Survey Paper on Clustering of Documents Based on Partitioning the Features

Prof. S. R. Durugkar[1], Madhuri Malode[2]

Assistant Professor[1], P.G. student[2]
Department of Computer Engineering
Late G.N.Sapkal College of Engineering, Anjaneri, Nasik [1, 2]
University Of Pune

Abstract :

Finding the appropriate number of clusters to which documents should be partitioned is crucial in document clustering.

In this paper we will focus on various clustering techniques and our proposed system to discover the cluster structure without requiring the number of clusters as input.

Document features or even we can say that the various attributes will be automatically partitioned into two groups, in particular, discriminative words and nondiscriminative words, and contribute differently to document clustering.

There is one variational inference algorithm which we have studied to infer the document collection structure as well as the partition of document words at the same time.

We will justify at the end in the conclusion how our approach will perform well on the data set. Then we will justify our system's accuracy and efficiency by describing what we have proposed with the predicted modules and features for the system.

Keywords— Database management, database applications-text mining, pattern recognition, clustering document clustering, feature partition

I.INTRODUCTION

1.1 Clustering

Clustering can be measured the most important *unsupervised learning* problem;

It contract with finding a *structure* in a collection of unlabeled data.

A general definition of clustering could be "the process of organizing given objects into certain number of 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.

If the query is general, it is extremely difficult to identify the specific document which the user is interested in. The users are forced to sift through a long list of off-topic documents. Moreover, internal relationships among the documents in the search result are rarely presented and are left for the user.

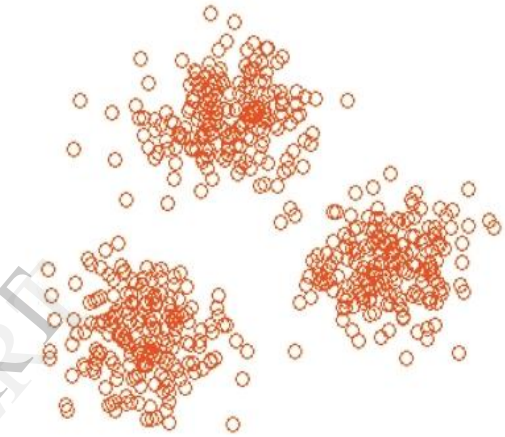


Fig 1.1 Clustering Overview

Search engines or any information retrieval application are an invaluable tool for retrieving information from the Web. In response to a user query, they return a list of results ranked in order of relevance to the query.

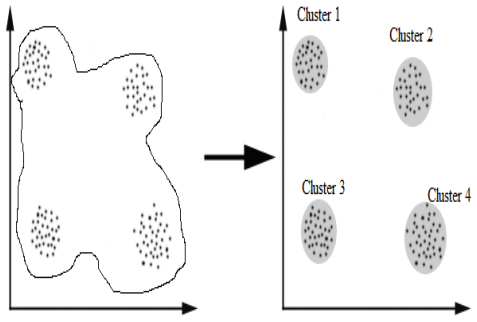


Fig 1.2 Formation of cluster of given data

The objective of clustering is to decrease the amount of data by categorizing or grouping similar data items and present them collectively. Such grouping is pervasive in the way humans process information, and one of the inspirations for using clustering algorithms is to

provide automated tools to help in constructing categories or taxonomies

The user starts at the top of the list and follows it down examining one result at a time, until the sought information has been found. third method is search results clustering, which consists of grouping the results returned by a search engine into a hierarchy of labeled clusters (also called categories). This method combines the best features of query-based and category-based search, Not only has search results clustering attracted considerable commercial interest, but it is also an active research area, with a large number of published papers discussing specific issues and systems. Search results' clustering is clearly related to the field of document clustering but it poses unique challenges concerning both the effectiveness and the efficiency of the underlying algorithms that cannot be addressed by conventional techniques.

1.2 Uses and possible Application Areas of Clustering

- ***Marketing*:** finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;
- ***Biology*:** classification of plants and animals given their features;
- ***Libraries***: book ordering;
- ***Insurance*:** identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
- ***City-planning***: identifying groups of houses according to their house type, value and geographical location;
- ***Earthquake studies***: clustering observed earthquake epicenters to identify dangerous zones;

1.3 Typical requirements or constraint for execution of any clustering algorithm

- Scalability of the objects or the system;
- Dealing with different objects;
- Discovering clusters with arbitrary pattern;
- Minimal requirements for specific system
- Ability to deal with noisy and outlier data;
- High dimensionality;
- Interpretability and usability.

1.4 Document Clustering

Document clustering has been investigated for use in a number of different areas of text mining and information retrieval.

Initially, document clustering was used for improving the precision or recall in information retrieval applications and as an efficient way of finding the nearest neighbors of a document so that system will return the max relevant document in response to user's query. Document clustering has also been used to automatically generate hierarchical clusters of documents

II. SURVEY ON VARIOUS CLUSTERING TECHNIQUES

There are many clustering techniques which are available in the market, and each of them may give a different grouping of objects.

The choice of a particular method will depend on the type of output desired that is it depends on the end user to select one of them as per his requirement and form the desired number of clusters. The known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset.

2.1 Single Pass Clustering Techniques: A very simple partition method, the single pass method creates a partitioned dataset as follows:

- ❖ In this first object will declare as a cluster representative of that cluster.
- ❖ Then subsequent objects after comparing the threshold value will be compared against the Cluster representative.
- ❖ In this way cluster will be formed of given objects.

2.2 Hierarchical Agglomerative Methods

The hierarchical clustering methods are most commonly used. The construction of this classification can be achieved by the following general steps.

1. Find the 2 nearest objects and merge them to form a new cluster
2. Find and merge the next two nearest objects where a point is either an individual object or a cluster of objects.
3. If more than one cluster remains , return to step 2

2.3 Partition Clustering:

It attempts to directly decompose the given data set or objects into a set of disjoint clusters. Typically the global criteria involve minimizing some measure of dissimilarity in the samples within each cluster, while maximizing the dissimilarity of different clusters.

2.4 The Complete Link Method

The complete link method is quite similar to the single link method except that it uses the least similar pair between two or more clusters to determine the inter-cluster similarity for generation of 'n' clusters. This method is characterized by small, tightly bound clusters. Therefore need arises to find out the similar

clusters either based on their distance or similarity.

And many more techniques present in the market for generation of clusters.

III. SURVEY ON DOCUMENT CLUSTERING

3.1 Document Clustering

Document clustering is automatic document collection or grouping, topic extraction, fat and effective information retrieval. It is closely related to data clustering.

Examples:

- Clustering will divides the results of a search for "cell" into groups like "biology," (which is a branch of science),"battery,"(cell used in battery) and "prison."

This approach will be very effective if we successfully form the clusters based on some similarity as we will retrieve the 'n' relevant documents within less steps.

Document clustering involves the use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster which contains the 'n' objects.

The application of document clustering can be categorized to two types, online and offline. Online applications are usually constrained by efficiency problems when compared offline applications.

Means we can store the searched results with respect to one query and will produce if that query in future will be executed once again.

If the number of clusters K is unknown before the clustering process, one solution is to estimate K first and use this estimation as the input parameter for those document clustering algorithms requiring K predefined.

3.2 Objective:

When the processing task is to be performed on the documents there is need to partition a given document collection into clusters of similar documents a choice of good features where what requires a good clustering algorithms to give better results.

A common task of text processing in many information retrieval applications is based on the analysis of word occurrences i.e. how many number of times a particular word is occurring across a document collection.

The number of words used by the application defines the dimension of a vector space (an information model) in which the analysis is carried out. Reduction of the dimension may lead to significant savings of computer resources and processing time. This reduction can be carried out by removing common words or those words which are occurring frequently in document.

However poor feature selection may dramatically degrade the information retrieval system's performance if we will fail to process that document effectively.

Even we can think about the similar words which are occurring repeatedly in the document which degrades the performance of information retrieval system.

3.3 Existing Dirichlet Process Mixture Model (DPM)

This elasticity of the DPM form makes it particularly capable for document clustering.

There is little work investigating this model for document clustering due to the high-dimensional representation of text documents. In the problem of document clustering, each document is represented by a large amount of words including discriminative words and nondiscriminative words.

Only discriminative words are useful for grouping documents. The involvement of nondiscriminative words confuses the clustering process and leads to poor clustering solution in return. When the number of clusters is unknown, the affect of nondiscriminative words is motivated.

Words in documents are partitioned into two groups, in particular, discriminative words and nondiscriminative words.

Each document is considered as a mixture of two components. The first component, discriminative words are generated from the specific cluster to which document belongs. The second component, nondiscriminative words, are generated from a general background shared by all documents present in that collection. Idea is to use only discriminative words to infer the document cluster structure.

There are two algorithms to infer DPM model parameters, in particular first one is the variational inference algorithm and second one is the Gibbs sampling algorithm. It is hard to apply the Gibbs sampling algorithm to document clustering since it needs long time to converge. Because of the high-dimensional representation of given text documents, it is even harder to be applied when the document data set is large.

The key concept here is the variational inference decreases posterior judgment to an optimization problem. Optimization is normally much quicker than approaches to posterior estimation

3.4 Mean Field Variational Inference

Mean field variational inference is a particular class of variational methods. Consider a model with a hyperparameter θ , latent variables W= { v1,v2,v3,……v_n} and data points x={ x1,x2,…Xd} . Mean field variance method starts from a family of distributions Q by using which both the mean field procedure and the subsequent inference procedures are easy to handle.

In order to yield a computationally effective inference method, it's very necessary and important to choose a reasonable family of distributions Q.

IV. PROPOSED SYSTEM

4.1 Basic initial flow of the system:

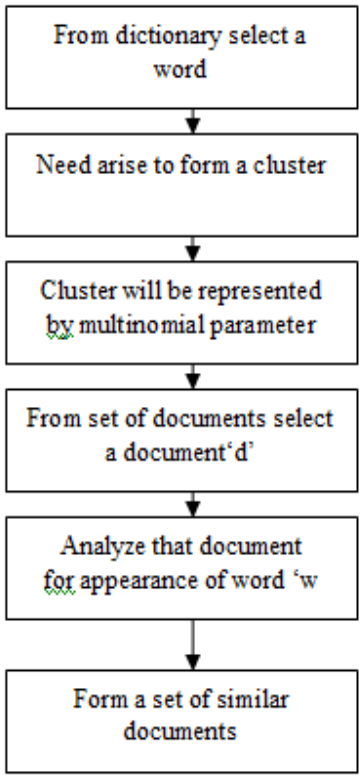


Fig 4.1 Flow of initial processing of system

As observe in the above flow chart first of all what we have to do is to prepare a list of words i.e. vocabulary or dictionary.

Then selection of document one by one to check whether the term or word is occurred in that particular document or not.

Based on this we will try to extract another features of the documents and will prepare a set of features.

Now by considering that set of features we will form a set document in which 'n' documents will poses that particular feature.

4.2 Use of Blocked Gibbs Sampling Algorithm

Another effective inference algorithm for our proposed model is the blocked Gibbs sampling algorithm.

Gibbs sampling or a **Gibbs sampler** is a Markov chain Monte Carlo algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution as we are using it for identifying the relationship between 'n' documents and trying to form a group of documents based some similar features.

It is commonly used as a means of statistical inference, especially Bayesian inference, Bayesian inference derives the posterior probability as a consequence of two antecedents, a prior probability and a "likelihood function" derived from a probability model for the data to be observed. Bayesian inference computes the posterior probability according to Bayes' rule which is as follows.

$$P(H|E) = P(E|H).P(H)/P(E)$$

It is a randomized algorithm (i.e. an algorithm that makes use of random numbers, and hence may produce different results each time it is run), and is an alternative to deterministic algorithms for statistical inference such as variation Bayes or the expectation-maximization algorithm .

4.3 Proposed Idea with example

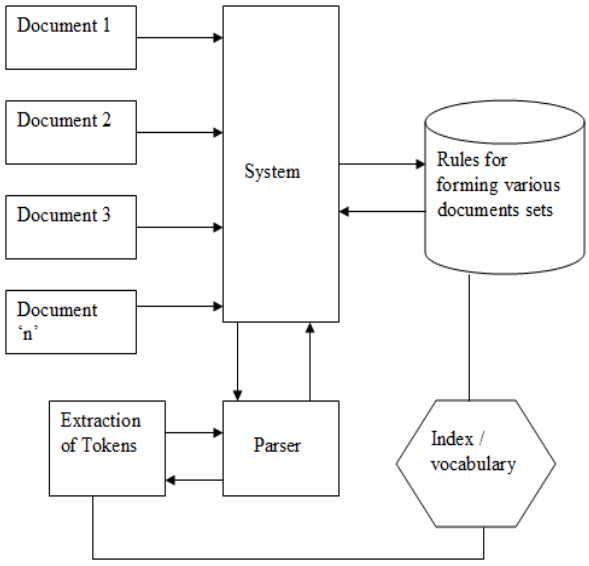
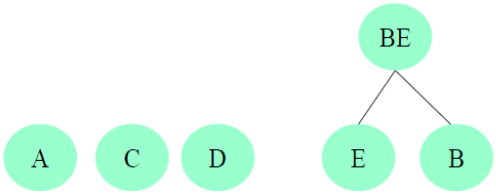


Fig 4.2 Proposed Architecture

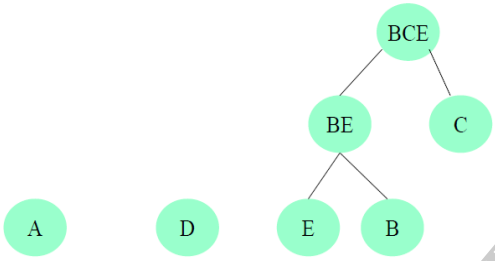
Above proposed architecture depicts that what we wish to perform after collecting 'n' no. of documents.

- Consider A, B, C, D, E as documents with the some similarities:
- Then we will have documents A,B,C,D,E

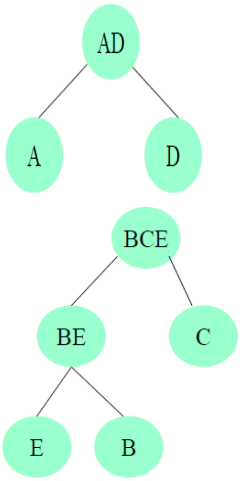
Now we will try to group the given documents based on some similar features and scenario will look like as follows:



at the next step it will be as follows :



and few more scenarios will be as follows:



Parser module which will handle prime task is to parse given documents into tokens & preparation of index or we can say the vocabulary will starts.

System will use this index/ vocabulary for further process by forming various rules so that it will form 'n' document sets.

In our proposed work what we wish to do is as follows:

First of all we have many documents as a input, then we will prepare a index or vocabulary for those documents.

Now we need identify various features i.e. actually we will predict some features and by considering those features as a base we will cluster the documents.

Additionally we can add an information model such as probabilistic model, fuzzy model, Boolean model, extended Boolean model etc. we will compare the features and form a cluster of documents.

CONCLUSION

We will show that following targets will definitely achieve as follows if we will form a set or clusters of given documents :

- Document clustering is the act of collecting similar documents into bins, where similarity is some function on a document.
- If a collection is well clustered, we can search only the cluster that will contain relevant documents.
- Searching a smaller collection should improve effectiveness and efficiency.

So it will very useful to have clusters of data based on some similarity. In our proposed system we will use Dirichlet Process Mixture Model, mean variance algorithm and blocked gibbs sampling algorithm. Our proposed system will be as discussed above.

REFERENCES

1. Michael Steinbach George Karypis Vipin Kumar,"A Comparison of Document Clustering Techniques" Department of Computer Science and Engineering, University of Minnesota
2. Inderjit Dhillon, Jacob Kogan,Charles Nicholas ,"Feature Selection and Document Clustering"
3. http://en.wikipedia.org/wiki/ Bayesian_inference
4. C. Elkan, "Clustering Documents with an Exponential-Family Approximation of the Dirichlet Compound Multinomial Distribution," Proc. Int'l Conf. Machine Learning, pp. 289-296, 2006.
5. R. Madsen, D. Kauchak, and C. Elkan, "Modeling Word Burstiness Using the Dirichlet Distribution," Proc. Int'l Conf. Machine Learning, pp. 545-552, 2005.