IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Text Clustering Techniques : A Review

Mukesh Kumar[1*], Amandeep Verma[2]

[1*]PG Dept. Of Computer Science, Mata Gujri College, Fatehgarh Sahib, Punjab, India.
[2] Dept. of Computer Science, Punjabi University Neighbourhood Campus, Mohali.

[*]*Corresponding Author: mukeshsun@rediffmail.com , Tel.: +91-81468-00126*

Available online at: www.ijcseonline.org

Accepted: 07/Jun/2018, Published: 30/Jun/2018

Abstract— Text clustering is an unsupervised data mining technique which involves the process of classifying an unlabeled dataset into the groups of similar data objects. These groups are known as clusters; each cluster consists of data objects such that the data objects are more similar within the same group and dissimilar to the data objects of other groups. There is a variety of text clustering techniques used to compute the similarity among the given unlabeled dataset patterns. Moreover, huge literature is available on clustering algorithms and a comprehensive survey would also be an immense task. The purpose of this paper is an attempt to explore the text clustering techniques and to facilitate the researchers for the future inventions. In this paper, literature survey of different text clustering techniques has been performed and presented an analysis of various studies in this area. After reviewing various text clustering techniques from different aspects, this paper suggests research directions for the researchers in this field that can be proved useful for the researchers. Survey of text clustering techniques is performed for the English text/documents as well as for the documents in vernaculars like Gurumukhi script.

Keywords— Text clustering, Clustering techniques, Data mining techniques, Unsupervised learning, Machine learning.

I. INTRODUCTION

I.II Motivation

In the recent years, there has been an immense increase of digital data due to significant increase in usage of Web and online communication. So, retrieving and managing this much huge data is not a simple task and therefore the data mining techniques were developed for managing the huge data. In the field of data mining, automatic classification of text has become an acute need to manage the huge digital data. The automatic classification of text is a process of distributing text in various categories (or set of classes) according to some common characteristics.

Clustering is different from classification as it deals with unsupervised learning of unlabeled data. The term clustering (i.e. unsupervised learning) designates the creation of classes (clusters) of a certain number of similar objects without prior knowledge. The term classification (i.e. supervised learning) deals with the assignment of text document to a class (with predefined classes).

This paper contributes a detailed survey of text clustering algorithms and presents an analysis of various studies in this area. After performing an analysis, the researchers are presented with a clear view of present research work on text clustering and suggested for the future inventions.

I.II Concept of Text Clustering

Text clustering is a collective term for processing of data items that includes several model formation tasks like: elimination of stop words, stemming, syntactic indexing based on term frequencies, semantic indexing based on term document correlations etc.

A good clustering method defines the clusters with high quality in which intra-cluster similarity of data sets is high and the inter-cluster similarity is low. In other words, the documents in one cluster consist the same topic, and the documents in different other clusters represent the different topics.

Thus, a cluster is represented as grouping of similar data sets around a center known as cluster centroid or it may be defined as prototype data instance nearest to the centroid. A cluster can be represented with or without a well defined boundary. The clusters represented with well defined boundaries are called as crisp clusters whereas the clusters represented without well defined boundaries are called fuzzy clusters. The clustering algorithms can be divided into four groups, namely: hierarchical clustering techniques, density based clustering techniques; grid based clustering techniques and partitional clustering techniques [1]:

I.III Outline

This paper is organized as follows. Section-II i.e. the next section of this paper summarizes various types of clustering techniques. Section –III presents the literature survey categorized in hard partitioning clustering algorithms and soft partitioning clustering algorithms. Section-IV i.e. the last section of this paper contains conclusion.

II. TYPES OF CLUSTERING TECHNIQUES

II.I Hierarchical Clustering Techniques

Hierarchical clustering techniques attempt to form a tree of clusters by grouping data objects. In this type of clustering, nested sequence of partitions with single big cluster is produced at the top and singleton clusters at the bottom. In this technique, intermediate level is represented as joining of two clusters or splitting of a cluster into two sub-clusters. The Hierarchical algorithms are either divisive (i.e. top down) or agglomerative (i.e. bottom up).

II.I.I. Divisive clustering

The divisive clustering top-down method. In this technique, the clustering process is started with one top most cluster and splited down the big cluster at each step (according to the similarity between data sets) until only singleton cluster of data sets is remained. Here, the decision to be taken is that which cluster is needed to be splitted down and on what basis.

II.I.II. Agglomerative clustering

The agglomerative clustering is bottom-up method. In this technique, the clustering process is started at the bottom to form a pair of data sets having similarity between them and merge the pair into a common representative. This process is iterated till only one cluster representing the entire original data sets.

Both of the methods stop (merge or split process) when it achieves ‘k’ number of clusters where ‘k’ is defined by user. The main advantage of Hierarchical clustering is ability to handle multiple similarity or distance and flexibility to any level of granularity. Whereas, unreliability and instability are the main disadvantages of hierarchical clustering.

II.II. Density Based Clustering Techniques

The density based clustering techniques use a density function to locate the clusters. In this technique, the clusters are observed as dense regions separated by noise i.e. regions of low density [2]. DBSCAN and NBC (neighbourhood based clustering) algorithms belong to the group of density based clustering [3].

II.III. Grid Based Clustering Techniques

The grid based clustering techniques perform the clustering operations on segmented data space rather than the original data objects. The grid based clustering is performed on the grids by mapping data streams to the defined grids. The typical grid-based algorithms are: STING (Statistical Information Grid), WaveCluster and CLIQUE. The STING approach divides the data space into rectangular cells using hierarchical structure [3]. The WaveCluster approach summarizes the data sets by using a multidimensional grid structure on the data space. In this approach, to find the dense regions in the transformed space, a wavelet transformation is used to transform the feature space [4]. The CLIQUE approach supports both a density and grid based algorithms. This approach works by moving from lower to higher dimensional data space [5].

II.IV. Partitional Clustering

Partitional clustering refers to the process of partitioning the given ‘n’ data sets into ‘k’ number of partitions where k<=n and each partition represents a cluster. In paritional clustering, the following criteria must be followed:

- At least one data object must be contained by each cluster.
- Each data object must belong to one cluster only.

The second criteria may be relaxed in soft clustering algorithms. There are various types of methods come under partitional clustering. The most widely used methods are iterative (or reallocation) and single pass [6]. In partitional clustering algorithms, in order to obtain better results, the distance between data object and the centroid should be minimum. The partitional clustering is of two types, namely: ***hard clustering*** and ***soft clustering***.

II.IV.I. Hard Clustering

In hard clustering (also known as traditional clustering) approach, the dataset is divided into a number of different clusters in such a way that every data item belongs to a single cluster. In this approach, the clusters have crisp sets for representing membership of an element. The membership of an element in a cluster is represented in terms of binary i.e. either an element belongs or does not belong to the cluster. In fact, crisp clustering (i.e. hard clustering) is the special case of fuzzy clustering in which a particular data-set is assigned a membership of ‘1’ (i.e. data-set belongs to cluster) for the cluster it actually belongs and assigns a membership value ‘0’ (i.e. data-set does not belong to cluster) for all the other remaining clusters. The clustering algorithms namely: k-Means, k-medoid and some of their variations are examples of hard clustering approach.

II.IV.II. Fuzzy Clustering

In traditional clustering (hard clustering) approaches, each pattern belongs to only one cluster. So, the clusters in this approach are disjoint. Whereas, in fuzzy clustering; each pattern belongs to more than one cluster using the degree of membership. Here the degree of membership (i.e. level of membership) indicates the strength of association between an element and the cluster. Thus, fuzzy clustering is the process of allocating the degree of membership according to its (element's) relevance with particular cluster and to use degree of membership for assigning data set to one or more clusters [7]. Fuzzy clustering provides more information about the data sets and overcomes the problem of overlapping. The fuzzy membership functions known as fuzzy sets can be either Type-I, Type-II, or Intuitionistic [8].

Fuzzy C-means (FCM) clustering algorithm is one the famous fuzzy clustering algorithm and some of the other algorithms such as possibilistic C-means known as PCM, conditional fuzzy c-means known as CFCM, FCM–σ, possibilistic fuzzy C-means known as PFCM, and DOFCM (Density Oriented Fuzzy C-Means) come under the head of type-I fuzzy set based clustering algorithms.

Whereas, Type-2 Fuzzy C-Means (T2FCM) clustering algorithm and kernelized type-2 fuzzy c-means (KT2FCM) clustering algorithms come under the head of type-II fuzzy set based clustering algorithms.

As well, Intuitionistic fuzzy C-means (IFCM), IFCM-σ and kernel-based fuzzy c-means (KFCM) clustering algorithms come under the head of intuitionistic fuzzy set based clustering algorithms. A fuzzy clustering can be converted to a hard clustering by assigning each pattern to the cluster with the largest measure of membership.

III. LITERATURE SURVEY

III.I. Hard Partition Clustering Algorithms

Jain, et al. (1999) proposed various types of text clustering algorithms which fall under the main two approaches namely: *hierarchical* and *partitional* approaches. Most of the hierarchical clustering algorithms are variants of single-link (also known as nearest neighbour), complete-link and *minimum-variance* algorithms. Out of these algorithms, single-link and complete-link algorithms are widely used algorithms. On the other hand, *partitional* clustering algorithms are variants of *square-error*, *graph-theoretic*, *mixture-resolving* and *mode-seeking* algorithms. Out of these algorithms, *k-means* algorithm is the simplest and most widely used algorithm employing a squared error criterion [13].

MacQueen, J. (1967) introduced most widely used clustering technique namely k-means clustering algorithm. This algorithm is a squared-error criterion falling under the partitioning clustering approach. The k-mean is non-deterministic, unsupervised, numerical, iterative algorithm. It divides the data objects in 'k' no. of clusters/groups in such a way that intra-cluster will posses high similarity of data and inter-cluster will posses low similarity. The Similarity is measured in term of mean value of data objects in a cluster.

Thus the goal of K-means clustering algorithm is to classify the data objects into various types of clusters in such a way that the data objects (i.e. data-sets) belong to the same cluster consist similar entities and the data objects which belong to the other cluster consist different entities [14].

Although the k-means algorithm consists best of the features like simplicity and acceptable computational time but some of the shortcomings of k-means algorithm are: sensitive to the selection of initial cluster centroids, requirement of specifying the optimal no. of clusters 'k' in advance, it may contain no. of empty clusters. So, it was quite crucial for the k-means algorithm to refine the shortcomings mentioned above. In view of this, several methods have been proposed to enhance the efficiency of k-means clustering algorithm.

Kaufman and Rousseeuw (1987) proposed k-medoids algorithm to overcome the limitations of k-means algorithm. The k-medoids algorithm defines each cluster by the most central medoid in which it is located.

The k-medoids algorithm uses medoids (rather than mean/centroids) to represent the clusters. The medoid is a statistic which represents the object of a particular cluster whose average dissimilarity to all the other objects in the cluster is minimal. Thus, a medoid unlike a mean is always a member of data set. It is the most centrally located data item in the cluster. The remaining steps of this algorithm are same as k-means algorithm.

The k-medoids algorithm has an advantage over k-means algorithm because in k-medoids algorithm, it is not required to calculate the distance between objects during each iteration. In terms of noise also the k-medoids algorithm is more robust in comparative to k-means algorithm. As the k-medoids is working directly with the medoids it is not that much influenced by the outliners like centroids calculation influences the k-means algorithm. Whereas, in terms of computational cost, the k-medoids algorithm incurs higher cost in comparative to k-means algorithm. Due this reason, k-medoids algorithm is efficient for small data sets [15].

Krishna and Murty (1999) proposed genetic algorithm (GA) based k-means to find the global optimal partition of given data set into 'k' no. of clusters. In this method, k-means operator is defined to be used as a search operator rather than as a crossover. In this proposed method, a biased mutation operator was also defined that was specific to clustering called distance based mutation. It purpose of defining the biased mutation operator was to help the k-means algorithm by avoiding local minima. After performing analysis and simulations it was resulted that almost every run of Genetic based k-means algorithm (GKA) converges to a globally optimal partition. The performance of GKA was

also analyzed by comparing with other algorithms, it was turned out that GKA is faster than other algorithms [17].

Aristidis Likasa, Nikos Vlassisb and JakobJ. Verbeekb (2002) presented global k-means algorithm that is known as an incremental approach to the text clustering. This algorithm constitutes an effective clustering by adding one cluster center dynamically for the minimization of the clustering error. For this process it employs the k-means algorithm as a local search process [18].

Arthur and Vassilvitskii (2007) proposed K-means++ initialization algorithm that enhances k-means with a simple, randomized seeding technique. The proposed technique improved speed and accuracy of k-means algorithm. In this technique, an initial set of centroids is obtained that is near optimal. It is done by adopting a way of initializing k-means in such a way that choose random starting centers with the specific probabilities. The inherent sequential nature of the proposed algorithm proved its main drawback because it limits the efficiency for high volume of data [19].

Kwedlo (2011) proposed a hybrid clustering algorithm combining the differential evolution algorithm (DE) with the k-means algorithm. This new clustering algorithm is known as DE-KM algorithm. This hybrid algorithm partitions a dataset into 'k' known no. of clusters using sum of squared errors criterion (SSE). In this algorithm, the resultant clustering solutions were corrected and fine-tuned using k-means algorithm. As result, it was found that the performance of DE-KM clustering algorithm was quiet good as compared to global k-means algorithm, genetic k-means algorithm and other two variants of k-means algorithm. The results also proved that DE-KM algorithm achieves solutions with lower SSE values if the no. of clusters is large [20].

Malinen, Mariescu-Istodor, and Franti (2014) proposed a new clustering approach named k-means* algorithm that generates an artificial dataset X* and fits the input data set into a given clustering model of equal size and dimension. Afterwards, an inverse of transformation of the artificial data set back to the original data set is performed by a series of gradual transformations. The key idea of this algorithm is to perform local fine-tuning of the clustering prototypes during the transformation. The main drawback of this algorithm is as the no. of clusters increases, it decreases the efficiency [21].

M. Ester, H.-P. Kriegel, J. Sander, X. Xu (1996) proposed a new clustering algorithm known as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm. This algorithm relies on density based notion of clusters; designed to find out the clusters of arbitrary shapes. In this proposed algorithm, the point density is attained by calculating the number of points in the specified region around that point. If the points with specified density are met then these points are constructed as clusters. The proposed algorithm has been proved its ability in contrast to other clustering algorithms because it does not require predefined number of clusters. Moreover, it has ability of processing very large databases [22].

Derya Birant, Alp Kut (2007) proposed an improved density based clustering algorithm known as ST-DBSCAN algorithm. This algorithm asserts an improvement over the DBSCAN algorithm in three aspects. First, the proposed algorithm has ability to cluster the spatial temporal data according to its spatial, non-spatial and temporal attributes. Second, the DBSCAN algorithm fails to detect noise points when clusters of different densities are met. The proposed algorithm overcomes this problem by assigning a density factor to each cluster. Third, the proposed algorithm solves the problem of variation in border-object values by comparing the average value of a cluster with new coming value [23].

Brendan and Dueck (2007) proposed a famous clustering algorithm known as Affinity Propagation (AP) i.e. based on probability graph models. The proposed algorithm finds clusters with very low rate of error in comparative to other clustering algorithms. Moreover, this algorithm needs very less time and space for computing and storing the predefined similarity matrix corresponding to the data set. The proposed algorithm simultaneously considers all data points as potential exemplars. Here the exemplars are the centers selected from actual data points [25].

Chen, et al. (2016) proposed a novel clustering algorithm known as CLUB (CLUstering based on backbone) to determine the optimal clusters. The proposed algorithm defines initial clusters based on 'K' Nearest Neighbors (KNN) method. Then the algorithm identifies the density backbones of clusters by taking initial clusters as an input. Finally, by assigning unlabeled point to the cluster with nearest higher density neighbor, the algorithm presents the final clusters.

The proposed algorithm has several drawbacks; as this algorithm uses KNN method that is not an efficient method to determine the k number of nearest neighbors. Moreover, this algorithm incurs more computational cost [26].

Kewen Chen, Zuping Zhang, Jun Long, Hao Zhang (2016) proposed a new term weighting scheme for text classification named as term frequency & inverse gravity moment (TF-IGM). The proposed text classification scheme is an alternative to the traditional text classification method named as TF-IDF (term frequency & inverse document frequency) i.e. not fully effective for text classification. The proposed method precisely measures the classes of text. The experimental results of the proposed method prove

outperformance in contrast to the famous TF-IDF and the state of the art supervised term weighting schemes [48].

Liangxiao Jiang at el. (2016) proposed an efficient feature weighting clustering algorithm known as deep feature weighting (DFW) that estimates the conditional probabilities of naïve Bayes by deeply computing feature weighted frequencies from the data sets. The experimental results of the proposed method show that it outperforms in contrast to the famous and standard naïve Bayes and achieved remarkable improvements [49].

Emre Gungor, Ahmet Ozmen (2017) proposed a new clustering algorithm known as Gaussian Density Distance (GDD) clustering algorithm i.e. based on the properties of sample space namely distance properties and density properties. The proposed clustering algorithm finds the best possible clusters without any prior information. Moreover this algorithm defines the clusters very close to human clustering perception [50].

Sharma & Gupta (2013) proposed Punjabi document clustering system which uses the karaka list to analyze the semantic relationship between words in a sentence. The karka list includes grammatical connectors for connecting nouns, pronouns, and verbs in a sentence; captures the semantic relations in a sentence. The proposed work is a first ever attempt in this direction to provide a solution for the text clustering of Punjabi documents [51].

III.II. Soft Partition (fuzzy) Clustering Algorithms

Anjana Gosain, Sonika Dahiya (2016) reviewed all major fuzzy clustering algorithms. This paper presents the performance and experimental analysis of all the fuzzy clustering algorithms. In this paper, the fuzzy membership sets are represented which can be either of Type-I, Type-II, or Intuitionistic [9].

III.II.I. Type-I fuzzy set based clustering algorithms

Bezdek, Ehrlich, & Full (1984) proposed FCM (fuzzy c-means) algorithm that is widely used fuzzy clustering technique in the field of data mining. In fuzzy clustering technique, a fuzzy partition is generated with the pre-defined number of clusters. In this method, degree of membership is expressed to each data object in a given cluster. So here, each data set belongs to all the clusters with its own degree of belonging i.e. degree of membership. The degree of membership of the data sets is set in the interval [0, 1] which means the degree to belong to the clustering center. In fuzzy clustering, the '0' value of degree of membership means that data set is not a member of fuzzy set. Whereas, the '1' value of degree of membership means that data set is fully a member of the fuzzy set. The partial value of degree of membership i.e. the value between '0' and '1' denotes that the data set partially belongs to fuzzy set. The fuzzy c-means algorithm proved more reliable and robust in comparative to other crisp clustering algorithms but with the inherent hindrance more computation time than other clustering techniques [16].

Krishnapuram and Keller (1993) proposed PCM (possibilistic C-means) algorithm that overcomes the problem of FCM in respect to handling outlier points (i.e. noise points). The proposed algorithm is based upon theory of possibility where the resulting values are represented in terms of degree of possibility of the points belonging to particular classes. The PCM algorithm improves the efficiency by identifying outlier points and by eliminating the outlier points [10].

R. Krishnapuram (1994) proposed important properties of the possibilistic clustering and explored how to approximate natural groupings of the data sets using parametric functions. The proposed approach determines the number of membership functions and the number of good clusters. This approach overcomes the traditional tedious process of determining number of clusters [11].

Pal and Bezdek (1997) proposed fuzzy possibilistic C-means (FPCM) algorithm. The proposed algorithm integrates the features of FCM and PCM in order to enhance the clustering model. This algorithm overcomes the coincident clustering problem of PCM algorithm by using the fuzzy values of the FCM and the typicality values of the PCM algorithm. This algorithm also solves the noise sensitivity deficiency of FCM but the noisy data influences the estimation of centroids [12].

Heiko Timm, Christian Borgelt, and Rudolf Kruse (2004) proposed an Extension of possibilistic fuzzy cluster algorithm that overcomes drawback of the conventional approach. In possibilistic fuzzy clustering the objective function is minimized only if all the cluster centers are identical. This undesired property is conquered by introducing a mutual repulsion of the clusters [24].

W.L. Cai, S.C. Chen and D.Q. Zhang (2007) proposed generalized fuzzy c-means clustering algorithms (FGFCM). The execution time of the proposed algorithm is fast. This algorithm is also more robust clustering algorithms that incorporate local information. This algorithm enhances the performance of clustering [27].

Krinidis et al. (2010) proposed a clustering algorithm known as robust fuzzy local information c-means (FLICM). In this algorithm, the local spatial information and gray level information are incorporated in fuzzy way. This algorithm overcomes the drawback of fuzzy c-means algorithm (FCM) and enhances the performance of clustering.

The proposed two above mentioned algorithms limit the flexibility of FCM algorithm because both of these algorithms need number of clusters to be specified in advance. But practically, usually it is not known well in advance about the number of clusters to be formed [28].

Li et al. (2012) proposed chaotic particle swarm fuzzy clustering (CPSFC) algorithm i.e. combination of chaotic particle swarm optimization (CPSO) method, gradient method and chaotic local search. The newly proposed algorithm provides best clustering results. To prove its superiority over the FCM clustering algorithm an experiment is demonstrated on several artificial and real data sets. The proposed algorithm overcomes the major limitation of FCM algorithm i.e. getting stuck at locally optimal values [29].

Du-Ming Tsai, Chung-Chan Lin (2011) proposed a new distance metric based clustering algorithm known as FCM-sigma (FCM-□). The proposed algorithm is an improvement in the effectiveness of FCM clustering algorithm. In FCM clustering algorithm, the conventional distance metric evaluates only the distance between two individual data points. Whereas, in the proposed algorithm, a new distance metric is introduced. The proposed algorithm, overcomes the problem of conventional FCM algorithm by concentrating on the global distance variation for all data points in a cluster. Whereas, conventional FCM clustering algorithm ignores the global distance variation for all data points in a cluster [30].

Du-Ming Tsai, Chung-Chan Lin (2011) also proposed a revised distance metric by introducing distance variation as the regularization in the mapped feature space. The proposed algorithm is known as KFCM-□ clustering algorithm. The proposed algorithm with the new distance metric proves better results in comparative to the conventional KFCM in feature space. Also the proposed algorithm performs extremely well for linearly non-separable data [30].

Witold Pedrycz (1995) proposed fuzzy C-means based clustering algorithm guided by a conditional variable known as conditional fuzzy c-means (CFCM) clustering algorithm. The proposed clustering algorithm is illustrated with the use of numerical illustrations representing its usefulness. Also it is emphasized that other approaches like fuzzy c-lines and FCM can be conditionalized [31].

Mainly the proposed algorithm focused on working efficiently with noisy data. The proposed algorithm reduced the effect of outliers on the resulting clusters. The proposed algorithm failed to find the accurate clusters for the data sets.

Prabhjot Kaur and Anjana Gosain (2010) proposed Density Oriented Fuzzy C-Means (DOFCM) algorithm that can detect the efficient clusters in the presence of outliers and noise. The proposed algorithm identifies the outliers and noise from the data set. In this algorithm, 'n+1' clusters are created with 'n' good clusters and one consisting noise as well as outliers. In this algorithm, density oriented approach is used for identifying the outliers and the clustering is performed considering the actual data points. Thus it results into more accurate clusters [32].

III.II.II. Type-II Fuzzy Set Based Clustering Algorithms

Rhee and Hwang (2001) proposed Type-2 Fuzzy C-Means clustering algorithm known as T2FCM. The proposed algorithm is an extension of the conventional fuzzy C-means algorithm. In this algorithm, the concept of type-2 fuzzy sets has been used. In this algorithm, the membership functions for the patterns are extended as type-2 membership sets by assigning membership grades to type-1 membership sets. The proposed algorithm has been proved an effective method for spherical structured datasets, but failed for the non-spherical and complex structured datasets [33].

Prabhjot Kaur, et al. (2011) proposed a novel kernelized type-2 fuzzy c-means (KT2FCM) clustering algorithm that is a generalization of conventional type-2 fuzzy c-means (T2FCM) clustering algorithm. This algorithm extends T2FCM (i.e. type-2 fuzzy C-means) by adopting a kernel induced metric in the data sets to replace the original Euclidean norm metric. Thus KT2FCM clustering algorithm overcomes the problem of T2FCM and gives the better segmentation of data sets over noisy data [34].

Aliev et al. (2011) proposed type-2 fuzzy inference system using type-2 clustering and fuzzy neural network based refinements. It is analyzed that in comparative to type-1 inference process, type-2 inference performs the best and yields higher accuracy. Moreover, type-2 inference system is capable of dealing with higher levels of uncertainty. Type-2 fuzzy neural networks have advantageous features of fuzzy clustering by defining small number of if-then rules. The comparative analysis of experimental results demonstrated quantified performance [35].

Ondrej Linda, Milos Manic (2012) proposed a novel approach for uncertain fuzzy clustering using general type-2 fuzzy C-means (GT2FCM) algorithm. The proposed algorithm overcomes the uncertainty associated with the FCM algorithm for choosing the fuzzification parameter. This algorithm performs well in the case of noisy data and clusters the variation of density and weights [36].

Zarandi, Gamasaee, Turksen (2012) proposed more reliable type-2 C-regression clustering algorithm. The proposed algorithm is an extended version of type-1 fuzzy C-regression clustering algorithm performed using interval type-2 fuzzy model. The experimental results show that the proposed algorithm is more effective and leads to better solutions and with less error [37].

Golsefid and Zarandi (2015) proposed dual-centers type-2 fuzzy clustering algorithm. The proposed algorithm is based on dual centers instead of single center and the degree of membership for the concerned cluster is described with lower and upper membership values. In this algorithm, the membership values of data sets in a cluster are defined by type-2 fuzzy sets and there is no any de-fuzzification process. The experimental results represents that the proposed algorithm determines the optimum number of clusters [38].

Golsefid and Zarandi (2016) also proposed multi-central general type-2 fuzzy clustering algorithm. The proposed algorithm mainly focuses on uncertainty associated with the cluster centers. In this algorithm, a data set is considered as the center of every cluster and uses type-2 fuzzy sets for defining the membership functions including primary and secondary variables. Also in this algorithm there is no any de-fuzzification process. The compatible indexes are defined for validation and verification of the clustering. The algorithm is experimented to evaluate the performance and demonstrated improved performance for defining clusters in comparative to other methods [39].

Sarkar, Saha and Maulik (2016) proposed a hybrid clustering technique with the use of type-2 fuzzy clustering algorithm. The proposed algorithm is known as rough possibilistic type-2 fuzzy C-means clustering method (RPT2FCM) with the integration of random forest. The rough set based clustering algorithm handles uncertainties and noisy data more efficiently. In this algorithm, various uncertainties and vagueness of data are handled by type-2 fuzzy sets and by rough set theories. The performance of this algorithm has been analyzed through the statistical significance test and demonstrated better performance in comparative to other methods [40].

III.II.III. Intuitionistic Fuzzy Set Based Clustering Algorithms

Zeshui and Junjie (2010) proposed Intuitionistic fuzzy C-means (IFCM) clustering algorithm to cluster intuitionistic fuzzy sets (IFS). The proposed algorithm is based on well known FCM algorithm and the basic distance measures between IFSs. In this algorithm, for each IFS, degree of membership for each cluster is estimated. Thus all the IFSs are clustered according to the estimated degree of membership. The experiments performed on both the real and simulated data sets result to more desirable advantages. Moreover, the IFCM algorithm has lower computational complexity than the other algorithms like FCM [41].

Prabhjot et al. (2011) presented Intuitionistic Fuzzy c-means (IFCM-σ) and kernel Intutitionistic Fuzzy C-Means(KIFCM-σ) clustering algorithms. The proposed algorithms avoid various problems of IFCM and KIFCM as well as provide the effective results. The experimental results represent that the proposed algorithms are highly robust to noise and outliers [42].

Tamalika Chaira (2011) presented intuitionistic fuzzy C-means clustering algorithm that uses intuitionistic fuzzy set theory. In the proposed method, performance of obtaining cluster centers is more desirable than obtaining the cluster centers using FCM clustering algorithm. The experimental results of proposed algorithm are proved more efficient in contrast to conventional fuzzy C-means and type-II fuzzy set based algorithms [43].

Dawaei et al. (2013) proposed a spectral data clustering technique using intuitionistic fuzzy information. In this paper, two new intuitionistic fuzzy similarity measures are defined for building an intuitionistic fuzzy similarity measure matrix. In this algorithm, it is found that the proposed algorithm can achieve more detailed clustering results as well as global optimal solution [44].

Kuo-Ping Lin (2014) proposed a novel data clustering algorithm named evolutionary kernel intuitionistic fuzzy c-means clustering algorithm (EKIFCM). The proposed algorithm combines intuitionistic fuzzy sets (IFSs) with kernel-based fuzzy c-means (KFCM), and genetic algorithms (GA) for the selection of parameters of EKIFCM. The experimental results presented in this paper prove that that proposed method performs effectively in comparison to other conventional clustering methods like as k-means, FCM, IFCM and KFCMG [45].

Zhong Wang et al. (2014) proposed clustering analysis technique under the intuitionistic fuzzy environment. In the proposed technique, an intuitionistic fuzzy implication operator is developed. In this technique an intuitionistic fuzzy triangle product and an intuitionistic fuzzy square product are also defined. Based on the proposed triangle product and square product, a direct method for intuitionistic fuzzy clustering analysis is developed. The proposed technique needs less calculation efforts and achieves the desirable clustering results [46].

Hanuman Verma, R. K. Agrawal and Aditi Sharan (2016) proposed a novel clustering approach using intuitionistic fuzzy named as an improved intuitionistic fuzzy c-means (IIFCM) clustering algorithm. The proposed algorithm uses the local spatial information in an intuitionistic fuzzy way. A non-parametric statistical analysis performed on the proposed algorithm shows its significant performance in comparison to other existing segmentation algorithms [47].

IV. CONCLUSION

In this paper, the authors have analyzed various types of text/document clustering techniques by arranging them in

various categories. After going through the detailed study, it is concluded that a lot of research work has been done on clustering of English text/documents. Much more research work is needed on clustering of documents specifically in vernaculars like Gurumukhi script (Punjabi language). As per review of literature done to carry out this study, a first ever attempt in this direction (i.e. clustering of Punjabi text/documents) has been done by Sharma and Gupta using hybrid approach. So, there is an urgent need to develop better machine learning based Gurmukhi script document clustering techniques so as to improve the performance of existing clustering technique.

REFERENCES

[1]. S.Prabha, K.Duraiswamy, M.Sharmila, "*Analysis of Different Clustering Techniques in Data and Text Mining*" International Journal of Computer Science Engineering (IJCSE), Vol. 3, PP. 107-116, 2014.

[2]. Han, J., Kamber, M., & Tung, A. K., "*Spatial Clustering Methods in Data Mining: A Survey*", Geographic Data Mining and Knowledge Discovery, Taylor and Francis, PP. 1–29, 2001.

[3]. Ester, M., Kriegel, H., Sander, J., & Xu, X., "*A Density-Based Algorithm for Discovering Clusters in Large Spatial Database with Noise*", Second International conference on Knowledge Discovery and Data Mining, Portland, PP. 226-231, 1996.

[4]. Gholamhosein Sheikholeslami, Surojit Chatterjee, Aidong Zhang, "*Wave Cluster: a wavelet-based clustering approach for spatial data in very large databases*", the VLDB Journal, Vol. 8, PP. 289-304, 2000.

[5]. Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan, "*Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications*", Data Mining and Knowledge Discovery, PP. 5-33, 2005,.

[6]. Michael Steinbach, George Karypis Vipin Kumar, "*A Comparison of Document Clustering Techniques*", Proc. Knowledge Discovery and Data Mining (KDD) Workshop Text Mining, PP.1-20, 2000.

[7]. Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn, "*Data clustering: a review*" *ACM computing surveys (CSUR,)* No. 3, PP. 264-323, 1999.

[8]. Anjana Gosain, Sonika Dahiya, "*Performance Analysis of Various Fuzzy Clustering Algorithms: A Review*", Prodedia Computer Science 79, Elsevier Science Ltd.,Vol. 79, PP. 100-111, 2016.

[9]. Anjana Gosain, Sonika Dahiya, "*Performance Analysis of Various Fuzzy Clustering Algorithms: A Review*", Prodedia Computer Science 79, Elsevier Science Ltd., PP. 100-111, 2016.

[10]. R. Krishnapuram, J.M. Keller, "*A possibilistic approach to clustering*", IEEE transactions on fuzzy systems, vol. 1, issue: 2, 1993.

[11]. R. Krishnapuram, "*Generation of membership functions via possibilistic clustering*" published in Fuzzy Systems, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the Third IEEE Conference on, 1994.

[12]. Nikhil R. Pal, Kuhu Pal, James M. Keller, and James C. Bezdek, "A Possibilistic Fuzzy c-Means Clustering Algorithm", IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, 2005.

[13]. Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn, "*Data clustering: a review*" *ACM computing surveys (CSUR,)* No. 3, PP. 264-323, 1999.

[14]. MacQueen, J., "*Some methods for classification and analysis of multivariate observations*", Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probabilities, 1, 281-296, 1967.

[15]. Kaufman L, Rousseeuw PJ, "*Clustering by means of medoids*". In: Dodge Y, editor. Statistical data analysis based on the L1 norm and related methods. Amsterdam: North Holland/Elsevier. pp. 405–416, 1987.

[16]. Bezdek, J. C., Ehrlich, R., & Full, W., "*FCM: The fuzzy C-means clustering algorithm*". Computers & Geosciences, 10 (2–3), 191–203, 1984.

[17]. Krishna, K., & Murty, M. N., "*Genetic K-means algorithm*". Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 29 (3), 433–439, 1999,

[18]. Aristidis Likasa, Nikos Vlassisb and JakobJ. Verbeekb, "*The global k-means clustering algorithm*", Pattern Recognition Society, Elsevier Science Ltd. 36, 451 – 461, 2002.

[19]. Arthur, D., & Vassilvitskii, S., "*K-means ++ : The advantages of careful seeding*". In Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms pp. 1027–1035, 2007.

[20]. Kwedlo, W., "*A clustering method combining differential evolution with the K-means algorithm*". Pattern Recognition Letters, Elsevier Science Ltd. 32 (12), 1613–1621, 2011.

[21]. Malinen, M. , Mariescu-Istodor, R. , & Fränti, "*K-means*: Clustering by gradual data transformation*". Pattern Recognition, 47 (10), 3376–3386, 2014.

[22]. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, "*A density-based algorithm for discovering clusters in large spatial databases with noise*", in: Proceedings of Second International Conference on Knowledge Discovery and Data Mining, Portland, OR 1996, pp. 226–231, 1996.

[23]. Derya Birant, Alp Kut, "*ST-DBSCAN: An algorithm for clustering spatial–temporal data*" Data & Knowledge Engineering, Elsevier Science Ltd. 60, 208–221, 2007.

[24]. Heiko Timm, Christian Borgelt, and Rudolf Kruse, "*An Extension of Possibilistic Fuzzy Cluster Analysis*" Fuzzy Sets and Systems, Elsevier Science Ltd. Volume 147, Issue 1, 1 October 2004, Pages 3-16, 2004.

[25]. Brendan J. Frey and Delbert Dueck, "*Clustering by Passing Messages Between Data Points*" Elsevier Science Ltd., vol. 315, pp. 972-976, 2007.

[26]. Chen, M., Li, L., Wang, B., Cheng, J., Pan, L., & Chen, X., "*Effectively clustering by finding density backbone based-on kNN*". Pattern Recognition, Elsevier Science Ltd. 60, 486–498, 2016.

[27]. W.L. Cai, S.C. Chen, D.Q. Zhang, "*Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation*", Pattern Recognition, Elsevier Science Ltd. 40 (3) 825–838, 2007

[28]. S. Krinidis, V. Chatzis, "*A Robust fuzzy local Information C-means clustering Algorithm*", IEEE Trans. Image Process. 19 (5) 1328–1337, 2010.

[29]. Li, C., Zhou, J., Kou, P., Xiao, J., "*A novel chaotic particle swarm optimization based fuzzy clustering algorithm*". Neurocomputing, Elsevier Science Ltd. 83, 98–109, 2012.

[30]. Du-Ming Tsai, Chung-Chan Lin, "*Fuzzy C-means based clustering for linearly and non linearly separable data*". Pattern Recognition, Elsevier Science Ltd. 44, 1750–1760, 2011.

[31]. Witold Pedrycz, "*Conditional Fuzzy C-Means*" Pattern Recognition Letters, Elsevier Science Ltd., Vol. 17 PP. 625-631, 1996.

[32]. Prabhjot Kaur and Anjana Gosain, "*Density-Oriented Approach to Identify Outliers and Get Noiseless Clusters in Fuzzy C – Means*", Fuzzy Systems (FUZZ), 2010 IEEE International Conference on, 2010.

[33]. Rhee, Frank Chung Hoon, and Cheul Hwang. "*A type-2 fuzzy C-means clustering algorithm.*" In IFSA World Congress and 20th NAFIPS International Conference, 2001. IEEE, Joint 9th, vol. 4, pp. 1926-1929, 2001.

[34]. Prabhjot Kaur, Dr. I. M. S. Lamba, Dr. Anjana Gosain, "*Kernelized Type-2 Fuzzy C-means Clustering Algorithm in Segmentation of Noisy Medical Images*" Recent Advances in Intelligent Computational Systems (RAICS), IEEE, 2011.

[35]. Rafik A. Aliev, Witold Pedrycz, Babek G. Guirimov, Rashad R. Aliev, Umit Ilhan, Mustafa Babagil, Sadik Mammadli, "*Type-2 fuzzy neural networks with fuzzy clustering and differential evolution optimization*", Information Sciences, Elsevier Science Ltd. 181 1591–1608, 2011.

[36]. Ondrej Linda, Milos Manic, "*General Type-2 Fuzzy C-Means Algorithm for Uncertain Fuzzy Clustering*", IEEE Transactions on Fuzzy Systems, Vol. 20, No. 5, 2012.

[37]. M.H. Fazel Zarandi, R. Gamasaee, I.B. Turksen, "*A type-2 fuzzy c-regression clustering algorithm for Takagi–Sugeno system identification and its application in the steel industry*", Information Sciences, Elsevier Science Ltd. 187,179–203, 2012.

[38]. Indices Samira Malek Mohamadi Golsefid, Mohammad Hossein Fazel Zarandi, "*Dual-centers type-2 fuzzy clustering framework and its verification and validation indices*" Applied Soft Computing, Elsevier Science Ltd. 1568-4946, 2015.

[39]. S. Malek Mohamadi Golsefid, M.H. Fazel Zarandi, I.B. Turksen, "*Multi-central general type-2 fuzzy clustering approach for pattern recognitions*", Information Sciences, Elsevier Science Ltd. Vol. 328, PP 172–188, 2016.

[40]. Jnanendra Prasad Sarkar, Indrajit Saha, Ujjwal Maulik, "*Rough Possibilistic Type-2 Fuzzy C-Means clustering for MR brain image segmentation*" Applied Soft Computing, Elsevier Science Ltd. Vol. 46, PP 527–536, 2016.

[41]. Zeshui Xu and Junjie Wu "*Intuitionistic fuzzy C-means clustering algorithms*", Journal of Systems Engineering and Electronics, IEEE, Vol. 21, Issue: 4, 2010.

[42]. Prabhjot Kaur, Dr. A. K. Soni, Dr. Anjana Gosain, "*Robust Intuitionistic Fuzzy C-Means Clustering for linearly and nonlinearly Separable Data*", International Conference on Image Information Processing (ICIIP 2011), IEEE, 2011.

[43]. Tamalika Chaira, "*A novel intuitionistic fuzzy C-means clustering algorithm and its application to medical images*", Applied Soft Computing, Elsevier Science Ltd., Vol. 11, PP. 1711–1717, 2011.

[44]. Dawei Xu, Zeshui Xu, Shousheng Liu, Hua Zhao, "*A spectral clustering algorithm based on intuitionistic fuzzy information*", Knowledge-Based Systems, Elsevier Science Ltd., Vol. 53, PP. 20–26, 2013.

[45]. Kuo-Ping Lin, Member, IEEE, "*A Novel Evolutionary Kernel Intuitionistic Fuzzy C-means Clustering Algorithm*", IEEE Transactions on Fuzzy Systems, Vol. 22, No. 5, 2014.

[46]. Zhong Wang, Zeshui Xu, Shousheng Liu, Zeqing Yao, "*Direct clustering analysis based on intuitionistic fuzzy implication*", Applied Soft Computing, Elsevier Science Ltd., Vol. 23, , PP 1–8, 2014.

[47]. Hanuman Verma, R. K. Agrawal and Aditi Sharan, "*An Improved Intuitionistic Fuzzy C-means Clustering Algorithm Incorporating Local Information for Brain Image Segmentation*", Applied Soft Computing, Elsevier Science Ltd., Vol. 46, PP. 543–557, 2016.

[48]. Kewen Chen, Zuping Zhang, Jun Long, Hao Zhang, "*Turning from TF-IDF to TF-IGM for term weighting in text classification*" Expert Systems with Applications, Elsevier Science Ltd., Vol. 66. PP. 245–260, 2016.

[49]. Liangxiao Jiang, ChaoqunLi, ShashaWang, LunganZhang, "*Deep feature weighting for naïve Bayesand its application to text classification*", Engineering Applications of Artificial Intelligence, Elsevier Science Ltd., Vol. 52, PP. 26–39, 2016.

[50]. Emre Gungor, Ahmet Ozmen, "*Distance and density based clustering algorithm using Gaussian kernel*", Expert Systems With Applications, Elsevier Science Ltd., Vol. 69, PP. 10–20, 2017.

[51]. Saurabh Sharma, Vishal Gupta, "*Punjabi Documents Clustering System*", Journal of Emerging Technologies in Web Intelligence, Vol. 5, No. 2, May, 2013.

Authors' Profile

Mr. Mukesh Kumar is working in PG Department of Computer Science, Mata Gujri Autonomous College, Fatehgarh Sahib, PUNJAB. He is currently pursuing Ph.D. from Punjabi University, Patiala. His area of research is document clustering for Gurumukhi script. He has published number of books with Kalyani Publishers. He has 16 years of teaching experience.

Dr. Amandeep Verma is working as Assistanct Professor in PG Department of Computer science, Punjabi University Regional Centre for Information Technology and Management, Mohali PUNJAB. He completed Ph.D. from Punjabi University, Patiala. His areas of research are: Ontological Engineering and Formal Methods, Image Processing, Natural Language Processing, and Cloud Computing. He has published number of research papers in reputed International journals.

