



Current Scenario of Powerful Clustering Technique

Sachin Sharma

M.E. I.T.: University Institute of Engineering and Technology,
Panjab University, Chandigarh, INDIA

Veenu Mangat

Asstt. Prof.: University Institute of Engineering and Technology,
Panjab University, Chandigarh, INDIA

Abstract:- This paper explains clustering, a data mining technique which mainly focuses on the grouping of similar objects, in this technique many researchers have done the work in different areas like web document clustering, text clustering and different models of clustering like connectivity models(hierarchical clustering), centroid model (k-means),distribution models, density models, subspace model, group model, graph based model, this paper provides the view of different techniques and models.

Index terms:-Text clustering, web document clustering, hierarchical clustering, fuzzy clustering, fuzzy c-mean clustering.

I. INTRODUCTION

Clustering is one of the techniques of data mining whose use is versatile, actually in this partitioning of data objects into subsets done, resulting subset known as cluster. Almost all the objects in the cluster are similar and dissimilar objects in the other cluster. Clustering leads to the discovery of previously unknown groups within the data. The application of clustering is in variety of areas such as business intelligence such as stock market surveillance is: prevention and investigation of abusive, manipulative or illegal trading practices in the securities markets. Market surveillance helps to ensure orderly markets, where buyers and sellers are willing to participate because they feel confident in the fairness and accuracy of transactions. Without market surveillance, a market could become disorderly, which would discourage investment and inhibit economic growth. Market surveillance can be provided by the private sector and/or the public sector. and network intrusion detection: is an intrusion detection system that attempts to discover unauthorized access to a computer network by analyzing traffic on the network for signs of malicious activity. Image pattern recognition: pattern recognition is the assignment of a label to a given input value and clustering efficiently compute the efficient class labels, web search: Organizing Web search results into clusters facilitates users' quick browsing through search results, biology: clustering data based on a measure of similarity (or dissimilarity) is a critical step in scientific data analysis and especially in current bioinformatics field, example is, with the advent of DNA Microarrays, clustering analysis becomes a powerful way to explore the expression profiles of all genes in the genome and security: clustering efficiently group the data, also used to find the outliers to detect the security threats. This paper explains the different aspects of clustering. In this paper there are three more sections, section II explains the techniques used in this field and till now what has been done in this area. Section III is the conclusion section, that will provide the summary of the whole work and last section explains the future work.

II. TECHNIQUE USED

Clustering used with the other techniques of data mining, clustering can be applied effectively to group the documents these documents can be web documents and text documents.

A. Document clustering

Regarding the document clustering there are number of techniques proposed, Fatima Ashraf et. al. [1] proposed the technique for automatic information extraction from the HTML pages in this HTML documents containing the semi structured data, a system called the information extraction try to solve the problem which is related to the huge amount of information present on the web and inconvenient human intervention by performing it automatically to extract the information as efficient as possible. In this ClusTex system that uses clustering techniques for information extraction from HTML Web pages is developed. It uses clustering for data extraction, which makes this technique different from others, an unsupervised learning technique and having no requirements of feedback from users after domain features have been accounted for. Odukoya et. al.[8] works on the mining of web documents, algorithm which is proposed, has improved data clustering in terms of preserving their conceptual similarities and high speed, accuracy. Formulation of Algorithm based on K-means

and fuzzy logic technique is used for the simulation in Matlab version 7.0. After simulation results have been compared with the existing algorithms on the behalf of accuracy, response time, adjusted rand index and entropy as performance parameters. Experimental results showed that proposed algorithm is better than K-means. Mei-Ling Shyu *et. al.*[9] gave concept to cluster the web documents by using affinity based similarity measures. This technique consider the user access patterns from web documents to define similarities using a probabilistic model, comparison experiments using a real data set and experimental results conducted, which demonstrate that the proposed similarity measures efficiently perform the cosine coefficient and Euclidean distance under different clustering algorithms. V. Mary Amala Bai *et. al.*[10] presented the results of an experimental study of some common document clustering technique. Comparison of Euclidean K-means (K-Means), Spherical K-means(SK-Means) and unsupervised Principal Direction Divisive Partitioning (PDDP) algorithms is done. Their comparative analysis is performed using Entropy and F-measure. Standard datasets have been taken to conduct the experiments. Result of K-means and SK-means strongly depend on their initialization. PDDP is comparatively difficult to implement since it is a hierarchical algorithm. Important point in it is its performance does not depend on initial clusters. For certain initial clusters the K-means and SK-means performed well than PDDP. Khaled M. Hammouda and Mohamed S. Kamel [11] proposed that document clustering has two key parts. In the first part there is a novel phrase-based document index model, the Document Index Graph, which provides the way to incremental construction of a phrase-based index of the document set concentrates on efficiency. It judges the similarity between documents by efficient phrase matching. The second part is an incremental document clustering algorithm, it maximize the tightness of clusters by monitoring the pair-wise document similarity distribution inside clusters. The combination of these two components creates an underlying model, which is robust and provides accurate document similarity calculation leads to much improved results in Web document clustering over traditional methods. Shen Huang *et. al.*[12] proposed a novel feature co-selection for Web document clustering, which is called Multitype Features Coselection for Clustering (MFCC). MFCC uses intermediate clustering results which are concerned with the one type of feature space and these are used to help the selection in other types of feature spaces. Experiments which are conducted in this context show that MFCC reduce the noise effectively which is introduced by “pseudoclass,” and further improves clustering performance, for most of the selection criteria. Christopher C. Yang and Tobun Dorbin Ng [13] investigated the density based clustering algorithms, a scalable distance-based technique proposed for web opinion clustering. In this experiments were carried out and density based algorithms used for benchmarks, these experiments show that algorithm has achieved greater accuracy. The said approach involves to enable the identification of themes in the development of web social networks and include the interaction of active participants. In this study visualisation tools developed to identify the topic cluster in visualised form. Jean-François Pessiot *et.al.*[14] proposed an unsupervised dimensionality reduction technique for document clustering. Technique is based upon the assumption that terms co-occurring in the same context with the same frequencies are semantically related. By considering this assumption, first of all term clusters using a classification version of the EM algorithm find. Then term clusters are used to represent the documents and document clusters build using multinomial mixture model.

B. Fuzzy Clustering

Clustering technique also used in combination of other techniques like fuzzy clustering, neural network and clustering etc. In this clustering data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters. William-Chandra Tjhi *et. al.* [2] proposed a new co-clustering algorithm called possibilistic fuzzy co-clustering (PFCC) which categorize a large set of data collection automatically. PFCC combines the possibilistic document clustering and fuzzy world ranking, it also include a fast iterative co-clustering procedure. This approach has so many benefits including robustness, rich representations, highly descriptive, performance, and sensitivity in the possibilistic clustering, these all are verified experimentally. Vicenc_Torra *et. al.* [3] extending the Gambal system for clustering and visualization of documents, which is based on the fuzzy clustering techniques. The proposed tool used a fuzzy hierarchical way to structure the set of documents and representing the resulting structure in a three-dimensional sphere graphical interface over which the user can navigate. Ridvan Sarac *et.al.* [5] presented study which includes the documents that belongs to more than one category. Document set which is used for the fuzzy clustering also used the said system. Mainly all the documents which belongs to more than one situations included by this system. The proposed approach consists of two stages to solve multi-categories problem. In the first stage all the documents belonging to more than one category are found. The second stage involves the determination of the categories to which these found documents belong to. Requirements for these stages, to set the value on the behalf of which categories are found, so a-threshold Fuzzy Similarity Classification Method (a-FSCM) and Multiple Categories Vector Method (MCVM) are proposed. With the Experiments it verified that proposed

system can distinguish the documents that belong to more than one category efficiently. Cui-xia Lia and Nan Linb[6] proposed a new clustering algorithm-attribute weighted fuzzy c-means algorithm. This algorithm involves the number of iterations, in the core of this algorithm important attributed have been found. Moreover, this algorithm can also find the hidden cluster structure which is not important. The algorithm simulation on test documents can prove the algorithm can gain a good computation speed, found the latent structure and can categorize the importance of each attribute. Sadaaki Miyamoto [7] proposed an algorithm for clustering in this similarity measures between two terms in a document set described, the core concept used for this is based on the fuzzy neighbourhood and algorithms for term clustering. According to algorithm properties which are related to the neighbourhood and some measures for similarity are used and studied theoretically. Agglomerative hierarchical as well as fuzzy/crisp -means clustering algorithms are proposed. Future scope of this includes numerical experiments of crisp/fuzzy -means on real documents and investigation of theoretical properties of these similarities and algorithms. Debao Chen and Chunxia Zhao[15] A novel data-driven fuzzy clustering method based on maximum entropy principle (MEP) and particle swarm optimization (PSO) is proposed. According to this algorithm, maximum entropy principle is used to predict the memberships of output variables and particle swarm optimization is used to optimize the centre of fuzzy rule based system. The computation cost here is very small if one compares it with training performed, the reason behind it is: number of parameters is less as compared with the deriving fuzzy rules. William-Chandra Tjhi and Lihui Chen [16] a new algorithm fuzzy co-clustering with Ruspini's condition (FCR) is proposed for co-clustering documents and words. Whenever there is a comparison of FCR with existing algorithms for co-clustering, FCR has an edge in terms of ability to generate fuzzy word clusters that capture the natural distribution of words, this concept especially beneficial for the information retrieval. E. Mehdizadeh1 and R. Tavakkoli-Moghaddam [18] presented a fuzzy decision-making approach which mainly deals with a clustering supplier problem in a supply chain system. In this approach there is need to cluster supplier into smaller subsets, subsets should be manageable. A hybrid approach has been developed in order to cluster supplier in fuzzy environment. Algorithm examined on the number of data sets showed the good results.

C. Fuzzy C-means Clustering

The fuzzy c-means (FCM) algorithm partitions a collection $X \{x_1, x_2, \dots, x_n\}$ of n data points specified by m -dimensional vectors $(x_{i1}, x_{i2}, \dots, x_{im})$ into $c (2 \leq c \leq n)$ fuzzy clusters. Here in it cluster centre has been found for each and also minimization of objective function done. The difference between the fuzzy c-means and hard c-mean is, in fuzzy c-mean any point can belong to several cluster with varying degree of membership. Raghu Krishnapuram [17] presented new algorithms—fuzzy-medoids (FCMdd) and robust fuzzy -medoids (RFCMdd)—which apply to relational data for fuzzy clustering. The objective functions involved in this to select c representative objects (medoids) from the data set, selection process should be in such a way that the total fuzzy dissimilarity within each cluster is minimize. When FCMdd compared with the well-known relational fuzzy -means algorithm (RFCM) shows that FCMdd is more efficient. There are several applications of these algorithms to Web mining, including Web document clustering, snippet clustering, and Web access log analysis Xu Yong-Feng et.al.[19] presented a new hybrid image clustering algorithm based on particle swarm optimization and fuzzy C-mean algorithm. Threshold for this are obtained by taking into consideration of FCM's local search ability and PSO's global optimization ability, beside these criterion function of FCM as the object function of PSO considered. By this optimum threshold and maximum entropy obtained, which is beneficial for clustering. Thomas A. Runkler and Christina Katz et. al.[20] proposed a fuzzy clustering by minimizing the fuzzy c-means (FCM) model, in their study they work on the FCM objective function, introduced two new methods for FCM objective function, to minimize the same using particle swarm optimization(PSO). These are PSO-V and PSO-U. In PSO-V approach the concerned particle at the center of the cluster, while it needs representation. In the second approach an unscaled and unnormalized value is used to represent the particle, value in this case is membership value. Wei Wang et. al.[21] presented a SFFCM, a new text clustering algorithm, in this FCM clustering algorithm and supervised feature selection methods are combined. A measure known as supervised feature selection developed, known as CRF-CHI. In this there is χ^2 statistic and category relative frequency involve. Data set for this: OHSUMED (medical abstracts corpus was created to assist information retrieval research), 20-Newsgroups (corpus is a publicly available data set) and Reuters-21578 (Text Categorization Test collection contains documents collected from the Reuters newswire in 1987) used, The experimental results show that the CRF-CHI measure is an effective one, for feature selection in text clustering. Mei Congli et. al.[22] used particle swarm optimization effectively to introduced novel velocity equation, this equation is based on the fuzzy c-means cluster analysis of the current position of particle. Proposed algorithm performed better than other PSO algorithms on the behalf of some bench marks. Chandrani Ray Chowdhury et. al.[23] proposed the system which deals with the information retrieval domain to retrieve the relevant documents. This involves the direct relation in between the number of documents retrieved and performance. More the number of documents retrieved more will be the performance.

D. Hierarchical Clustering

In this technique of clustering hierarchy of documents involve, hierarchical clustering generally involves two types; Agglomerative: this is a bottom up approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive: This is a top down approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. Chun-Ling Chen et. al. [4] presented an effective Fuzzy Frequent Itemset- Based Hierarchical Clustering (F2IHC) approach, in the said approach fuzzy association rule mining and Hierarchical clustering method applied to improve the clustering accuracy of frequent itemset. In this approach, document set responsible for the key terms, after this each document is to be pre-processed so that it can be represented into the form in which mining process can process it. Then, a fuzzy association rule mining algorithm is applied for text to discover a set of highly-related fuzzy frequent itemsets, mainly these highly related itemsets contain key terms, the function of these key terms to be regarded as the labels of the candidate clusters. Finally, all the candidate clusters present used to cluster the documents into a hierarchical cluster tree. Yongxin Liu and Zhijng Liu proposed a simple agglomerative hierarchical K-Means clustering (SAHKC) algorithm, this technique based on the hierarchical k-means algorithm, new approach also introduced for web documents known as multiple feature vector space model. This technique improves the clustering of web documents. Jingxuan Li et. al.[25] studied the technique to cluster the musical data, concerned study involve the application of hierarchical co-clustering (HCC) methods. In this they conducted the experiments on previous hierarchical co-clustering method and proposed a new hierarchical clustering method, which includes the bottom-up approach of clustering, also involve the co-clustering. Then these are used to organize the musical data with effective relationship between artist and – related music information.

III. CONCLUSION

Various clustering techniques have been discussed in this paper. One thing is clear from the above discussion is scope of clustering is versatile than any other field of data-mining. While we mine the knowledge from data by considering clustering, we have the choice to apply any clustering depending upon the data-set and result needed. For document clustering we can apply it on the text documents, web documents, further if demand is there then fuzzy technique, hierarchical clustering can also be applied. So all depends upon the datasets and results needed.

IV. FUTURE WORK

Potential of clustering is huge in almost every field of data mining. As for as personalization of information is concerned, clustering can be applied on it in order to group the users, so that when user appears on the Internet, server may provide the set of documents which are of his/her interest. Other fields like hierarchical co-clustering can be improved in order to cluster the web documents effectively. Clustering for musical, geographical and other complex data may also have the potential in future. As neural network training is time consuming, introducing neural network in clustering in such a way so that effectiveness of clusters over time improves, is also a future direction.

REFERENCES

- [1] Fatima Ashraf, Tansel O'zyer, and Reda Alhajj, "Employing Clustering Techniques for Automatic Information Extraction From HTML Documents," *IEEE Transactions on Systems, Man, and Cybernetics—PART C: Applications and Reviews*, Vol. 38, No. 5, September 2008.
- [2] William-Chandra Tjhi, Lihui Chen, "Possibilistic fuzzy co-clustering of large document collections," *Pattern Recognition*, Vol.40, No.12, p.3452-3466, December, 2007.
- [3] Vicenc Torra a, Sadaaki Miyamoto b, Sergi Lanau a, "Exploration of textual document archives using a fuzzy hierarchical clustering algorithm in the GAMBAL system," *Information Processing and Management*, Vol. 41, No. 3, p. 587-598, 2005.
- [4] Chun-Ling Chen a, Frank S.C. Tseng b, Tyne Liang a, "Mining fuzzy frequent item sets for hierarchical document clustering," *Information Processing and Management*, Vol. 46, No. 2, p. 193-211, January 2010.
- [5] Ridvan Sarac, og'lu Kemal Tu'tu'ncu', Novruz Allahverdi, "A new approach on search for similar documents with multiple categories using fuzzy clustering," *Expert System with Applications*, New York, Vol. 34, No. 4, p. 2545-2554, 2008.
- [6] Cui-xia Lia, Nan Linb, "A Novel Text Clustering Algorithm," *Conference on Energy Systems and Electrical Power*, Singapore, Vol. 13, p. 3583-3588, 2011.
- [7] Sadaaki Miyamoto, Erina Kataoka, "Algorithms for Clustering Terms in Document Set Based on Fuzzy Neighborhoods," *IEEE international conference on Fuzzy System*, p. 979-984, 2005.
- [8] Odukoya, O.H, Aderounmu, G.A. and Adagunodo, E.R., "An Improved Data Clustering Algorithm for Mining Web Documents," *International Conference on Computational Intelligence and Software Engineering (CiSE)*, p. 1-8, 2010.
- [9] Mei-Ling Shyu, Shu-Ching Chen, Min Chen, Stuart H. Rubin, "Affinity-Based Similarity Measure for Web Document Clustering," *IEEE International conference on Information Reuse and integration*, p. 247-252, 2004.

- [10] V. Mary Amala Bai, Dr. D. Manimegalai, "An Analysis of Document Clustering Algorithms," IEEE International Conference on Communication Control and Computing Technologies, p. 402-406, 2010.
- [11] Khaled M. Hammouda, Mohamed S. Kamel, "Efficient Phrase-Based Document Indexing for Web Document Clustering," IEEE Transaction on knowledge and data engineering, Vol. 16, No. 10, p.1279-1296, 2004.
- [12] Shen Huang, Zheng Chen, Yong Yu, and Wei-Ying Ma, "Multitype Features Coselection for Web Document Clustering," IEEE Transactions on knowledge and data engineering, Vol. 18, No. 4, 2006.
- [13] Christopher C. Yang and Tobun Dorbin Ng, "Analyzing and Visualizing Web Opinion Development and Social Interactions With Density-Based Clustering," IEEE Transactions on System, Man, and Cybernetics—Part A: System and Humans, Vol. 41, No. 6, 2011.
- [14] Jean-François Pessiot, Young-Min Kim, Massih R. Amini, Patrick Gallinari, "Improving document clustering in a learned concept space," Information processing and management, Vol. 46, p. 180-192, 2010.
- [15] Debao Chen, Chunxia Zhao, "Data-driven fuzzy clustering based on maximum entropy principle and PSO," Conference on Expert Systems with Applications, Vol. 36, p. 625-633, 2009.
- [16] William-Chandra Tjhi, Lihui Chen, "A partitioning based algorithm to fuzzy co-cluster documents and words," Pattern Recognition Letters, Vol. 27, No. 3, p. 151-159, 2006.
- [17] Raghu Krishnapuram, Anupam Joshi, Olfa Nasraoui, Liyu Yi, "Low-Complexity Fuzzy Relational Clustering Algorithms for Web Mining," IEEE Transactions on Fuzzy Systems, Vol. 9, No. 4, 2001.
- [18] E. Mehdizadeh, R. Tavakkoli-Moghaddam, "A Hybrid Fuzzy Clustering PSO Algorithm for a Clustering Supplier Problem," IEEE conference on Industrial Engineering and Engineering Management, p. 1466-1470, 2007.
- [19] Xu Yong-Feng, Zhang Shu-Ling, "Fuzzy Particle Swarm Clustering of Infrared Images," International Conference on Information and Computing Science, 2009.
- [20] Thomas A. Runkler, Christina Katz, "Fuzzy Clustering by Particle Swarm Optimization," IEEE International Conference on Fuzzy Systems, Vancouver, BC, Canada, 2006.
- [21] Wei Wang, Chunheng Wang, Xia Cui, Ai Wang, "Fuzzy C-Means Text Clustering with Supervised Feature Selection," Fifth International Conference on Fuzzy Systems and Knowledge Discovery, p. 57-61, 2008.
- [22] Mei Congli, Zhou Dawei, "An Improved Particle Swarm Optimization With Fuzzy c-means Clustering Algorithm," International Conference on Intelligent Human-Machine Systems and Cybernetics, 2009.
- [23] Chandrani Ray Chowdhury, Prachet Bhuyan, "Information Retrieval using Fuzzy c-means Clustering and Modified Vector Space Model," Third IEEE International conference on Computer Science and Information Technology, Vol.1, p. 696-700, 2010.
- [24] Yongxin Liu, Zhijng Liu, "An Improved Hierarchical K-Means Algorithm for Web Document Clustering," International Conference on Computer Science and Information Technology, 2008.
- [25] Jingxuan Li, Bo Shao, Tao Li, Mitsunori Ogihara, "Hierarchical Co-Clustering: A New Way to Organize the Music Data," IEEE Transactions on Multimedia, Vol. 14, No. 2, April 2012.