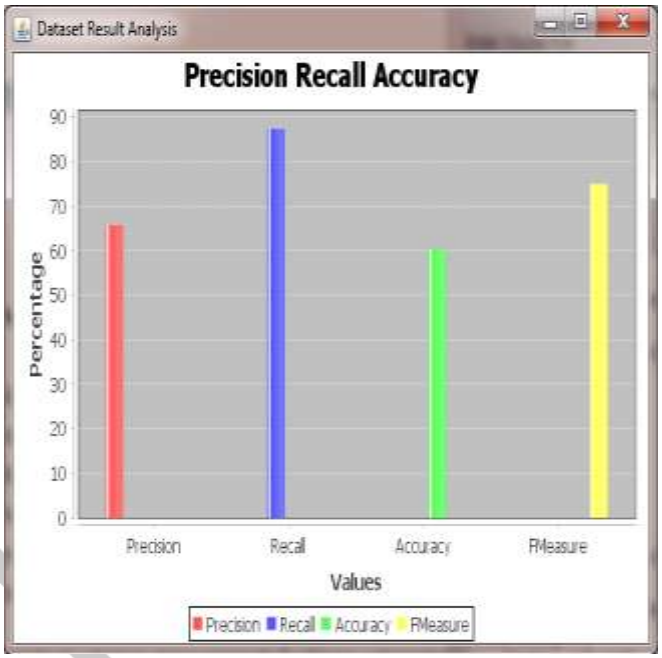
Algorithm for K-Means

The k-means clustering algorithm
Input: D:{d1,d2….dn}\\set of n items
K //Number of desired clusters
Output: A set of k-clusters.
Steps: 1. Arbitrarily choose k-data items from D as initial centroid;
2. Repeat assigns each item di to the cluster which has the closest centroid, Calculate new mean for each cluster; until convergence criteria are met.

A cluster is an ordered list of objects that have common characteristics. The objects belong to an interval [a, b], in our case [0, 1]. The distance between two clusters involves some or all of the elements of the two clusters. The grouping method determines how the distance should be calculated. The k-means clustering technique begins with a description of the basic algorithm. Choosing k initial centroid is the first step, where k is a parameter specified by the user, that is, the number of desired clusters. The second step is to assign each point to the nearest centroid, and each collection of points assigned to a centroid is a cluster. The centroid of each group is updated based on the points assigned to the cluster. The assignment steps are repeated and updated until there are no point change clusters, or equivalently, until the centroid remain the same.



Screen Shots 1: Executing K-means on Reuters 21578 Dataset



Screen Shots 2: K-means Graphical Results on Reuters 21578 Dataset

Table 1: PRECISION, RECALL, OVERALL F-MEASURE (%) using K-means Clustering ON "REUTERS-21578, DISTRIBUTION 1" DATASET.

K	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Precision	67.47	66.88	66.80	66.80	66.80	66.88	67.22	67.47	65.71
Recall	87.68	87.34	87.34	87.34	87.34	87.68	87.68	87.68	87.34
Fmeasure	75.00	75.11	75.11	76.11	75.11	75.11	75.11	75.11	75.00
Accuracy	62.17	60.33	60.33	60.33	60.33	60.33	62.17	62.17	60.17

From above overall result of K-means clustering we observe that In K-Means Difficult to predict the number of clusters (K---Value) and initial seeds have a strong impact on the final results. However results generated using this algorithm are mainly dependent on choosing initial cluster centroid Measurable and efficient in large data collection Disadvantages of k-means algorithm.

IV. NAME ENTITY RECOGNISION AND NATURAL LANGUAGE PROCESSING

Proposed system use a Natural Language Processing (NLP) for the Automatic Classification of documents. By classifying text, we aim to assign a document or piece of text to one or more classes or categories making it easier to manage or sort. A Document Classifier often returns or assigns a category "label" or "code" to a document or piece of text. Depending on the Classification Algorithm or strategy used, a classifier might also provide a confidence measure to indicate how confident it is that the result is correct.

Stanford NER is a Java implementation of a Named Entity Recognizer. Named Entity Recognition (NER) labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names. It comes with well-engineered feature extractors for Named Entity Recognition, and many options for defining feature extractors.NER categorized words are generated to cluster the data based on different categorized topic like in given example topic the hole topic contents are organize in different categories such as PERSON, LOCATION, PLACE, ORGANIZATION, DATE etc.

V. TERM FREQUENCY AND INVERSE DOCUMENT FREQUENCY

Fuzzy clustering is use Term Frequency (TF)**,** which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length as a way of normalization.

Inverse Document Frequency (IDF), which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones.

Tf-idf stands for term frequency-inverse document frequency, and the Tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Tf-idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1]. The name derives from the term "direction cosine": in this case, note that unit vectors are maximally "similar" if they're parallel and maximally "dissimilar" if they're orthogonal (perpendicular). This is analogous to the cosine, which is unity (maximum value) when the segments subtend a zero angle and zero (uncorrelated) when the segments are perpendicular.

$$\text{Eq.1 is} \ldots\ldots\ldots\ldots\ldots\ldots \text{Similarity} = \cos(\theta) = \frac{A.B}{||A||2\ ||B||2}$$

Propose use the cosine similarity for calculating similarity score between documents. If the similarity score between documents is '1' then it will display result as a similar document and if the similarity score between document is '0' then it will display dissimilar document.

VI. Proposed Architecture

The proposed work is motivated by a research paper [6]. Based on the findings and the literature review, the following key issues and challenges are addressed to improve the traditional text-gathering technique.

1. The length of textual records is not similar; therefore, evaluation of individual textual content requires a significant amount of computing resources.
2. Extracting functionality from different documents is different in nature and length, so that measuring the similarity of a data object to another object is a complex task.
3. Cluster formation of documents must select some centroid for accurate group formation, but the random and fluctuating Centroid selection in text documents can increase the process time and clustering accuracy.
4. The approximation of similarity in text extraction must compare the text document with their important characteristics, but directional information on similarity is still calculated to optimize clustering performance.

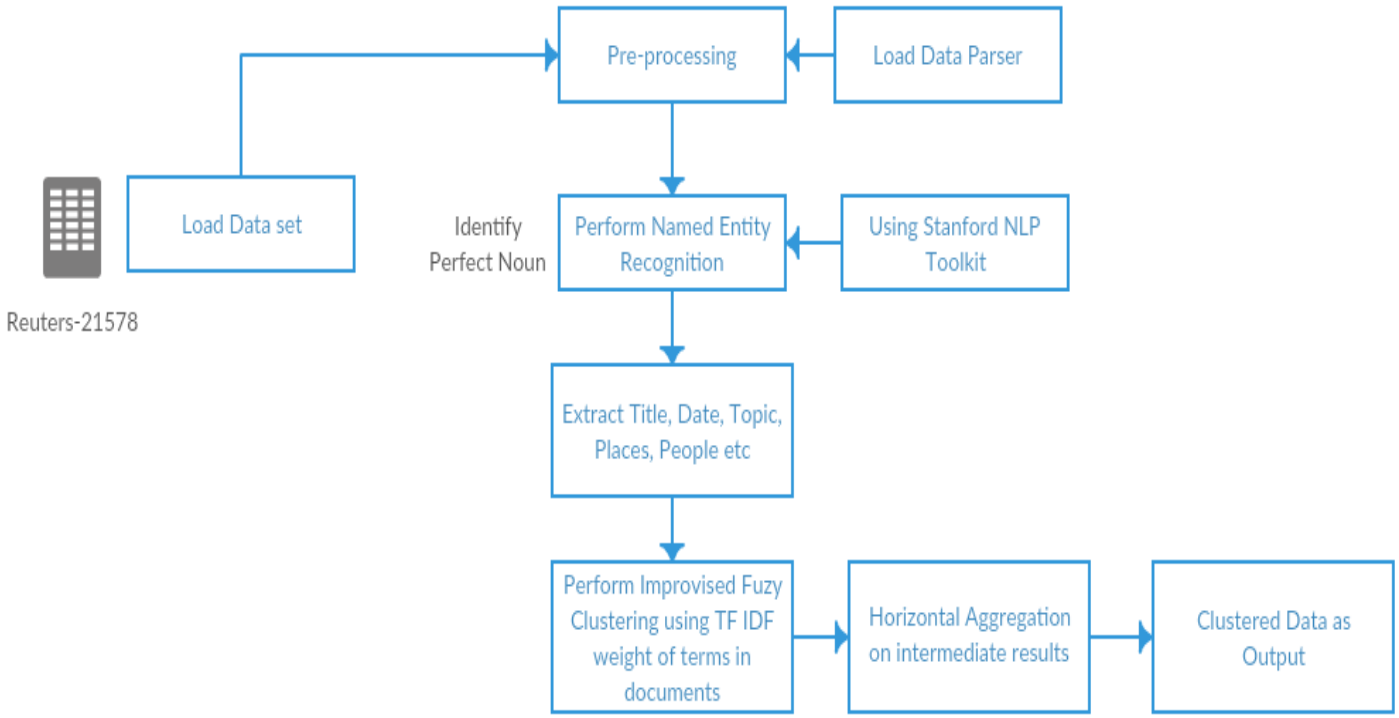


Figure1. Proposed Architecture

In order to solve the obtained issues and challenges for document clustering the following solution is proposed for further investigation and design.

1. Design of an strong pre-processing technique for refining the noisy contents form the documents learning set.
2. Design a new feature extraction and selection technique for optimizing the performance of document content analysis and their comparisons.
3. Enhance the traditional fuzzy c-means in order to achieve higher accuracy over the text content analysis and their clustering.
4. Implement the modifications on the fuzzy c-means clustering to demonstrate the hierarchical relationship among the documents.

Psuedo code for proposed module:

1. Start
2. Load data set Reuters 21578
3. Perform data parsing and segregate data attributes.
4. Identify Perfect Noun, i.e. Name, Place, People, Country etc using Stanford NLP Toolkit.
5. Perform Named Entity Recognition to create a level 0 simplex with all named entities.
6. Calculate Term frequency / Inverse Document Frequency of document along with weight score.
7. Using TF-IDF score, perform clustering .

Require: V = {x1, x2, • • • , xn} be the vertex set of all
reserved named entities generated fromW associated with
their categories _ in a collection of documents.
Ensure: H is the hierarchy of connected components.
Let S = {{x1}, {x2}, • • • , {xn}} be the set of all 0-
simplexes initially.
TF - IDF Given two weights a and b
Let k ← 0.
while k ⩽ n do
Let Si and Sj be two n-simplexes in S.
while delta(Si, Sj) ⩾ a and sigma(Si, Sj) ⩽ b do
S′ ← Si ∪ Sj .
Add S′ to S
end while
k ← (k + 1)
end while

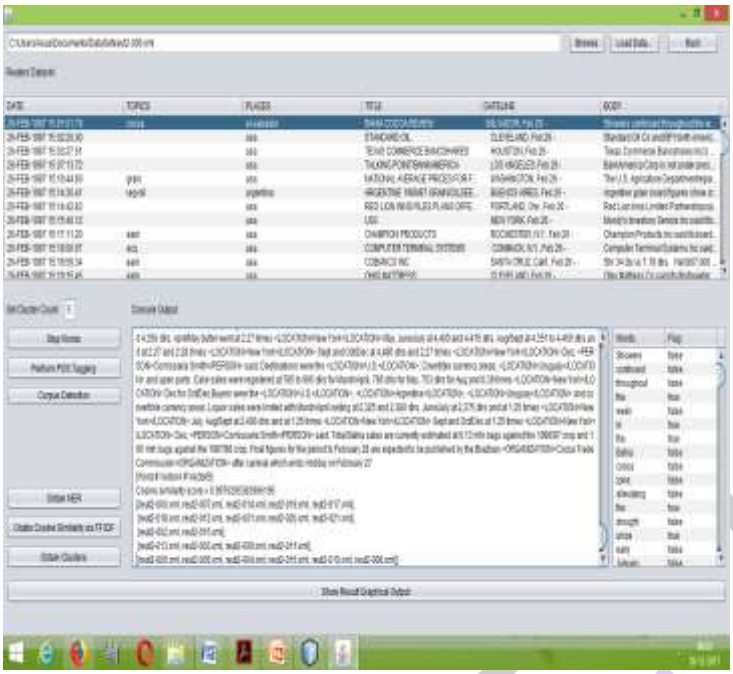
8. Perform horizontal aggregation on results.
9. Collect output and calculate Precision, Recall and F measure.
10. The input from the matrix is the generated score. We calculate the smallest and biggest scores. We calculate exactly five ranges starting from smallest the value and ending with the largest value. Now assign the score to score calculated in master matrix step and check the score in these five ranges. Once, the scores have been calculated a threshold of say '2' is set. The file having threshold more than two is added to the cluster and discards the file which fails to satisfy the condition.
11. END

VII. Experimental Results

Experimental evaluation of document pooling approaches usually measures its overall effectiveness instead of adding up to its efficiency. In other words, measuring the ability of an approach is to make a correct categorization. To identify and discriminate the correct topics in a collection of documents, the combinations of characteristics and their co-occurring relationships are the clue, and the possibilities of showing how meaningful it will be. All document features compose a topologically probabilistic space, more specifically a simplified complex associated with probabilistic measures to denote the underlying structure. The complex can be geographically decomposed into inseparable components at various levels (at various levels of skeletons) that each component corresponds adequately to the themes of a collection of documents. Of course, subjects that an induced component are topologically distinguishable or perfectly included in other induced topics.

Table 2: Contingency Table

Topic zi (Expert Judgment)	**Clustering Results**	
	Yes	***No***
Yes	TP	FN
No	FP	TN



Screen Shot 3: Executing Fuzzy Clustering on Reuters 21578 Dataset

Screen Shot 4: Fuzzy Clustering Graphical on Reuter 21578 Dataset

Table 3: PRECISION, RECALL, OVERALL F-MEASURE (%) using Improvised Fuzzy Clustering(IFCNER) ON "REUTERS-21578, DISTRIBUTION 1" DATASET.

K	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Precision	96.61	97.91	96.15	92.85	97.43	97.56	95.34	97.56	96.96
Recall	91.93	92.15	86.20	92.86	86.26	86.95	87.23	88.88	82.05
Accuracy	94.21	94.94	90.90	91.66	91.56	91.95	91.11	93.02	88.88
Fmeasure	90.27	92.30	86.48	92.85	89.06	89.55	88.40	90.90	86.66

Key Index Parameters for Result Classification

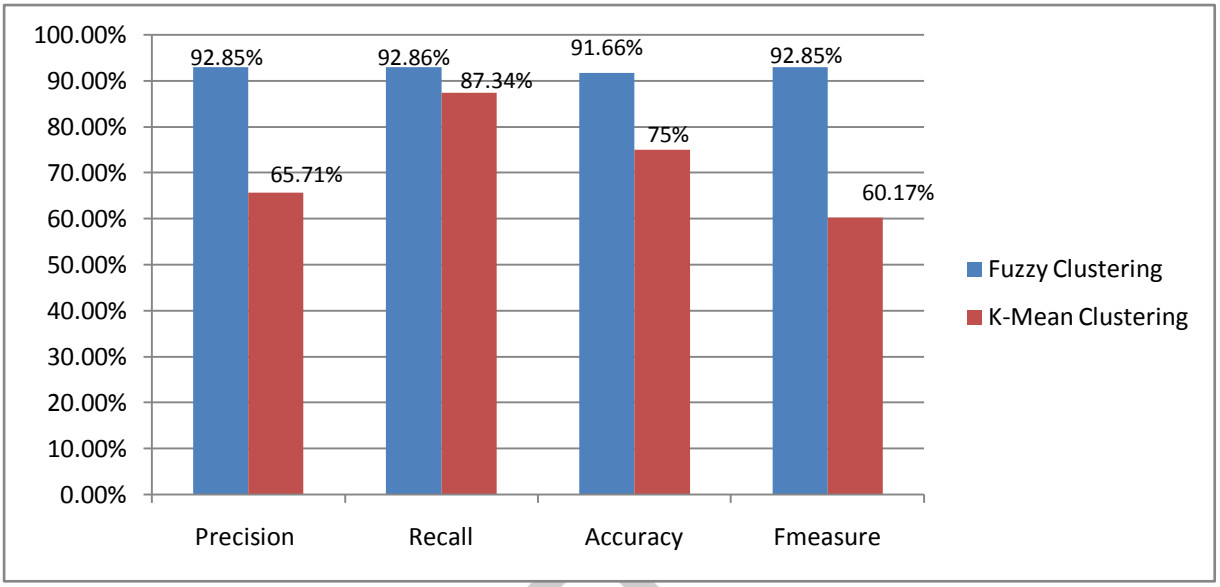
In information clustering with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also called sensitivity) is the fraction of the relevant instances that are retrieved. Precision and recall are therefore based on understanding and measuring relevance. In simple terms, high accuracy means that an algorithm returns significantly more relevant than irrelevant results, while a high recall means that an algorithm has yielded the most relevant results.

The most important category measurements for binary categories are:

Table 4: Calculating Methods

Precision	Recall	F Measure
$P = TP/(TP + FP)$	$R = TP/(TP + FN)$	$tp + tn/tp + tn + fp + fn$

F1 measure used in this paper was obtained when _ is set to be 1, which means precision and recall are equally weighted for evaluating the performance of clustering. Because many categories would be generated and need to be compared, the overall precision and recall were calculated as the average of all precisions and recalls belonging to categories, respectively. F1 is calculated as the mean of individual results. It is a macro average of the categories.



Graph 1: Comparison Graph for Fuzzy Clustering and K-Means Clustering

In our experimental results performed on document clustering, we intend to show that through topic and data clustering the overall performance of our method can be improved. Precision score 92.85% is good as compare to k means clustering that is 65.71%.The overall accuracy of proposed method is 91.66% as compare to existing once is 60.17%. That means fuzzy clustering give most semantic result as compare to existing system.

In the experiments, we have also demonstrated that clustering can help infer more coherent topics and can differentiate topics into group-specific ones and group-independent ones.

VIII FUTURE SCOPE

Fuzzy clustering for web document is one of the best documents clustering method used in medical documents as such document as it includes many acronyms of clinical data. Web document are very complex and difficult to identify particular document from large set of documents to overcome this problem fuzzy clustering is used to identify the particular document and make its cluster which is having similar set of documents. It is also used in social media, trend analysis, market analysis, banking sector and so forth To Identify Similar Pages in Web Applications This approach is based on a process that first computes the dissimilarity between Web pages using latent semantic indexing, and then group's similar pages using clustering algorithms. Fuzzy clustering having a good scope in various application as web technology is growing new web search engines are developed to retrieve as accurate data as possible.

IX CONCLUSION

We propose a improvised clustering topic model to simultaneously perform document clustering. Experiments on Reuters datasets demonstrate the fact that the tasks of mining and clustering are closely related and can mutually promote each other. In our experimental results performed on document clustering, we intend to show that through topic and data clustering the overall performance of our method can be improved. In the experiments, we have also demonstrated that clustering can help infer more coherent topics and can differentiate topics into group-specific ones and group-independent ones.

After successfully implementation of the proposed technique of document clustering approach the following outcomes are expected.

1. An improved approach of fuzzy c-means clustering for making accurate document clustering using weighted technique
2. A comparative performance study with fuzzy clustering with K means clustering and strength evaluation of the proposed methodology
3. A new technique for document domain identification with less resource consumption (running time) as compared to traditional document clustering approach.

REFERENCES

[1] Charu C Aggarwal and ChengXiang Zhai. A survey of text clustering algorithms. Mining Text Data, pages 77–128, 2012.

[2] Amr Ahmed and Eric P Xing. Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1140–1150. Association for Computational Linguistics, 2010.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. the Journal of Machine Learning Research, 3:993–1022, 2003.

[4] Jonathan Boyd-Graber, Jordan Chang, Sean Gerrish, Chong Wang, and David Blei. Reading tea leaves: how humans interpret topic models. In Proceedings of the 23rd Annual Conference on Neural Information Processing Systems, 2009.

[5] Deng Cai, Xiaofei He, and Jiawei Han. Locally consistent concept factorization for document clustering. Knowledge and Data Engineering, IEEE Transactions on, 23(6):902–913, 2011.

[6] Chaitanya Chemudugunta and Padhraic Smyth Mark Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference, volume 19, page 241. MIT Press, 2007.

[7] Scott Deerwester, Susan T. Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. Journal of the American society for Information Science, 41(6): 391–407, 1990. Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 2, pages 524–531. IEEE, 2005.

[8] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 42(1):177–196, 2001.

[9] Yue Lu, Qiaozhu Mei, and ChengXiang Zhai. Investigating task performance of probabilistic topic models: an empirical study of plsa and lda. Information Retrieval, 14(2):178–203, 2011.

[10] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: analysis and an algorithm. Advances in Neural Information Processing Systems, 2: 849–856, 2002.

[11] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(8):888–905, 2000.

[12] Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In Proceedings of the 17th international conference on World Wide Web, pages 111–120. ACM, 2008.

[13] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. Foundations and Trends R in Machine Learning, 1(1-2):1–305, 2008.

[14] Hanna M Wallach. Structured topic models for language. Unpublished doctoral dissertation, Univ. of Cambridge, 2008.

[15] Wei Xu and Yihong Gong. Document clustering by concept factorization. In Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval, pages 202–209. ACM, 2004.

[16] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Informaion Retrieval, pages 267–273. ACM, 2003.

[17] Jun Zhu, Li-Jia Li, Li Fei-Fei, and Eric P Xing. Large margin learning of upstream scene understanding models. Advances in Neural Information Processing Systems, 24, 2010.