RESEARCH ARTICLE OPEN ACCESS

Review on Text Clustering Using Statistical and Semantic Data

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

ABSTRACT
The explosive growth of information stored in unstructured texts created a great demand for new and powerful tools to acquire useful information, such as text mining. Document clustering is one of its the powerful methods and by which document retrieval, organization and summarization can be achieved. Text documents are the unstructured databases that contain raw data collection. The clustering techniques are used group up the text documents according to its similarity. As there is a huge amount of unstructured data and there is a semantic correlation between features of data it is difficult to handle that. There are large no of feature selection methods that are used to used to improve the efficiency and accuracy of clustering process. The feature selection was done by eliminate the redundant and irrelevant items from the text document contents. Statistical methods were used in the text clustering and feature selection algorithm. The semantic clustering and feature selection method was proposed to improve the clustering and feature selection mechanism with semantic relations of the text documents.

Keywords:- Clustering, CHIR, CHIRSIM, K-means algorithm

I. INTRODUCTION

With the huge growth of information that is stored in unstructured texts created a great demand for new and powerful tools, such as text mining techniques, to acquire interesting information. Text mining is a process of extracting useful information and knowledge from unstructured textual documents [1].However, Text clustering technique is the process of grouping from a set of objects. The objects within a cluster are similar to each other, but they are dissimilar to objects in other clusters. Some text clustering techniques don't work efficiently in high dimensional data. For this dimensionality reduction is done. Dimensionality reduction is the process of decreasing the dimension of the dataset to a manageable size keeping as much as possible the original information. Dimensionality reduction techniques can be divided into feature extraction and feature selection techniques. The first method refers to the mapping of the original high-dimensional data onto a lower dimensional space. The second method chooses an optimal subset of features according to an objective function.

II. CLUSTERING USING STATISTICAL AND SEMANTIC DATA

The clustering process cannot be directly applied on textual documents. An indexing procedure, mapping each textual document of the dataset into a compact representation of its content, needs to be uniformly applied. There are several studies that are focused on feature selection methods. Authors in [8], uses a new frequency feature selection method called Third moment which has an ability to enhance rare features. Similarly, a strategy called CMFS was proposed in [9], it comprehensively measures the significance of a term by calculating its significance from both inter-category and intra-category. Another statistical method was introduced in [6]; based on the well known chi-square metric, the CHIR statistic keeps only terms those are positively dependent to the categories. However, all these methods are

frequency feature selection techniques. An interesting research topic is to perform a clustering document using not only a statistical feature selection method but also a semantic one as that improves the clustering that had already been performed using statistical feature selection method [11].
In [2] it is proposed that a document clustering is possible through two stages: a statistical clustering followed by a semantic one. The statistical weight of each term is estimated by the CHIR-statistic [5] and the semantic weight is estimated using a new measure based on the mutual information. The clustering and feature selection processes are done in parallel order.
There are different feature selection methods used in[2] which are described as:

A. Vector Space Model

In this model each document is considered as a term-weight vector. For determining the weight of a term in a document, we choose the standard Term Frequency-Inverse Document Frequency function. With the help of these functions similarity between two documents is measured, for similarity cosine function is used.

Cosine (di, dj) = di*dj/||di|||dj||

When two documents are identical, the cosine value is equal to 1, and it takes the value 0 when the documents are totally dissimilar and have nothing in common. Larger cosine values indicate that these two documents are likely similar since they share most of their words.

B. CHIR Method

CHIR is a new feature selection technique introduced in [5] and based on the X2 statistic. In order to select only relevant terms that have strong positive dependency on certain categories in the dataset, [5] propose a new measure that determines whether the dependency between two attributes is positive or negative. It is defined as:

$$R(w,c) = O(w,c)/E(w,c)$$

Where O(w,c) is the observed frequency of the documents that belong to the category c and contain w, E(w,c)is the expected frequency of the category c and the term w. If the term w and the category c are independent, R(w,c) is close to 1. If there is a positive dependency, R(w,c) is larger than 1 and when there is a negative dependency R(w,c) is smaller than 1.

C. SIM Method

In this similarity of terms is measured two terms are considered similar if their mutual information with all terms in the vocabulary is nearly the same. These are the supervised feature selection techniques however if there is a unsupervised data it will be difficult to implement them because class label information is required before implementation.

D. CHIRSIM Method

In order to implement supervised feature selection techniques in text clustering we first implement k-mean algorithm to get initial clusters and centroïds. We then apply the feature selection method and perform the clustering iteratively.

a) Initial step:
1. Perform the k-means algorithm on the corpus to get initial clusters and centroids.

b) Statistical step: TCFS algorithm:

2. Perform the CHIR method by using the previous clustering result. The weight of each unselected feature is reduced.

3. Recalculate k centroids in the new feature space.
4. For each document in the dataset, calculate its similarity with each centroid and assign it to the
Closest cluster.
5. Repeat steps 2, 3, and 4 until convergence.

c) Semantic step: the initial clusters and centroids are obtained from the CHIR clustering:
6. Perform the SIM similarity by using the previous clustering result. The weight of each unselected feature is reduced by f (- （0 ,1）.
7. Recalculate k centroids in the new feature space.
8. Reassign the documents to the closest cluster
9. Repeat steps 6, 7, and 8 until convergence.

We first build clusters by performing the k-means algorithm as initial clustering step. We then categorize the documents starting by the k clusters guided by the k-means and using the CHIR-statistical measure to select relevant features. Finally, we use the clusters and centroids obtained from the statistical clustering as input to carry out the semantic clustering in which the mutual information measure SIM (a semantic measure) is used as feature selection method. The two feature selection methods use the current clusters and their centroids to estimate the relevance of each term to the dataset. If a term is considered relevant, it is kept in the feature space. Otherwise, the term weight is reduced by a factor f which is a predetermined factor in the range of [0, 1].

III. ENHANCEMENT

K-means algorithm is a popular clustering technique and it was successfully applied to many of practical clustering problems since it suffers from several drawbacks due to its choice of initializations. Recent advancements in clustering algorithm introduce the evolutionary computing such as genetic algorithms [3] and particle swarm optimization [4, 5].

- **Swarm Intelligence**

Swarm Intelligence (SI)[7] is an innovative, method for solving optimization problems that organized from the study of colonies, or swarm of social organisms. The state of the art clustering algorithms based on SI tools are Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO).

- **Particle Swarm Optimization**

It is an optimization approach based on the social behavior of animals such as a flock of birds, a school of fish, or a swarm of bees or a group of people who have a common goal in their lives. PSO is a population-based search procedure where the individuals, referred to as particles, are grouped into a swarm. Each particle in the swarm represents a candidate solution to the optimization problem. In a PSO system, each particle is "flown" through the multidimensional search space, adjusting its position in search space according to its own experience and that of neighboring particles. A particle therefore makes use of the best position by itself and the best position of its neighbors to position itself toward an optimal solution. It has a fast convergence when compared with many global optimization algorithms like Genetic Algorithms (GA), Simulated Annealing (SA) and others.

- **Ant Colony Optimization**

ACO deals with artificial systems that is inspired from the foraging behavior of real ants, which are used to solve discrete optimization problems [20]. The main idea is the indirect communication between the ants by means of chemical pheromone trials, which enables them to find shortest path between their nest and food. The more the number of ants traces the given path, the more attractive this path (trail) becomes and is followed by other ants by depositing their own pheromone. This auto catalytic and collective behavior results in the establishment of the shortest route. Ants find the shortest path based on intensity of pheromone deposited on different paths. There exist a large number of clustering algorithms in the literature. No single algorithms is suitable for all types of objects, nor all algorithms appropriate for all problems. Many of clustering algorithms have a number of drawbacks. Recently, algorithms inspired by nature used for clustering. It is claimed that ant-based clustering algorithms can overcome these drawbacks. These algorithms have advantages in many aspects, such as self-organization, flexibility, robustness, no need of prior information, and decentralization. Research on ant-based clustering algorithms is still an on-going field of research. Clustering is done using groups of ants which are as many as the number of clusters. The goal of each group is to collect members of one cluster.

These algorithms are not consistent because these are dependent on parameter values of application and provide less optimal solutions. Also, These algorithms takes more time to achieve optimality. So to improve this new algorithm that is firefly is developed.

- **Firefly Algorithm**

For global optimization there are various meta heuristics algorithms. A subset of meta heuristics are often referred to as swarm intelligence (SI) based algorithms. These algorithms are based on characteristics of biological agents such as birds, fish, humans and others. For example, particle swarm optimization was based on the swarming behavior of birds and fish [15], while the firefly algorithm was based on the flashing pattern of fireflies. Among these new algorithms, it has been shown that firefly algorithm is very efficient in dealing with multimodal, global optimization problems.
FA uses the following three idealized rules:
• Fireflies are unisex so that one firefly will be attracted to other fireflies regardless of their sex.
• The attractiveness is proportional to the brightness, and they both decrease as their distance increases.
• The brightness of a firefly is determined by the landscape of the objective function.

Firefly algorithm is used for feature selection and showed that firefly algorithm produced consistent and better performance in terms of time and optimality than other algorithms [14].
It is found that FA can outperform PSO and obtained global best results [16].

- **K-MEANS Clustering Algorithm [12]**

K-means clustering is a well known partitioning method. In this objects are classified as belonging to one of K-groups. The results of partitioning method are a set of K clusters, each object of data set belonging to one cluster. In each cluster there may be a centroid or a cluster representative. In case where we consider real-valued data, the arithmetic mean of the

attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases.[12,13]. As mentioned previously, clustering means the division of a dataset into a number of groups such that similar items falls or belong to same groups. In order to cluster the database, K-means algorithm uses an iterative approach.

- **Hierarchical Clustering [13]**

Hierarchical methods are well known clustering technique that can be widely useful for various data mining tasks. A hierarchical clustering scheme produces a sequence of clustering in which each clustering is nested into the next clustering in the sequence. Since hierarchical clustering is a greedy search algorithm based on a local search, the merging decision made early in the agglomerative process are not necessarily the right ones. One possible solution to this problem is to refine a clustering produced by the agglomerative hierarchical algorithm to potentially correct the mistakes made early in the agglomerative process. Hierarchical methods are commonly used for clustering in Data Mining.
Since k-means takes less execution time than h-means but h-means provide better quality as compared to k-means. As no of records increases then execution time increases in case of hierarchical clustering.

IV. CONCLUSION

Nature-inspired meta heuristic algorithms have gained popularity, which is partly due to their ability of dealing with nonlinear global optimization problems. These algorithms are feature selection algorithms that determine the optimal solutions. Feature selection is a valuable preprocessing technique for applications involving huge data. It refers to the problem of selecting minimal attribute set that are most predictive to represent the original attributes in data set. The paper discussed the strengths and weaknesses of various existing feature selection methods. These methods either fail to find optimal data reductions or require more time to achieve better results. In this paper it is discussed that as k-means algorithm is used for clustering to find clusters that have some drawbacks so instead of this if h-mean is implemented and optimized using any of feature selection method may provide better results. When feature selection methods are used out of all, firefly algorithm according to its characteristics can outperform. When comparison is being made of all algorithms for optimal solutions then experimental results prove that our algorithm exhibits consistent and better performance in terms of time and optimality as compared to other methods.

REFERENCES:

[1] M. Thangamani and P.Thangaraj, “Survey on Text Document Clustering”, International Journal of Computer Science and Information Security, 2010, volume 8(2), pp. 174-178, 2010.
[2] A. Benghabrit, B.Ouhbi, H.Behja,’’Text Clustering Using Statistical and Semantic Data’’, World Congress on Computer and Information Technology, volume 1, 22-24 June 2013.
[3] K. J. Kim and H. Ahn, "A recommender system using GA K-means clustering in an online shopping market," *Expert Systems with Applications*, volume. 34, pp. 1200-1209, 2008.
[4] S. Paterlini and T. Krink, “Differential evolution and particle swarm optimization in partitional clustering,"*Computational*

Statistics & Data Analysis, volume. 50, pp. 1220-1247, 2006.

[5] S. Rana, S. Jasola, and R. Kumar, "A review on particle swarm optimization algorithms and their applications to data clustering," *Artificial Intelligence Review*, volume. 35, pp. 211-222, 2011.

[6] Y.Li, C.Luo and S.M.Chung, "Text Clustering with Feature Selection by using Statistical Data Knowledge and Data Engineering", IEEE Transactions on Know and Data Eng., volume 20(5), pp.641–651, 2008.

[7] M.Thangamani and P.Thangaraj, "Integrated Clustering and Feature Selection Scheme for Text Documents", Journal of Computer Science, volume 6(5), pp. 536-541, 2010.

[8] F.Peleja, G.P.Lopes, and J.Silva, "Text Categorization: A Comparison of Classifiers, Feature Selection Metrics and Document Representation"; Proceedings of the 15th Portuguese Conference in Artificial Intelligence, pp.660-674, 2011.

[9] J.Yang, Y.Liu, X.Zhu, Z.Liu, and X.Zhang, "A New Feature Selection Base on Comprehensive Measurement both in Inter-category and Intracategory for text categorization"; Information Processing & Management, volume 48(4), pp.741-754, 2010.

[10] M.J.Meena, K.R.Chandran and J.M.Brinda, "Integrating Swarm Intelligence and Statistical Data for Feature Selection in Text Categorization", International Journal of Computer Applications, volume 1(11), pp.16-21, 2010.

[4] S. Paterlini and T. Krink, "Differential evolution and particle swarm optimisation in partitional clustering," Computational Statistics & Data Analysis, volume 50, pp. 1220-1247, 2006.

[6] S. Rana, S. Jasola, and R. Kumar, "A review on particle swarm optimization algorithms and their applications to data clustering," Artificial Intelligence Review, volume 35, pp. 211-222, 2011.

[11] A. Abraham, He Guo and Hongbo Liu, "Swarm Intelligence:Foundations, Perspectives and Applications", Swarm Intelligence in Data Mining, A. Abraham, C. Crosan, V. Ramos (Eds.), Studies in Computational Intelligence (series), Springer, Germany, 2006.

[12] Kehar Singh, Dimple Malik and Naveen Sharma " Evolving limitations in K-means algorithm in data mining and" IJCEM International Journal of Computational Engineering & Management, Volume 12, April 2011

[13] Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur" *Efficent K-means Clustring Algorithm Using Ranking Method In Data Mining*"ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012.

[14] H. Banati and M. Bajaj, Firefly based feature selection approach, *Int. J. Computer Science Issues*, 8(2), 473-480 (2011).

[15] J. Kennedy and R. Eberhart, Particle swarm optimisation, in: *Proc. of the IEEE Int.*
Conf. on Neural Networks, Piscataway, NJ, pp. 1942-1948 (1995).

[16] M. A. Zaman and M. A. Matin, Nonuniformly spaced linear antenna array design
using firefly algorithm, *Int. J. Microwave Science and Technology*, Volume 2012, Article
ID: 256759, (8 pages), 2012. doi:10.1155/2012/256759