Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

A Survey On Clustering Techniques For Movie Recommendation

B.Rajeswari[1] , Dr. S. Shajun Nisha[2] , Dr. M. Mohamed Sathik[3]

[1]Research Scholar PhD, PG & Research Department of Computer Science, Sadakathullah Appa College, Tirunelveli, India. Affliation of Manonmaniam Sundaranar University 627012 , Reg.No:18211192162020

[2]Assistant Professor & Head , PG & Research Department of Computer Science, Sadakathullah Appa College, Tirunelveli, India.

[3]Principal, PG & Research Department of Computer Science, Sadakathullah Appa College, Tirunelveli, India.

Abstract:

With the big data era, the amount of information accessed in everyday is increasing in most of the industry. All these data do not possess useful information and may be utilized if it has suitable information. It is necessary to investigate, and extract significant learning from such gigantic amount of data. Data mining is an essential method in obtaining important knowledge from huge set of data. Recommendation system helps humans in making decisions. Clustering is one of the important steps in data mining which is used to build efficient recommendation system. Clustering reduces the complication of data mining that has different attributes of many types of large database. Thus unique computation is required for each clustering approaches. In recent years, variety of clustering algorithms have been emerged to satisfy the above requirements and found to be successful when applying to practical recommendation issues. The comprehensive reviews of various clustering techniques that are utilized in different recommendation systems are discussed in this survey.

I. INTRODUCTION

Information sources such as movies, audio, image and text are used as a technological tool to share information in order to satisfy customers requirements. With the development of technology and increased knowledge dissemination, customer's needs become more complex. Service providers have found it difficult in fulfilling the customer's need to offer services and products as there exists an active competition in business. Information filtering systems such as recommendation system

are helpful for users in choosing their preferences according to their consideration and behavior. They started to emerge as a research topic in the late twentieth century. With the ever increasing number of options available online, recommendation system are becoming indispensable. Further the system does not require all the details due to the explosion of overwhelming information and acts as a decision making tool in assisting humans without internet.

Clustering is an unsupervised learning based data mining which is used to group samples of similar and dissimilar data types. Even though it was not broadly studied, it is a research subject in recommendation system area. The drawback of existing recommendation systems such as content based recommendation, collaborative filtering and knowledge based system are overcome by clustering approach. Cluster based and cluster only are the two major approaches of clustering in recommendation system domain. This survey work is emphasized on various clustering techniques involved in various applications of recommendation system.

II. LITERATURE SURVEY

QUANTITATIVE ANALYSIS

Reshma M. Batule et al [2016] presented a hybrid recommendation system based on collaborative filter and clustering. The overview of the combined techniques is also presented to prove the proposed method as efficient recommendation system.

Mingjing Du et al [2016] proposed a novel DPC-KNN algorithm to overcome the loss of some clusters of conventional DPC structure. To preprocess real time high dimensional data, PCA is integrated with the proposed method. This method is compared with spectral clustering and k-means algorithm to show its effectiveness in terms of accuracy.

Bhagyashree Pathak et al [2017] made a survey on five clustering algorithms for information extraction in data mining. The basic study of each algorithm is described and comparative analysis is also done.

George Lekakos et al [2008] proposed a movie recommendation system based on hybrid filtering methods such as collaborative filtering and content based filtering to monitor. The performance of the proposed method is evaluated by measuring prediction accuracy, predictive coverage and prediction run time.

Eugene Seo et al [2010] made a comparison between k-means and SVM classifier based movie recommendation system and the experiment is conducted on Netflix dataset. The dataset consists of 480 thousand training data and 2.8 million of test data. The performance of SVM classifier is found to be superior to k-means.

Sobia Zahra et al [2015] proposed a novel centroid selection method in k-means clustering to improve the accuracy of movie recommendation system using MovieLens, LastFM and FilmTrust

dataset. A comparative analysis is made with FCM and EM to prove the robustness of the proposed approach.

Mohammed Nazimuddin et al [2009] presented a hybrid recommendation system based on diverse module selection method to choose different items between the recommended modules of collaborative filter and fed as input to content based filter.

Pei-pei Wang et al [2019] proposed a key user identification strategy based on density peak cluster method. This recommendation system identifies each significant user as key user. In this way the accuracy and diversity performance of recommendation system is improved.

Shreya Agrawal et al [2017] proposed a hybrid recommendation system based on content based filter, collaborative filter using SVM classifier which is enhanced by GA. Comparatively the proposed method has better accuracy and diversity over individual approach.

Gilda Moradi Dakhel et al [2011] developed a new recommendation algorithm based on k-means clustering and voting algorithm to improve the performance of existing collaborative filtering algorithm. The dataset used is MovieLens.

Chih-Lun Liao et al [2016] developed a dimensionality reduction based novel self constructing method to improve the efficiency of collaborative filtering based recommendation system. Re-transformation approach is used to speed-up the recommendation process.

Jinyin Chen et al [2017] developed an enhanced spectral clustering based recommendation system to overcome the existing sparsity problem. The recommendation over most frequent pair of clusters is generated to reduce the computation time.

Zhou Zhao et al [2017] represented a multi model neural network network for developing heterogeneous social aware recommendation system by using text description, movie image, user ratings and social relationship. A real time social aware large dataset is used to evaluate the performance of the system in terms of sparsity problem.

Siti Rofiqoh Fitriyani et al [2016] proposed a mini batch k-means algorithm for topic recommendation of social media. The accuracy has to be compensated to increase the computation time.

Donghui Yan et al [2018] developed spectral clustering based novel system for distributed data to reduce the communication traffic with less computation time. Experiment was conducted on UV Irvine dataset to show its accurate performance and privacy concern.

Ali Feizollah et al [2014] used MalGenome dataset to measure the performance of k-means and mini batch k-means algorithm. This paper is intended to analyze the network traffic for detecting malwares applications in mobile phones.

Jiang Wang et al [2018] proposed two stage clustering technique (density and distance based k-means clustering) to overcome the challenges faced conventional spatial clustering approach. The effectiveness of the proposed approach is verified with DBSCAN.

Minh D. Nguyen et al [2019] proposed density based clustering approach using spatial textual data of social media to detect geo-tagged record of single data type. The proposed method performs well in terms of f1 score and its derivatives.

Samina Kausaret al [2018] developed a personalized e learning system to obtain information from educational dataset using CFSFDP-HD method. The learning capability can be improved by identifying optimal settings.

Zheng Cao et al [2010] proposed an efficient video similarity search method where the feature extraction is done by image attribute code based on spatio-temporal statistics for large dataset. As the implementation is easy, it can be employed in storage gadgets for video search.

Veena K.M. et al [2017] modified the traditional CA algorithm with density weighted FCM approach to estimate the membership function thereby enhancing the web based recommendation by fetching most frequently used web sites for the users.

S.Ephina Thendral et al [2016] proposed cross domain algorithm that focuses on transferring the user's knowledge from high rated site into sparse domain recommendation system. Thus the proposed method associates the user suggestion on different domains.

Mingjing Du et al [2017] proposed a new DPC-MD clustering algorithm for real time data clustering. This new similarity measurement approach avoids changes in feature and parameter variation and outperforms the existing methods.

Juha Vesanto et al [2000] developed a two stage clustering approach that consists of agglomerative and partition based clustering in a large dataset. This approach performs better than direct data clustering in reducing the computation time.

Sueli A. Mingoti et al [2006] made a comparison between hierarchical (SVM) and non-hierarchical clustering (Fuzzy c-means) and conclude that non- hierarchical approach has good stability and performs well for different dataset.

Kashif Hussain Memon et al [2018] proposed a generalized KWFLICM algorithm for segmenting M-dimensional input images that are highly corrupted by noise. The experimental verification has been done on real time data to show its effectiveness.

S. V. Vimala et al [2019] developed KLD-FCM clustering method to enhance the stability and robustness of existing collaboration filtering based recommendation system. MAE, RMSE, Recall, Standard Deviation and Accuracy are the experimental measures involved in this paper.

Hamidreza Koohi et al [2016] proposed fuzzy c-means clustering for user defined collaborative filter which is a novel approach whose performance is superior to other clustering methods for movieLens dataset. This is further enhanced by the combining CoG defuzzified cluster and Pearson correlation coefficient.

QUALITATIVE ANALYSIS

S. No	Authors & Year	Method used	Advantages	Disadvantages
[1]	Reahma M. Batule, S.A Itkar, 2016	Combination of clustering and collaborative filtering	Relatively good accuracy and low cost.	increased complexity, dependent on human ratings, limited scalability for large datasets
[2]	Mingjing Du, Shifei Ding, Hongjie Jia, 2016	DPC-KNN-PCA	Outperforms well for UCI dataset and low dimensional dataset. Performs well for real time data.	Does not perform well when there is an accumulation of focuses forming vertical streaks in dataset. Need improvement on manifold dataset.
[3]	Bhagyashree Pathak, Niranjan Lal, 2017	Hierarchical methods, Partitioning methods, Density based clustering, Grid based clustering, Model based clustering, Soft computing clustering	Versatility, easy analysis, less processing time, linear complexity	No back tracking capability, data have order dependency; work with mathematical model, high computation cost.
[4]	George Lakekos, Petros Caravelas, 2008	Content based filtering, collaborative filtering	Comparatively the mean absolute error is reduced with 100% accuracy.	Execution time is high, accuracy depends on weight of threshold value
[5]	Eugene Seo and Ho-Jin Choi, 2010	SNM and k-means clustering	Prediction accuracy of SVM is good. RMSE is not affected.	Computation time is long. Some values get missed with large dataset.

[6]	Sobia Zahra, Mustansar Ali Ghazanfar, Asra Khalid, Muhammad Awais Azam, Usman Naeem, Adam Prugel-Bennett, 2015	Novel centroid selection based k-means clustering	Cost effective, accurate prediction, less MAE	The performance depends on number on clusters and decreases with increasing cluster number.
[7]	Mohammed Nazimuddin, Jenu Shrestha, Geun-Sik Jo, 2009	Diverse item selection algorithm	Good prediction accuracy. Diverse item recommendation enhances the content based filter. Cold start problem is eliminated.	When number of active user ratings reduced, the performance will degrade.
[8]	Pei-pei, Pei-yu Liu, Ru Wang, Zhen-fang Zhu, 2017	Density peak clustering based on key user determination method. HHM method	Key users are effectively differentiated from false users. Highly accurate Algorithm complexity reduced.	When number of key users increased, diversity becomes bitter.
[9]	Shreya Agrawal, Pooja Jain, 2017	SVM, Genetic Algorithm	Less computation time. Accuracy, diversity and quality are improved.	Memory requirement is high. Verified for single dataset [movieLens] only.
[10	Gilda Moradi Dakhel, Mehregan Mahdavi, 2011	k-means clustering, Minkowski distance based voting algorithm	More accurate	Consumes more time
[11	Chih-Lun Liao, Shie-Jue Lee, 2016	Self constructing clustering algorithm, re-transformation approach	Recommendation time is reduced. Efficiency is improved. Clusters are formed automatically and a	Same number of cluster is used for grouping. Practical application is restricted to specific dataset.

			pre-determined number of clusters provided by the user is not required.	
[12	Jinyin Chen, Yangyang Wu, Lu Fan, Xiang Lin, Haibin Zheng, Shanqing Yu, and Qi Xuan, 2017	node2vec algorithm based spectral clustering	Sparsity & information loss problems are solved.	The parameter variation has impact on accuracy.
[13	Zhou Zhao, Qifan Yang, Hanqing Lu, Tim Weninger, Deng Cai, Xiaofei He and Yueting Zhuang, 2017	Random walk based learning method	Performs better than existing system by solving sparsity problem.	The proposed method cannot be directly applied for learning the multimodal ranking metric.
[14	Siti Rofiqoh Fitriyani, Hendri Murfi, 2016	Mini batch k-means algorithm	The whole dataset is not required to determine the centroid. Much faster than standard algorithm.	Accuracy is slightly less than k-means method.
[15	Donghui Yan, Yingjie Wang, Jin Wang, Guodong Wu, and Honggang Wang, 2018	Distortion minimizing local transformations	Speed increased by twice. Privacy concern. Feasible.	There was little loss of information.
[16	Ali Feizollah, Nor Badrul Anuar, Rosli Salleh and Fairuz Amalina, 2014	k-means and mini batch k-means algorithm	Mini batch k-means algorithm has good accuracy and less computation time.	Does not suits for large dataset.
[17	Jiang Wang, Cheng Zhu, Yun Zhou , Xianqiang Zhu, Yilin Wang	Density and distance based k-means clustering technique	The clusters with different shapes and densities are effectively identified.	Cores and noise are difficult to determine.

	And Weiming Zhang, 2108		Data size and dimensions are scaled efficiently.	Variation of local point in decision graph. Results are unstable.
[18	Minh D. Nguyen and Won-Yong Shin, 2019	Fuzzy DBSTexC algorithm	Superior performance over existing DBSCAN method.	Computationally complex.
[19	Samina Kausar, Xu Huahu, Iftikhar Hussain, Zhu Wenhao and Misha Zahid, 2018	Clustering by fast search and finding of density peak via heat diffusion method.	Efficiently analyze big data to make robust educational system. Accurate clusters are formed in less time.	The performance of system depends on student's collaboration.
[20	Zheng Cao, Ming Zhu, 2010	Clustering index table search method.	Twenty times faster than other methods. The search efficiency is improved for large database.	The algorithm needs frequent updating whenever new videos are added.
[21	Veena K.M, Radhika M.Pai, 2017	Density weighted FCM	Advantages of both hierarchical and partition clustering are utilized.	Data have order dependency. Computation time is more.
[22	S.Ephina Thendral and C.Valliyammai, 2016	Cross domain collaborative filtering. Hierarchical agglomerative cluster.	95% of accuracy is achieved.	Does not suits for large dataset.
[23	Mingjing Du, Shifei Ding and Yu Xue, 2017	DPC-MD clustering algorithm.	Highly stable. Less sensitive to parameter variation. Efficiently handle mixed data.	There is a need to know parameter values.

[24	Juha Vesanto and Esa Alhoniemi, 2000	Two stage SOM based clustering approach	Less computation cost, low noise, less memory requirement.	Valid only if SOM cluster is same as input data. Requires more time for training.
[25	Sueli A. Mingoti and Joab O. Lima, 2006	SOM, Fuzzy c-means clustering algorithm.	Fuzzy c-means performs well compared to other methods.	SOM method is not evaluated completely. The comparison is restricted with overlapping and outliers parameters.
[26	Kashif Hussain Memon and Dong-HoLee, 2018	Generalised KWFLICM algorithm	Applicable for m-dimensional input data. Suitable for clusters having different size and density. Maximum efficiency of 94.55% is achieved.	Computation time and cost are high.
[27	S. V. Vimala, K. Vivekanandan, 2019	KLD-FCM clustering method	Improved prediction. Low cost.	Time consuming.
[28	Hamidreza Koohi and Kourosh Kiani, 2016	User based clustering based on FCM approach. and Pearson correlation coefficient.	High accuracy. Less computation time	No much improvement in precision and recall for average prediction type.

III. CONCLUSION

In this work a survey is made on numerous clustering techniques involved in recommendation system. The traditional efficient collaboration filtering methods are combined with clustering to enhance their performance. With the aid of technology, CR based recommendation techniques are replaced with machine learning based clustering technique. By analyzing a number of techniques, a comparative study is presented. Evaluation of traditional techniques is also neatly explained to

identify the effective recommender system. By utilizing various modified clustering algorithms, more effective results are provided than the existing methods. The computational time, accuracy, recall, MSE and several other parameters are considered for comparing the proposed techniques. As clustering is the challenging task in movie recommendation system, sparsity and diversity have gained major significant. So an efficient technique which highlights the important criteria for the enhancement of clustering methodology is much needed.

REFERENCES

[1] Reshma M Batule, Prof. Dr. Mrs.S.A.Itkar, "A Survey Paper on different Clustering techniques for Collaborative Filtering for services recommendation", International Journal of Computer Science and Information Technologies, Vol. 7 (3) , 2016, 1410-1413

[2] Mingjing Du , Shifei Ding , Hongjie Jia, "Study on Density Peaks Clustering Based on k-Nearest Neighbors and Principal Component Analysis", knowledge based system, Elsevier ISSN 0950-7051

[3] Bhagyashree Pathak, Niranjan Lal, "A Survey on Clustering Methods in Data Mining", International Journal of Computer Applications (0975 – 8887) Volume 159 No 2, February 2017

[4] George Lakekos, Petros Caravelas, "A hybrid approach for movie recommendation", multimedia tool application, Springer, 2008

[5] Eugene Seo and Ho-Jin Choi, "Movie Recommendation with K-Means Clustering and Self-Organizing Map Methods", International Conference on Agents and Artificial Intelligence, 2010

[6] Sobia Zahra, Mustansar Ali Ghazanfar, Asra Khalid, Muhammad Awais Azam, Usman Naeem, Adam Prugel-Bennett, "Novel Centroid Selection Approaches for KMeans-Clustering Based Recommender Systems", Information Science, Elsevier, 2015

[7] Mohammed Nazim uddin, Jenu Shrestha, Geun-Sik Jo, "Enhanced Content-based Filtering using Diverse Collaborative Prediction for Movie Recommendation" IEEE conference on Intelligent Information and Database Systems, 2009

[8] Pei-pei, Pei-yu Liu, Ru Wang, Zhen-fang Zhu, "A Recommendation Algorithm Based on Density Peak Clustering and Key Users", International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, 2017

[9] Shreya Agrawal, Pooja Jain, "International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)", 2017

[10] Gilda Moradi Dakhel, Mehregan Mahdavi, "A New Collaborative Filtering Algorithm Using K-means Clustering and Neighbors' Voting, International Conference on Hybrid Intelligent Systems (HIS), 2011

[11] Chih-Lun Liao, Shie-Jue Lee, "A Clustering Based Approach to Improving the Efficiency of Collaborative Filtering Recommendation", Electronic Commerce Research and Applications, Elsevier, 2016

[12] Jinyin Chen, Yangyang Wu, Lu Fan, Xiang Lin, Haibin Zheng, Shanqing Yu, and Qi Xuan, "Improved Spectral Clustering Collaborative Filtering with Node2vec Technology", IEEE International Workshop on Complex Systems and Networks, 2017

[13] Zhou Zhao, Qifan Yang, Hanqing Lu, Tim Weninger, Deng Cai, Xiaofei He and Yueting Zhuang, "Social-Aware Movie Recommendation via Multimodal Network Learning", IEEE Transactions on Multimedia, 2017

[14] Siti Rofiqoh Fitriyani, Hendri Murfi, "The K-Means with Mini Batch Algorithm for Topics Detection on Online News", IEEE International Conference on Information and Communication Technologies, 2016

[15] Donghui Yan, Yingjie Wang, Jin Wang, Guodong Wu, and Honggang Wang, "Fast Communication-efficient Spectral Clustering Over Distributed Data", IEEE Transactions on Big Data, 2018

[16] Ali Feizollah, Nor Badrul Anuar, Rosli Salleh and Fairuz Amalina, "Comparative Study of K-means and Mini Batch K-means Clustering Algorithms in Android Malware Detection Using Network Traffic Analysis", IEEE International Symposium on Biometrics and Security Technologies, 2014

[17] Jiang Wang, Cheng Zhu, Yun Zhou , Xianqiang Zhu, Yilin Wang and Weiming Zhang, "From Partition-Based Clustering to Density-Based Clustering: Fast Find Clusters With Diverse Shapes and Densities in Spatial Databases", IEEE access, 2018

[18] Minh D. Nguyen and Won-Yong Shin, "An Improved Density based approach to Spatio-Textual clustering on social media", IEEE access, 2019

[19] Samina Kausar, Xu Huahu, Iftikhar Hussain, Zhu Wenhao and Misha Zahid, "Integration of Data Mining Clustering Approach in the Personalized E-Learning System", IEEE access, 2018

[20] Zheng Cao, Ming Zhu, "An Efficient Video Similarity Search Algorithm", IEEE Conference on Intelligent Computing and Intelligent Systems, 2010

[21] Veena K.M, RadhikaM.Pai, "Clustering of web user's access patterns using a modified competitive agglomerative algorithm, International Conference on Advances in Computing, Communications and Informatics, 2017

[22] S.Ephina Thendral and C.Valliyammai, "Clustering based transfer learning in cross domain recommendation system", IEEE International conference on advance computing, 2016.

[23] Mingjing Du, Shifei Ding and Yu Xue, "A novel density peaks clustering algorithm for mixed data", Elsevier, Pattern Recognition Letters, 2017

[24] Juha Vesanto and Esa Alhoniemi, "Clustering of the self- organizing map", IEEE Transaction on neural networks, Vol. 11, No. 12, 2000

[25] Sueli A. Mingoti and Joab O. Lima, "Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms", Elsevier, Pattern Recognition Letters, 2006

[26] Kashif Hussain Memon and Dong-HoLee, “Generalised kernel weighted fuzzy c-means clustering algorithm with local information”, Elsevier, Fuzzy sets and systems, 2018

[27] S. V. Vimala, K. Vivekanandan, “A Kullback–Leibler divergence-based fuzzy C-means clustering for enhancing the potential of an movie recommendation system”, SN applied science, Springer, 2019

[28] Hamidreza Koohi and Kourosh Kiani, “User Based Collaborative Filtering using Fuzzy C-Means”, Elsevier, International measurement confederation, 2016