A Review of Clustering Methods and its Challenges in Document Clustering

[1]P Gopala Krishna, [2]D LalithaBhaskari

[1]Associate Professor, [2]Professor
[1]Dept of IT,[2] Dept of CS&SE,
[1]Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India
[2]Andhra University College of Engineering (A), Andhra University, Visakhapatnam, India

***Abstract*—The explosive growth of the Internet and its services with massive amounts of text documents demands for collecting most similar documents together for diverse applicationsdrawing the attention of researchers in this area.Document clustering can simplify the works of document organization needed for various data mining activities for the search engines, data filtering, information classification etc. However, there have been some attempts to build up well-organized document clustering algorithms, but most clustering methods have had difficulty dealing with high dimensional, scalability, accuracy, and important cluster label problems. It contributes a significant function in "machine learning, data mining, information retrieval and pattern recognition". This paper presents a reconsider on different clustering methods in data mining initially. It will focus on the documents clustering mechanism, its applicability, and limitations in documents clustering. Finally, we consider the related works proposed on document clustering and its outcomes.**

***IndexTerms*—Clustering methods, Document Clustering, Data Mining.**

I. INTRODUCTION

As information about the Internet grew, web mining was the focus of information retrieval. The Internet of the day is currently utilized broadly and leads to large document repositories. Existing information is accumulated in a text database, which is a group of documents commencing diverse resources such as:"news articles, research papers, books, digital libraries, e-mail messages, blogs, and web pages". The concept of information, and consequently the transfer of information, has transformed significantly over the past few decades[1], [2] [3], [4].This is for the reason that it needs to "know", "collaborate" and "contribute" is still being recognized and all of these tasks involve in information requirement. However, it has become clear that the awareness must grow continuously and the corresponding information increase should organize the information so that everyone can easily access various types of information effortlessly. The word "organize" means establishing anorder between various information sources.

Clustering is the basic task of data mining, which is often used in unsupervised learning methods [6], [7], [8]. The purpose of clustering is to determine a novel group of categories, so the novel group is interested in itself and the assessment is essential. Data instances are grouped into small groups in such a way that related instances are clustered, but other instances belong to diverse groups. It is an unchecked classification process that groups objects into "classes" or "clusters" so that objects contained by the same cluster are extremely related to objects of other classes and extremely related to every one other. Therefore, clustering can efficiently visualize documents in collections by grouping similar and relevant documents in one group or cluster.Unlike classification, clustering must not be mystified with classification. Classified documents are not provided in clustering [4], [10] and should not be confused. The matter of clustering has been extensively studied in many fields and is being actively studied in the area of data mining.

Document Clustering is an automated document composition and text extraction for fast information retrieval. Automatic clustering is a widely applied method for analyzing information because it requires no human intervention and does not require prior knowledge of thetext. It is not just that a group of documents is more similar and different groups of documents constitute a document in meaningful groups in different ways [6]. The similarity between documents can be measured using similarity measures [11]. Clustering of uncertain data, one of the critical responsibilities of uncertain data mining, poses a significant challenge in modeling uncertainty between individuals and developing proficient computational methods.

Document clustering is able to be inspected from dissimilar viewpoints, depending on the document representation, the processing of the document, the method used, and the method used in the application. From the perspective of the information retrieval (IR) community and to a smaller coverage the machine learning community, traditional methods for document representation are used, and there is a significant trend towards vector space models.

The remaining paper organized as, Section-II discuss the various clustering methods, Section-III discuss the various document clustering techniques and its applicability in real-time, Section-IV discuss the challenges in document Clustering and Section-V presents the conclusion of the paper.

II. STUDY OF CLUSTERING METHODS

Clustering is the generally frequent outline of self-learning and is a key dissimilarity among "clustering" and "classification". The absence of a supervision means that no specialist has assigned the document to the class. In clustering, determining the cluster affiliated is the distribution and configuration of data. Clustering is often misrepresented as "automatic classification". However, the clusters established are not identified before processing, this is not accurate because the class is predefined in the situation of classification.

In clustering, the classifier determines the allocation and nature of the data that determines cluster association against classifications that learn the relationship among "objects" and "classes" which is called "training sets", *i.e.*a data collection is properly marked with hand, and afterwardit imitates the behavior learned on unlabeled data.

Hierarchical Methods

This methodconstructs a cluster by "recursively partitioning" the instances moreover "top-down" or "bottom-up" [29]. These methods are separated into the subsequent:

- *Agglomerative hierarchical clustering*: Every object is primarily represented by its own cluster. These clusters are merged sequentially until they receive the required cluster formation.
- *Divisive hierarchical clustering*: The entire objects related to individual cluster at the beginning. Later the cluster is separated into sub-groups, which are successfully separated into their sub-groups. This process will continue till it receives the needed cluster composition.

The outcome of "hierarchical methods" is a "dendrogram", that represents a collection of objects and resembles the group's change. A collection of data objects can be acquired by splitting the dendrogram at the required resemblance level of comparison.

Generally, thehierarchical methods are featured with the subsequent advantages:

- *Adaptability*: It is " single-link methods", it keeps the high-quality performance of data collections with non-isotropic clusters, together with well-separated "chain-like" and "dense" clusters.
- *Several partitions*: The "hierarchical methods" are not a partition, but many local partitions agree to dissimilar users to decide dissimilar partitions with respect to the needed resemblance stage. The hierarchical distribution is provided by means of the "dendrogram".

The foremost shortcomings of "hierarchical methods" are being identified as, "Incapability to scalability" and "time complexity", it means that the hierarchical protocols should have "$O(m^2)$" process,where m is the "total number of instances occurrences" which will be not linear with the number of objects. Even this methods can never be completed to restore what was previously done. That means there is no surveillance capability to this method.

Partitioning Methods

Partitioning methods move instances with changing them from one cluster partition to another partition [22]. Such methods naturally necessitate that the number of clusters is preset by the user. To accomplish overall most favorable in partitioned clustering, a complete details procedure of the entire probable partitions is necessary. Since this is not possible, assured "greedy heuristics" are utilized in the structure of an iterative for most advantageous. A "displacement method"repeatedly moves the points among the k-clusters.

The "K-means algorithm" [7] can be seen as a "gradient-descent procedure" that commences through a preliminary collectionof "K-cluster" centric and repeatedly revises to reduce the inaccuracy function. This "linear complexity" is the causes in favor of the recognition of the "K-means algorithm". Since, if the numerous of instances is very huge, this algorithm is computationally prominent. Thus, the "K-means algorithm" has advantages over other clustering methods with non-linear complexity. The algorithm also distinguishes between noisy data and anomalies. Applies only if an average is defined that is, for numeric attributes. If it does not contain the prior knowledge, you need a pre-cluster number that is not trivial.

Another partitioning algorithm that challenges to decrease the "Sum of Squared Error (SSE)" is the "K-medoids". This algorithm is extremely related to the "K-means algorithm". It diverges from last mostly in the demonstration of other clusters. Every one cluster is symbolized as the main central object in the cluster, not the restricted meaning that it might not related to a cluster. The "K-medoidsmethod" is further powerful than the "K-means algorithm" when there are noises and outliers because the method is less affected with outliers or other intense assessments than the average. However, the process is added expensive than the "K-means method". In both methods, the consumer has to indicate "K", as the number of clusters required [9].

Density-based Methods

The "density-based method" believes that the points related to every cluster are extracted from a particular "probability distribution" [28]. The total division of data is supposed to be a combination of various divisions. The purpose of this method is to recognize the cluster and its division parameters. Much of the work in this area is supported on the hypothesis that component density is "multivariate Gaussian" or"multiple nominal". The satisfactory explanation in this situation is to utilize the "maximum likelihood principle". In case of to this standards, ithas to decide the clustering organization and factors to maximize the possibility of the data created with these clustering structures and factors.

The "expectation maximization (EM)" algorithm, a universal maximum probability algorithm for misplaced data difficulties, thathas been practical to the parameter estimation problem. The "DBSCAN algorithm" a "density-based spatial clustering" of noisy functions finds clusters of any type and is well-organized for bulky "spatial databases". The algorithm looks for the database for contiguous areas of each object to search the cluster and make sure that it contains at least the least amount number of objects. Density-based clustering can also use "non-parametric methods" such as retrieving a huge number of stores from a "multidimensional histogram" of the input instance space.

Model-Based Methods

This method attempts to improve the suitability of certain data and a few "mathematical models" [5]. In contrasting traditional clusters that define collections of objects, the "model-based clustering method" as well describes properties for every group in which every group characterizes a conception or category. The commonly utilized methods are,"induction methods", "decision trees" and "neural networks".

In "Decision Trees", the data is symbolized by a "hierarchical tree", where every node referring to a "perception and a probabilistic" explanation of this concept. The "Neural Networks algorithm" characterizes every cluster with a "neuron" or "prototype". The entered data are to correspond to by "neurons" associated to prototype neurons. Every such link has an influence that is adaptively learned in the period of learning. An extremely well-liked "neural algorithm" in support of clustering is the "self-organizing map (SOM)".

Soft-computing Methods

In general conventional clustering methods create partitions. In a partition, each instance related to merely single cluster. Therefore, clusters in hard clustering are disconnected. The extensively utilized "fuzzy based clustering algorithm" [8] is the "fuzzy C-means (FCM)" algorithm. The intent of associated purposes is the main significant difficulties in fuzzy clustering. Other preferences include selection supported on "similarity decomposition" and "cluster centering". The simplification of the "FCM algorithm" is recommended through the objective function group. The "Fuzzy C-shell algorithms" and "adaptive variations" for identifying "circular" and "elliptical boundaries"are includedin the proposals [15]. It has been observed that the performance is enhanced than the "*K*-means algorithm" and the "fuzzy c-means algorithm". However, all these advance are overly responsive to parameters. Therefore, for every particular difficulty, the user must adjust the factor values to ensemble the application.

III. DOCUMENT CLUSTERING AND ITS APPLICABILITY

Document clustering techniques [2], [3], [16], [21]maximize the distance linking clusters with using appropriate distance measurements among documents while minimizing the inter-cluster distances between documents. So distance measurement [20] or similarity measurement is at the core of document clustering. With a diversity of documents, it is approximately unfeasible to generate a generic algorithm that works best with all kinds of data sets. Document clustering is a small group of data clustering in data mining technology that includes the concepts of "information retrieval", "natural language processing", and "machine learning". Document clustering consists of multiple groups of documents, called clusters, where documents in every one cluster contribute to a common attribute according to a "defined similarity measures". The "Fast" and"high-quality document clustering algorithms"describes an essential responsibility in serving users for "navigation", "summarization", and "organize information effectively" [17].

It is important to emphasize that from document collections to clustering of collections is not simply a single task. It includes several steps. It normally consists of three most important steps for doing the clustering as: "Document Representation", "Feature Extraction and Selection", and "Similarity Clustering".

Document Representation

The most widespread approach to characterize a document is a collection of "keywords", where the keywords can be simple words or phrases, such as using "part-of-speech tagging", named "object recognition", and so on. In some cases, the document will also be represented as a "vectors of features", which can be the name of an entity, place, and so on. It can be represented by other form models as described below.

A. *Vector Space Model:*In this model, document D can be interpreted as a collection of terms "$\{t_1 . . . t_n\}$". Each of these terms can be weighted by some criteria. Given the weighting scheme " $W(D)$", D can be represented by an "n-dimensional vector w". The strong motivation for vector space representation is effortless to design and tell. Documents that discuss a general idea, but use different vocabulary, which is deemed to be different, are really genuine. The "vector public space model" [3],

[5] was introduced in this regard. The idea is to extend the "vector space model" to a time-bound relationship. When calculating the similarities between documents, it introduces the weight of contact on the similarity metric.

B. *Graph Model:*A graph is a collection of vertices (or nodes) and edges, usually denoted "$G = \{V, E\}$",where V is a "vertex" and E is an "edge collection". An edge represents the relationship between vertices. K. M. Hammouda et al. [21] proposes a "document index graph (DIG)". A "DIG" is a "directed graph" where each vertex v_i, characterizedan exclusive word in the corpus. Each edge is between the vertex pairs "(v_i, v_j)" only, if v_j follows the v_i in the corpus. Vertices in the graph track documents that contain words. Sentence path information, *i.e.*,which index is moved to a document and kept in a separate index. In a directed graph, the degree of phase matching between documents is used later to determine the similarity of the document.

Feature Extraction and Selection

Feature extraction starts by means of the parsing of every one document to generate a collection of features and excludes a predefined list of "stop words", that is not related to the semantic aspect. Representative features are then selected from a collection of extracted features [13]. Feature selection is a prerequisite for eliminating noisy features. This declines the high dimensionality of the characteristic space and presents superior data perceptive, which revolve in the move ahead clustering results, effectiveness and achievement. This is widely utilized for "supervised learning" such as "text classification" [14]. Therefore, it is important to increase clustering competence and efficiency. Commonly used feature selection metrics are "term frequency (TF)", "inverse document frequency (IDF)", and the "hybrid".

Feature selection efforts to find the attributes of the dataset largely appropriate to the data mining operation. This is a powerful technique commonly used to reduce the level of problems to an additional convenient level. Feature selection engages searching during different feature small groups and estimatingevery one of these small groups by means of several criteria [10], [13], [19]. The most accepted search policy is"greedy sequential searches" that search forward or backward through the feature space. The "wrapper model technique" uses a data mining algorithm that will be used ultimately to evaluate the data collection. Therefore, it encompasses a selection process centered on data mining algorithms and examines the fundamental properties of the data to evaluate a small group of features prior to data mining. Feature extraction performs thetransformation from M-dimensional space to K-dimensional space.

The "Stop-word filtering" is the most popular and perhaps simplest method used in many document clustering applications. Stop words are terms that occur very often, such as usually, or have modest or no contextual meaning, such as articles, prepositions, and so on. In terms of information words, stop words have very little information about the document context. The idea is to get higher search precision by removing stop words.

The "Feature transformations" are normally utilized in high-dimensional data sets. These methods include techniques such as "principal component analysis" and "singular value decomposition". Variations regularly preserve their unique relative distance between objects. In this way, itwants to create a linear combination of properties to conclude the data collection and reveal the potential structure. Feature conversion is often a pre-processing step, so you can only utilize a few of the newly created features in the clustering algorithm.

In [10], [13], various approaches in related to feature selection similar to "MultipleFeature co-selection for Clustering (MFCC)", "Weighted Semantic Features and Cluster Similarity" with means of "non negative matrix factorization (NMF)", "Local Feature Selection for partitioned hierarchical text clustering approach" supported on "Expectation Maximization (EM)" and "Cluster Validity", supported on "Ant Colony Optimization" are being considered.

Similarity Clustering

Clustering can create separate or overlapping partitions. Overlapping partitions can display documents in multiple clusters, but in non-clustered clustering, every document becomes visible exactly as one cluster. The cluster analysis method is supported on the measurement of similarities among a couple of objects. The purpose of the resemblance among a couple of objects involves three main phases: (i) "the selection of the variables to be used to characterize the objects", (ii) "the selection of a weighting method for these variables", and (iii) "the selection of a similarity coefficient to establish the extent of resemblance between the two attribute vectors" [20].

Precise clustering necessitates specific definition of proximity among the pairs of objects in conditions of pair similarity or distance. Various similarity or distance measurements such as "cosine similarity", "Jaccard correlation coefficient", "Euclidean distance" and "relative entropy" have been suggested and extensively functional. There are many similarity measures are specified in [11].

Applicability of Document Clustering

Clustering is the general appearance of "unsupervised learning"and it is the main instrument in many areas of business and domains. This segment summarizes the essential applicability for clusteringis used.

- *Discovery Similar Documents*: This characteristic is frequently utilized when a user finds a "good" document in search results and needs further. What is fascinating is that clustering is a conceptually similar document that can be found, unlike the search-based approach, where documents can be found to contribute to several of the similar words.
- *Organizing Large Document*: Document retrieval aims at discovery documents related to an exacting query but does not resolve the difficulty of creating a huge quantity of unclassified documents. The confront at this point is to systematize these documents into the same nomenclature that humans have created over the period and utilize them as a "browsing interface" to the novel collection of documents.
- *Duplicate Content Recognition*: In numerous applications, it might require to locate duplicates or few duplicates in a huge numeral of documents. Clustering is also adopted for "plagiarism search", "grouping of related news articles", and "reordering of search results" to ensure a superior assortment along with top-level documents.
- *Recommendation System*: In this kind of system, it suggested article depend on articles you have previously understand. Clustering of articles is probably in instantaneous and great progress the quality.
- *Search Optimization*: Clustering helps you improve search engine quality and efficiency by allowing user queries to be compared to clusters first, instead of directly comparing them to documents, and it is able to simply and sort search results efficiently.

Challenges in Document Clustering

The primary challenge in all clustering approach is to determine what features of a document are considered discrimination, in which case a document model is required. However, most clustering approaches use vector space models to represent each document as a vector. Document clustering has been around for decades, but there are certain concerns that necessitate being deal withed for rapid and efficient clustering.

The previous research on clustering uncertain data is a broad extension of "traditional clustering algorithms" premeditated for specific data. Because objects in a particular data collection are particular points, the allocation of the objects themselves is not measured in the "traditional clustering algorithms". Thus, the study of clustering uncertain data by extending traditional algorithms is inadequate to with "geometric distance-based similarity" measurements and cannot confine differences among indecisive objects with, unlike allocations. In particular, there are three main kinds in the literature: clustering approaches [24], [25], [26], "density-based clustering approaches" [28], [29], and "possible world approaches" [30]. The preliminary and second are in line with the clustering method classification for specific data [23]. Possible world approaches apply only to uncertain data that follows popular global semantics for uncertain data [27].

IV. RELATED WORKS

Clustering is the fundamental of the data mining assignment. Clustering of specific data has been considered for years in "data mining", "machine learning", "pattern recognition", "bioinformatics", and several former fields [1], [2], [4], [10], [11]. However, there is merely a preliminary study of clustering uncertain data. There are several applications that need to cluster a huge gathering of patterns. The explanation of 'large' is ambiguous. In document search, it must cluster millions of instances with 100 or more dimensions to accomplish data generalization. Most of the approach and algorithms proposed in [1 - 15] incapable to handle such huge datasets. Conceptual clustering optimizes some baseline features and is generally expensive to calculate.

J.-P. Mei et al. [1] presented a new effort for "fuzzy clustering" of huge and high-dimensional data that is particularly suitable for document classification. In order to consider the large and high dimensional nature of the formulation of the problem, itsmain idea is to integrate document-tailored fuzzy clustering. As an effective scheme for dealing with large-scale problems, We have identified three representative approaches: "Sampling-Extension", "Single- Pass" and "Divide-Ensemble". Experimental studies of real-world large document datasets have been conducted, and the outcomes illustrate that the suggested approach achieves consistently better than the existing one in document categorization.

J. E. Judith et al. [2] proposed algorithm uses an optimal center for "K-Means clustering" supported on "Particle Swarm Optimization (PSO)". PSO [22] is used to acquire benefit of global search capabilities that provide an optimal focus that helps create more compact clusters with improved search accuracy.H. Jaber et al. [4] proposed clustering algorithms to improve "collaborative decision" making in novel product progressing projects. The goal is to facilitate the collaborative decision-making process by grouping actors into decision-directed relationships. These groups are formed using a unique approach that combines several classic clustering algorithms.

M.Ailem et al. [3] presents a new generation mixed representation for co-clustering these data. This model, "Sparse Poisson Latent Block Model (SPLBM)", is thefoundation on the "Poisson distribution", which occurs obviously for unintended tables such as "document term matrices". There are two improvements of "SPLBM". First, "this is arigorous statistical model which is also very parsimonious", and subsequent, "it was designed from the beginning to address the problem of data sparseness". As a result, in addition to finding "homogeneous blocks", it filters identical but noisy filters because of the scarcity of the data, just like any other algorithm available. The "SPLBM algorithm" offered here accomplish something in predicting the expected cluster construction of a complicated and unstable dataset that previous famous algorithms cannot effectively be handled.

Natthakan et al. [12] present an investigation suggesting that clustering problem can bring to the superiority of the clustering results and suggests a novel "link-based approach" improve the traditional matrix through finding unidentified entries through similarities among clusters in the collection. In particular, the proficient "link-based algorithms" are proposed for fundamental similarity evaluation. Although there have been attempts to explain the difficulty of clustered unqualified data through cluster collections, the outcomes have been aggressive with existing algorithms, but unfortunately, these procedures create the concluding data partitions supported on imperfect information. The default group information template provides merely cluster data point relationships where numerous items are unidentified. To obtain the final clustering results, apply the graph segmentation technique to the graph with the weighted formulas formulated in the refined template.

P. Blomstedt et al. [5] introduces a "model-based approach" for clustering feature vectors of mixed types, allowing each function to take both categorical and real types simultaneously. Such data can be found in, for example, "chemical and biological analysis", "survey data analysis" and "image analysis". The proposed model is formulated within the "Bayesian prediction framework" where the clustering solution corresponds to any partition of data. Using a conjugate analysis, you can use an efficient computer search strategy to find posterior optimal segmentation, since the posterior probability for each possible segmentation can be analytically determined. The derived model is described using multiple synthetic and real datasets.

Banerjee et al. [18] present theoretical analysison "k-means" such as the "iterative relocation clustering algorithm" supported on "Bregman divergences", which is a common situation of "KL divergence", was analyzed. They concludeda "generalized iterative relocation clustering framework" for different resemblance measurements from earlier studies in terms of information theory. They demonstrated that discovering the most favorable clustering is corresponding to diminish the failure function of the "Bregman information" equivalent to the particular "Bregman divergence" utilized as the default similarity computation.

Dhillon et al. [19] utilized "KL divergence" to determine the similarity among words to gather words from the document to diminish the number of features in the document classification. They developed "k-means" such as clustering algorithms and demonstrated that the algorithm minimizes the "Jensen-Shannon divergence" between clusters while monotonically reducing the objective function and minimizing the "Jensen-Shannon divergence" between clusters as in [9]. Since the application data is in the form of text and every word is an isolated variable in the document environment.

W.K. Ngai et al. [24] anticipated a "UK-means method" to expand the "k-means method". The "UK-means method" computes the distance among an uncertain target and the cluster center by an estimated distance. H.-P. Kriegel et al. [28] presented the "FDBSCAN algorithm", which is a probabilistic adding up of the established"DBSCAN algorithm" for clustering specific data. The "DBSCAN" has been extended to a "hierarchical density-based clustering method" called "OPTICS". Kriegel et al. [29] build up a probabilistic edition of "OPTICS" called "FOPTICS" for clustering uncertain data objects. The "FOPTICS" effectiveness of a hierarchical instruct in which data objects that are not clustering members determined for each object are c lustered.

V. CONCLUSION

Clustering of uncertain data is especially important in cluster investigation. Various applications necessitate the investigation of documents including a huge numeral of features or dimensions. Clustering uncertain data is difficult caused with the complexity of the dimensions, and various dimensions could not be appropriate. As the number of dimensions increases, the data becomes gradually more sloppy, making distance measurements between point pairs worthless and lowering the standard density of all points in the data. Therefore, an efficient clustering methodology is required to be developed for uncertain data clustering. In this paper, we present insights into the different clustering methods, document clustering challenges, and their applicability. This includes anextensivereviewof basic data models to various clustering algorithms. Our foremost focus was to consider thedifferent document processing methods to improve clustering results.

REFERENCES

[1] J.-P. Mei, Y. Wang, L. Chen, C. Miao, "Large Scale Document Categorization With Fuzzy Clustering", IEEE Transactions on Fuzzy Systems, Volume-25, Issue-5,Pg.1239 - 1251, 2017.

[2] J. E. Judith, J. Jayakumari, "Distributed document clustering analysis based on a hybrid method", China Communications,Volume-14, Issue-2, Pg.131 - 142, 2017.

[3] M. Ailem, François Role, Mohamed Nadif "Sparse Poisson Latent Block Model for Document Clustering", IEEE Transactions on Knowledge and Data Engineering Volume: 29, Issue: 7 Pages: 1563 - 1576, 2017.

[4] H. Jaber, Franck Marle, and MarijaJankovic, "Improving Collaborative Decision Making in New Product Development Projects Using Clustering Algorithms" IEEE Transactions On Engineering Management, Vol. 62, No. 4, Nov. 2015.

[5] P. Blomstedt, Jing Tang, JieXiong, Christian Granlund, and JukkaCorander, "A Bayesian Predictive Model for Clustering Data of Mixed Discrete and Continuous Type", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 37, No. 3, March 2015.

[6] T.-H. T. Nguyen and V.-N. Huynh, "A k-Means-Like Algorithm for Clustering Categorical Data Using an Information Theoretic-Based Dissimilarity Measure", Springer International Publishing Switzerland, 10.1007/978-3-319-30024-5, pp. 115-130, 2016.

[7] J. Wang, Jingdong Wang, Jingkuan Song, Xin-Shun Xu, Heng Tao Shen, and Shipeng Li, "Optimized Cartesian K-Means", IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 1, Jan. 2015

[8] T. C. Glenn, Alina Zare, and Paul D. Gader, "Bayesian Fuzzy Clustering", IEEE Transactions On Fuzzy Systems, Vol. 23, No. 5, October 2015.

[9] A, M.Almalawi, A. F., ZahirTari, M. A. Cheema, and Ibrahim Khalil "k-NNVWC: An Efficient k-Nearest Neighbors Approach Based on Various-Widths Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 28, No. 1, Jan. 2016.

[10] Z. Li, Jing Liu, Yi Yang, Xiaofang Zhou and Hanqing Lu, "Clustering-Guided Sparse Structural Learning for Unsupervised Feature Selection", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 9, Sept. 2014.

[11] Jiang, Jian Pei, Yufei Tao and Xuemin Lin, "Clustering Uncertain Data Based on Probability Distribution Similarity", IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 4, April 2013.

[12] N. Iam-On, TossaponBoongoen, Simon Garrett, and Chris Price, "A Link-Based Cluster Ensemble Approach for Categorical Data Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 3, March 2012.

[13] Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data", In Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 333-342, 2010.

[14] A. A. Elsayed and E. M. Nour, "A Novel Parallel Algorithms for Clustering Documents Based on the Hierarchical Agglomerative Approach", Int'l Journal of Computer Science & Information Technology, vol.3, issue 2, pp.152, Apr.2011.

[15] J.-P. Mei, Y. Wang, L. Chen, and C. Miao, "Incremental fuzzy clustering for document categorization", in Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2014, pp. 1518–1525, 2014.

[16] Wang, C. Tan, P. Li, and A. C. Knig, "Efficient document clustering via online nonnegative matrix factorizations", in Proceedings of the 2011 SIAM International Conference on Data Mining, pp. 908–919, 2011.

[17] V. D'Orangeville, M. Andre Mayers, M. Ernest Monga, and M. Shengrui Wang, "Efficient Cluster Labeling for Support Vector Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 11, November 2013.

[18] A. Banerjee, S. Merugu, I.S. Dhillon, and J. Ghosh, "Clustering with Bregman Divergences", J. Machine Learning Research, vol. 6, pp. 1705-1749, 2005.

[19] I.S. Dhillon, S. Mallela, and R. Kumar, "A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification", J. Machine Learning Research, vol. 3, pp. 1265-1287, 2003.

[20] L. Wu, Steven C.H. Hoi, Rong Jin, Jianke Zhu, and Nenghai Yu, "Learning Bregman Distance Functions for Semi-Supervised Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 3, March 2012.

[21] K. M. Hammouda and M. S. Kamel, "Efficient phrase-based document indexing for web document clustering", IEEE Transactions on Knowledge and Data Engineering, 16:1279–1296, 2004.

[22] J.J. Li, X. Yang, Y.M. Ju and S.T. Wu, "Survey of PSO clustering algorithms", Application Research of Computers, vol. 26, no.12, Dec.2009.

[23] K. Ericsson and S. Pallickara, "On the performance of high dimensional data clustering and classification algorithms",Future Generation Computer Systems, June 2012

[24] W.K. Ngai, B. Kao, C.K. Chui, R. Cheng, M. Chau, and K.Y. Yip, "Efficient Clustering of Uncertain Data", Proc. Sixth Int'l Conf. Data Mining (ICDM), 2006.

[25] B. Kao, S.D. Lee, D.W. Cheung, W.-S. Ho, and K.F. Chan, "Clustering Uncertain Data Using Voronoi Diagrams", Proc. IEEE Int'l Conf. Data Mining (ICDM), 2008.

[26] S. D. Lee, B. Kao, and R. Cheng, "Reducing Uk-Means to k- Means", Proc. IEEE Int'l Conf. Data Mining Workshops (ICDM), 2007.

[27] N. N. Dalvi and D. Suciu, "Management of Probabilistic Data: Foundations and Challenges", Proc. ACM SIGMOD-SIGACTSIGART Symp. Principles of Database Systems (PODS), 2007.

[28] H.-P. Kriegel and M. Pfeifle, "Density-Based Clustering of Uncertain Data", Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD), 2005.

[29] H.-P. Kriegel and M. Pfeifle, "Hierarchical Density-Based Clustering of Uncertain Data", Proc. IEEE Int'l Conf. Data Mining (ICDM), 2005.

[30] P. B. Volk, F. Rosenthal, M. Hahmann, D. Habich, and W. Lehner, "Clustering Uncertain Data with Possible Worlds", Proc. IEEE Int'l Conf. Data Eng. (ICDE), 2009.