Review on Text Clustering Algorithms

Manpreet Kaur[1] Navpreet Kaur[2]
[1]M.Tech & Research Scholar [2]Assistant Professor
[1, 2] Department of Computer Science and Engineering
[1, 2]Sri Guru Granth Sahib World University Fatehgarh Sahib, Punjab, India

Abstract---A clustering algorithm finds a partition of a set of objects that fulfills some criterion based on these conditions. Clustering is an unsupervised method of learning. Most text clustering algorithms are based on the vector space model which has the advantages of simple concept ad convenient applications. While the Latent Semantic analysis method takes the relationship between words into account and supposed to be an improved model of VSM. K-Mean clustering algorithm has shortcoming, which depend on the initial clustering center and needs to fix the number of clusters in advance. Vector Space model has problems, such as high dimensional and sparse. This can be optimized using various optimization techniques such as Genetic algorithm, PSO, pollination Based optimization. Pollination based optimization is inspired by natural flower pollination.

Keywords: Text clustering, Vector Space Model, Latent Semantic Analysis, K-means clustering algorithm, clustering optimization.

I. INTRODUCTION

Data Mining is the process of discovering interesting patterns and knowledge from large amounts of data. Data mining is often treated as term, Knowledge Discovery in Databases (KDD). By observing large data sets over a period of time, we can deduce previously-unknown and useful information concerning patterns, models, trends, and rules in the area of application. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining uses information from past data to analyze the outcome of a particular problem or situation that may arise.

A. Data mining involves six common classes of tasks in Data Mining

1) Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.

2) *Association rule learning (Dependency modeling)* – Searches for relationships between variables. For example, the supermarket can determine which products are frequently bought together and use this information for market analysis. This is sometimes referred to as market basket analysis.

3) *Clustering* – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

4) *Classification* – is the task of generalizing known structure to apply to new data. For example, an e-mail program used to classify an e-mail as "legitimate" or as "spam".

5) *Regression* – Attempts to find a function which models the data with the least error.

6) *Summarization* – providing a more compact representation of the data set, including visualization and report generation.

B. Text Clustering

The objective of clustering is to partition an unstructured set of objects into clusters (groups). One often wants the objects to be as similar to objects in the same cluster and as dissimilar to objects from other clusters as possible.

To use most clustering algorithms following two points are necessary:

- An object representation,
- A similarity (or distance) measure between objects.

1) Applicability of clustering for a number of tasks Document Organization and Browsing:

The hierarchical organization of documents into coherent categories can be very useful for systematic browsing of the document collection. A classic example of this is the Scatter/Gather method, which provides a systematic browsing technique with the use of clustered organization of the document collection.

2) *Corpus Summarization:* Clustering techniques provide a coherent summary of the collection in the form of cluster-digests or word-clusters, which can be used in order to provide summary insights into the overall content of the underlying corpus... The problem of clustering is also closely related to that of dimensionality reduction and topic modeling. Such dimensionality reduction methods are all different ways of summarizing a corpus of documents.

3) *Document Classification:* While clustering is inherently an unsupervised learning method, it can be used in order to improve the quality of the results in its supervised variant. In particular, word-clusters and co-training methods can be used in order to improve the classification accuracy of supervised applications with the use of clustering techniques.

C. Vector Space Model

Most clustering algorithms use the vector space model of IR, in which text documents are represented as a set of points in a high dimensional vector space. Similarity between two texts is traditionally measured by the cosine of the angle between their vectors, though Cartesian distance. Documents judged to be similar by this measure are grouped together by the clustering algorithm. These algorithms are notoriously slow, because the time required computing a similarity score for each pair of documents is proportional to the size of the smaller document. One method involves keeping only N most important terms from each document (as judged by the chosen term weighting scheme), where N is much smaller than document size. Latent Semantic Analysis (LSA) applies to the document collection to create a new abstract vector space, which has a property that

vectors describing this vector space are ordered according to their importance to describing the document collection.

1) Document Preprocessing

Before representing documents as tf-idf vectors, some preprocessing we need. There are commonly two steps: First, we need to remove stop words, such as 'a', 'any', 'what', 'I', etc., since they are frequent and carry no information. A stop words list can be found online. Secondly, stem the word to its origin, in which we only consider the root form of words.

2) tf-idf Matrix

The Vector Space Model is the basic model for document clustering, upon which many modified models are based. In this model, each document, dj, is represented as a term-frequency vector in the term-space:

$$\boldsymbol{d}_{j_{tf}} = (tf_{1j}, tf_{2j}, ..., tf_{Vj})' \qquad j = 1, 2, ..., D$$

Where tf-ij is the frequency of the ith term in document dj, V is the total number of the selected vocabulary, and D is the total number of documents in the collection. Next, weight each term based on its inverse document frequency (IDF). The basic idea is that if a term appears frequently across all documents in a collection, its discriminating power should be discounted. So finally, we obtain a tf-idf vector for each document

$$\boldsymbol{d}_j = (tf_{1j} \times idf_1, tf_{2j} \times idf_2, ..., tf_{Vj} \times idf_V)' \qquad j = 1, 2, ..., D$$

Put these tf-idf vectors together, to get a tf-idf matrix:

$$(tf - idf)_{ij} = tf_{ij} \times idf_i \qquad i = 1, 2, ..., V; \qquad j = 1, 2, ..., I$$

3) Similarity Measure and Clustering Algorithm

The most commonly chosen measure is the cosine similarity. The choice of a similarity measure can be crucial to the performance of a clustering procedure. And the Euclidean distance behave poorly in some situations, while the cosine similarity captures the 'directional' characteristic which is intrinsic of the document tf-idf vectors. There are two categories of clustering algorithms are , the partitioning and the agglomerative.

4) The Baseline and its Problems

Document clustering by using the K-means algorithm with cosine similarity (spkmeans) to the full space VSM has been considered as a baseline when doing performance comparison. This algorithm is straightforward and easy to implement, but it need high computational cost, which makes it less appealing when dealing with a large collection of documents. The curse of dimensionality problem is that : V is large. Generally, there are more than thousands of words in a vocabulary, which makes the term space high dimensional. Hence, various dimensionality reduction techniques have been developed to make improvements above the baseline.

D. Latent Semantic Analysis

Latent Semantic Analysis has been proposed as a tool for uncovering common patterns of word usage across a large number of documents. It is similar to the Principal Component Analysis used by statisticians to analyze data sets composed of many highly correlated variables and to represent them in terms of a few uncorrelated variables. LSA uses a mathematical technique called Singular Value Decomposition (SVD) to create a new abstract vector space that is the best representation of the document collection in the least-squares sense. SVD also computes two sets of independent vectors that form a basis of this vector space. One set of vectors provides a transformation from the original vector space of terms to a new LSA-space, and the other set provides a transformation from the document space to the new LSA-space. Moreover, SVD returns an ordering of these vectors, so that the first vector is the most informative, that is it describes the strongest regularities of word usage in the document collection. The second vector describes those aspects of word usage not captured by the first vector, and so on to the last and least informative vector

E. K-mean clustering algorithm

1) Algorithm: K-means:-The K-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

2) Input:

K: the number of clusters,

D: a data set containing n objects.

3) Output:

A set of k clusters.

4) Method:

Randomly choose k objects from D as the initial cluster centers;

5) Repeat

(Re) assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

Update the cluster means, which means to calculate the mean value of the objects for each cluster;

6) Until no change;

The main objective of standard k-means clustering is to minimize the mean-squared error.

$$E = \frac{1}{N} \sum_{\mathbf{x}} \|\mathbf{x} - \mu_{k(\mathbf{x})}\|^2$$

Where k(x) = $\arg \min_{k \in \{1,...,K\}} \|\mathbf{x} - \mu_k\|$ is the index of the closest cluster centroid to x, N is the total number of data vectors. It is well-known that the underlying probability distribution for the standard k-means algorithm is Gaussian [7].

F. PBO

Optimization is a natural process embedded in the living beings. Pollination is used to transfer the pollen from male parts of flower called anther to the female part called stigma of a flower. Some flowers will develop seeds as a result of self-pollination, in self-pollination pollen and pistil are from the same plant, often (but not always) from the same flower. Other plants require cross-pollination: in cross pollination pollen and pistil must be from different plants. Pollination is beneficial for plants because the movement of pollen allows them to reproduce by setting seeds. However, pollinators don't know or care that the plant benefits. They pollinate to get pollen and/or nectar from flowers to meet their energy requirements and to produce offspring. If pollination process is above normal the plants reduce expenditure on resources for producing nectar, floral display and fragrance in the flowers. If the pollination process is below normal, plants increase the resource expenditure so that more fragrance, floral display, and nectar to attract pollinator.

II. ENHANCEMENTS ON CLUSTERING ALGORITHM

Praveen and Dora Babu sudarsa [1] proposed an efficient algorithm for document clustering. Document clustering is the task of grouping a set of documents into clusters so that the documents in the same cluster are similar to each other than to those in other clusters. K-means is a commonly used algorithm for document clustering, but K-means has some limitations : 1) The number of clusters K has to be given as input and 2) Based on the initializations it converges to different local minima. 3) It is very slow and cannot be used for large number of data novel algorithm to eliminate all these basic drawbacks of K-means.

Canyu Wang and Xuebi Guo [2] proposed LSA to measure the similarities between topics and Crime Prototype Vector, and some similarities will be used as the weights of the paths in the network hierarchies and calculate the suspicious degrees. In the criminal cases, the investigators or the police have to make full use of the messages or spoken documents data that they record in files. In Information Retrieval area, Latent Semantic Analysis (LSA) is an important method for query matching which can discover the underlying semantic relation or similarity between words and topics.

Wang Chun-hong and Nan Li-Li; [3] presented a text clustering algorithm based on Latent Semantic Analysis and Optimization is proposed. This algorithm overcome the problems of Vector Space Model, and also can avoid the shortcomings of k-means algorithm. And when compared with the text clustering algorithm based on Latent Semantic Analysis and the text clustering algorithm based on Vector Space Model and optimization, their algorithm is proved which can improve the effect of text clustering, and upgrade the precision ratio and recall ratio of text.

Kristina Lerman [4] investigated the use of Latent Semantic Analysis to create a new vector space that is the optimal representation of the document collection. Documents are projected onto a small subspace of this vector space and clustered. They compare the performance of clustering algorithms when applied to documents represented in the full term space and in reduced dimension subspace of the LSA-generated vector space. They report significant improvements in cluster quality for LSA subspaces with optimal dimensionality. They discuss the procedure for determining the right number of dimensions for the subspace

Michael Steinbach [5] presented the results of an experimental study of some common document clustering techniques. In particular, they compare the two main approaches to document clustering are agglomerative ,hierarchical clustering and K-means. (For K-means we used a "standard" K-means algorithm and a variant of K-means, we use "bisecting" K-means.) Hierarchical clustering is often portrayed as the better quality clustering approach, but it is limited because of its quadratic time complexity. In contrast sometimes K-means and agglomerative hierarchical approaches are combined so as to "get the best from both worlds." However, our results indicate that the bisecting K-means technique is better than the standard K-means approach and as good as or better than the hierarchical approaches that we tested for a variety of cluster evaluation metrics.

Bohdan Pavlyshenko [6] analyzed the clustering of text documents in the vector space of semantic fields and in the semantic space with orthogonal basis. It is shown that using the vector space model with the basis of semantic fields is effective of author's texts in English fiction in the cluster analysis algorithms. The analysis of the author's texts distribution in cluster structure showed the presence of the areas of semantic space that represent the author's idiolects of individual authors. SVD factorization of the semantic field's matrix makes it possible to reduce significantly the dimension of the semantic space in the cluster analysis of author's texts.

Xiaohui Cui and Thomas E. Potok [7] presented a hybrid Particle Swarm Optimization (PSO)+K-means document clustering algorithm that performs fast document clustering and can avoid being trapped in a local optimal solution as well. For comparison purpose, we applied the PSO+K-means, PSO, K-means, and other two hybrid clustering algorithms on four different text document datasets. The number of documents in the datasets ranges from 204 to over 800, and the number of terms range from over 5000 to over 7000. The results illustrate that the PSO+K-means algorithm can generate the most compact clustering results than other four algorithms.

Andreas Hotho and Steffen Staab [8] discussed a way of integrating a large thesaurus and the computation of lattices of resulting clusters into common text clustering in order to overcome these two problems. As its major result, their approach achieves an explanation using an appropriate level of granularity at the concept level as well as an appropriate size and complexity of the explaining lattice of resulting clusters. Common text clustering techniques offer rather poor capabilities for explaining to their users to achieve particular result. They have the limitation that they do not relate semantically nearby terms and that they cannot explain how resulting clusters are related to each other.

Algorithm	**Advantages**	**Disadvantages**
K-means algo	1) If variables are huge,then K-means most of the times computationally faster than hierarchical clustering, if k smalls. 2)K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.	1) Difficult to predict K-Value. 2)It didn't work well with global cluster. 3) Different initial partitions can result in different final clusters. 4) It does not work well with clusters (in the original data) of Different size and Different density
Vector Space Model	1)Simple model based on linear algebra. 2)Term weights not binary. 3)Allows ranking documents according to their	1)Longer documents are poorly represented because they have poor similarity values. 2)The order in which the terms appear in the document is lost in

	possible relevance. 4)Allows partial matching	the vector space representation. 3)Theoretically assumes terms are statistically independent.
Latent Semantic Analysis	1)LSI overcomes two of the most problematic constraints of Boolean keyword queries: synonymy and polysemy. 2)LSI is also used to perform automated document categorization.	1)LSI requires relatively high computational performance and memory in comparison to other information retrieval techniques. 2) Difficulty in determining the optimal number of dimensions to use for performing the SVD.

III. CONCLUSION

A text clustering algorithm based on latent semantic analysis to overcome the short comings of vector space model and k-mean algorithm. K-Mean clustering algorithm has shortcoming, which depend on the initial clustering center and needs to fix the number of clusters in advance. Vector Space model has problems, such as high dimensional and sparse. Genetic algorithm has several advantages over the vector space model for text clustering like as high-dimensional and sparse problem.Hence an enhanced Pollination based optimization with k-mean clustering for text clustering can be used for text efficient retrieval.

REFERENCES

[1] M. Praveen, Dora Babu Sudarsa, "A Customized Vector Space Model Implementation in Document Clustering to Enhance the Performance" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013.
[2] Canyu Wang, Xuebi Guo, Hao Han, "Crime Detection Using Latent Semantic Analysis and Hierarchical Structure" IEEE 3rd International Conference on Software Engineering and Service Science (ICSESS), pp. 337 – 340, 22-24 June 2012.
[3] Wang Chun-hong, Nan Li-Li; Ren Yao-Peng, "Research on the text clustering algorithm based on latent semantic analysis and optimization" 2011 IEEE International Conference on (Volume:4), pp. 470 – 473, 10-12 June 2011.
[4] Kristina Lerman, "Document Clustering in Reduced Dimension Vector Space", University of Minnesota,2009.
[5] Michael Steinbach, George Karypis, Vipin Kumar, "A Comparison of Document Clustering Techniques", University of Minnesota,2008.
[6] Bohdan Pavlyshenko,"The Clustering of Author's Texts of English Fiction in the Vector Space of Semantic Fields", 2012.
[7] Xiaohui Cui, Thomas E. Potok, "Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm", Journal of Computer Sciences :27-33, 2005.
[8] Andreas Hotho, Steffen Staab, Gerd Stumme, "Explaining Text Clustering Results using Semantic Structures", Institute of Applied Informatics and Formal Description Methods AIFB,University of Karlsruhe,May 2008.

