Semiconductor Optoelectronics
ISSN: 1001-5868

E-COMMERCE BIG DATA CLASSIFICATION BASED ON ITERATIVE CLUSTERING ALGORITHM

[1]Anima P, [2]Dr. A.S. Aneeshkumar

[1] Research Scholar, Department of Computer Science, AJK college of Arts and Science, Navakarai,

[2] HOD and Assistant Professor, Department of computer Science, AJK college of Arts and Science, Navakarai

ABSTRACT

This study presents an AI-based big data classification method to solve the issues of poor recursion efficiency and excessive redundancy in data classification, while taking into consideration the context of modern e-commerce big data. Adopting the lightning-fast Spark architecture, the method sets a vertical sequence governed by the data jurisdiction dimension, greatly improving data mining efficiency. We use data from e-commerce websites to simulate and validate the suggested method. These results demonstrate that the algorithm is both fast and accurate when it comes to data categorization.

Keywords; big data classification, clustering algorithms, E-commerce, Fuzzy, data mining efficiency

INTRODUCTION

Internet log files are used by web servers to record information about user interactions with the server. There is a wealth of information in online log files that may be used to analyse user needs, provide personalised services, optimise websites, support web professionals, and conceal users' access habits and interests. As a result, researchers and businesses are becoming more interested in online log mining. One common use of clustering algorithms in web log mining is the analysis of user interest on online sites. Users' interests in the web sites are dynamic, meaning they alter over time. Accordingly, both the user clusters and the pages undergo changes throughout time. Many different types of clustering algorithms exist, including partition, hierarchical, density-based, grid-based, model-based, and many more. For the most part, intra-class distance and inter-class distance are the two measurement indices used to evaluate the clustering proces

To have a good clustering effect, an algorithm has to be able to do two things: first, it has to have strong similarity inside classes or objects, and second, it needs weak similarity across classes or objects. As a species, we cluster as a fundamental cognitive process. The only way for humans to learn the underlying rules of objects and conduct easy study on them is via proper clustering. According to [1], the goal of clustering is to group items into classes according to shared characteristics in a way that maximises inter-class similarity while minimising intra-class similarity. To that end, we provide the Fuzzy Clustering Approach (FCA), a new method of clustering, to identify groups of commercial websites' visitors and pages. Users and pages are clustered using this method, and their mobility within the cluster over time is also studied.

Businesses in the e-commerce space want to learn about consumers' online habits in order to tailor their advertising to them, develop a better market strategy, and improve the online shopping experience overall [2]. We can do this with the help of the FCA. Web access records from various websites may provide light on user habits and site architecture. As a result, website structure design is enhanced.

Related Work

It is possible to classify clustering techniques as either "hard partitioning" or "soft partitioning" [3]. A hard partition clustering technique may first be used to segment a picture. The basic idea is to divide the image into smaller parts according to different factors like colour, texture, and greyscale. Then, using techniques like the H-means algorithm, global K-means algorithm, and K-means algorithm, among others, to minimise the objective function, we can find the best solution or partition. One such approach is K-means clustering, which is both quick and easy to use; nevertheless, it has an obvious structure and may easily fall to a local minimum while optimising the segmentation process [4].

According to what Dunn stated in 1947, a soft partitioning clustering technique splits comparable pixels indirectly using pixel characteristics or probability, and then finds the ideal decomposition by minimising the objective function or maximising the likelihood function of the parameters. The FCM fuzzy C-means clustering technique was first suggested in 2001. Then, in 1981 [5], Bezdek compared fuzzy mean clustering to mean clustering in order to prove metric theory. With the validation of its convergence, the establishment of fuzzy clustering theory, and the promotion of its development, the fuzzy mean clustering method became an important branch of fuzzy theory. The clustering algorithm's flexibility was enhanced by incorporating this idea, and this technique is presently frequently utilised [6,39].

In [7], the authors suggested a technique for picture segmentation that makes use of subspace clustering. In order to decrease storage and computation needs, they first created assessment criteria and search algorithms to filter out characteristics that are beneficial for clustering. Then, they clustered the original data set in separate subspaces.

In the 1990s, the idea of a recommendation system that is tailored to each individual user was proposed. Since then, researchers from all around the world have poured a lot of time and energy into this area, and they've accomplished a lot. Since its introduction, the recommendation system has seen extensive application in e-commerce platforms, greatly benefiting their bottom lines. Marketing materials for Amazon state that the company's competitive advantage does not lie in offering customers the lowest prices, but in meeting their most pressing product needs. Nearly a third of Amazon's revenues have come from the implementation of suggested technologies [8,32, 40].

According to the literature [9,33], search technology is the foundation of search engine service providers' technological core, which is available as distinct products. To enhance the user experience and boost the conversion rate of e-commerce websites, recommendation technology is an integral component of these systems. These days, there are businesses that provide recommendation services specifically for online retailers, thanks to the explosion of e-commerce. In order to suggest products to consumers, literature [10,41] makes use of item similarity. Whenever a customer makes a purchase or browses an Amazon product, the company remembers their history, so when they log in again, they can see what other things they have bought.

Based on system analysis, users may get recommendations for both current news messages and information that they are likely to find interesting, as described in literature [11,34]. Google Video likewise uses the same item-based collaborative filtering suggestion for their video recommendations. Each of these algorithms is unique in its design, the data sets it can work with, and the conditions in which it may operate. There are a lot of algorithms that can make recommendations more accurate. People are also interested in variety and originality as markers to measure the efficacy of recommendation systems.

According to the literature [13], the recommendation system's output is associated with the user's degree and preference, and it is based on an energy propagation model that is bipartite graph based. Literature [14,35] suggests integrating the user's commodity purchase connection into the bipartite graph's network structure for energy transmission and heat conduction in order to give consumers diversity suggestions, following attention to diversity in recommendations. Incorporating content data into recommendations is the goal of the user item label, a three-part graph approach based on the paper's described network structure. Migration based on user desire time impacts are considered.

According to literature [15,36], which analyses how user emotion classification impacts e-commerce personalised recommendation results, the goal is to use AI and ML to categorise user emotions, and then use that information to provide users with recommendations tailored to their specific needs. The IMDB and Twitter systems use several algorithms, including Naïve Bayes, SVM, decision trees, and others. We compare the decision tree classification algorithm to personalised recommendation technology and conduct tests on four datasets of hotel and Amazon reviews to confirm that it has a superior emotional classification impact.

The outcomes enhance the impact of tailored suggestion. For the system to be able to fulfil the functional requirements, this paper's design process begins with creating a system framework that takes those needs into account. From there, it analyses and develops the system level in accordance with those needs [16,37]. After that, the system's foreground and backdrop are used to design and define each process analysis in depth. We discuss and explain in detail the key data tables and associated fields of the database using the fuzzy c-means clustering technique, which is one of the essential pieces of the system. From a few of the system's most typical and crucial functions, this article details and evaluates the implementation of the system. By analysing user input and system data, this paper's Personalised Recommendation-based e-commerce recommendation system can accomplish the e-commerce system's primary functions, provide users with personalised product recommendations, and reach the paper's stated goal [17,38].

INTRODUCTION TO FUZZY CLUSTERING

Clustering is a way of data processing. Its purpose is to divide the disordered data into several data groups according to the similarity, so that the data feature similarity in the same group is the highest. Clustering is an unsupervised classification. There is no rule limit before classification, but takes similarity as the only criterion for classification [18]. Clustering algorithm has been proposed for a long time. Clustering technology is generally divided into two categories: hard clustering technology and soft clustering technology. Soft clustering technology is the fuzzy clustering mentioned in this paper. Hard clustering is to completely divide the items into a certain class, which is "either 0 or 1" according to mathematics.

However, for some samples with unclear membership, it may belong to both this class and another class. Using hard clustering to divide such data may cause problems. Fuzzy clustering can solve this problem well. It allocates the membership interval of 0 to 1 for each sample. Instead of completely assigning a sample to a class, it allocates the membership of different classes, which can cluster the sample data between classes more effectively [19]. This paper uses the fuzzy c-means clustering algorithm (FCM), deeply studies and analyzes the FCM clustering principle, introduces it into the recommendation technology, and completes the recommendation function together with the collaborative filtering technology.

Data Preprocessing

Data Cleaning

Data cleansing is necessary since these online shoppers generate over 100,000 of behavioural data per month. The first step is to handle data that has either no value or an unusual one, such as zero-cost data, data with the purchase date as the idle value, or data with plainly incorrect expenditure. The second step is to deal with duplicate data. Right down to the hour, the user's spending habits are spot on. Processing this kind of data is necessary since a small number of customers may make many purchases or add favourites in a single hour. The data consistency is addressed lastly. The indicator R incorporates characteristics of time. Since the time data already contains the date and hour in a single field, it is separated into two columns. The Timestamp field's field type is also changed to year, month, and day to make time calculations easier [20].

Data Sparsity and Cold Start Problem

Due to the data sparsity issue, which significantly affects RS performance in real-world practice, the gathered user ratings for the relevant objects are exceedingly sparse. To that end, a second battery of tests was run to ascertain how well the suggested model reduced the effect of the data sparsity issue.

With the goal of creating a sparse matrix with four distinct densities in mind, we randomly removed certain entries from the training set. To make a sparse user-item matrix with an 80% data density, for instance, we would utilise 80% of the user-item entries as a training set and 20% as a testing set, selected at random. We also established sparse user-item matrices with varying densities in the collection (20%, 40%, 60%) using the same method. Theoretically, a high data density is indicative of a sparse training set, and this holds true in reverse as well. The procedure works iteratively by relocating the K centroids and re-classifying data points in Figure 1 [21].

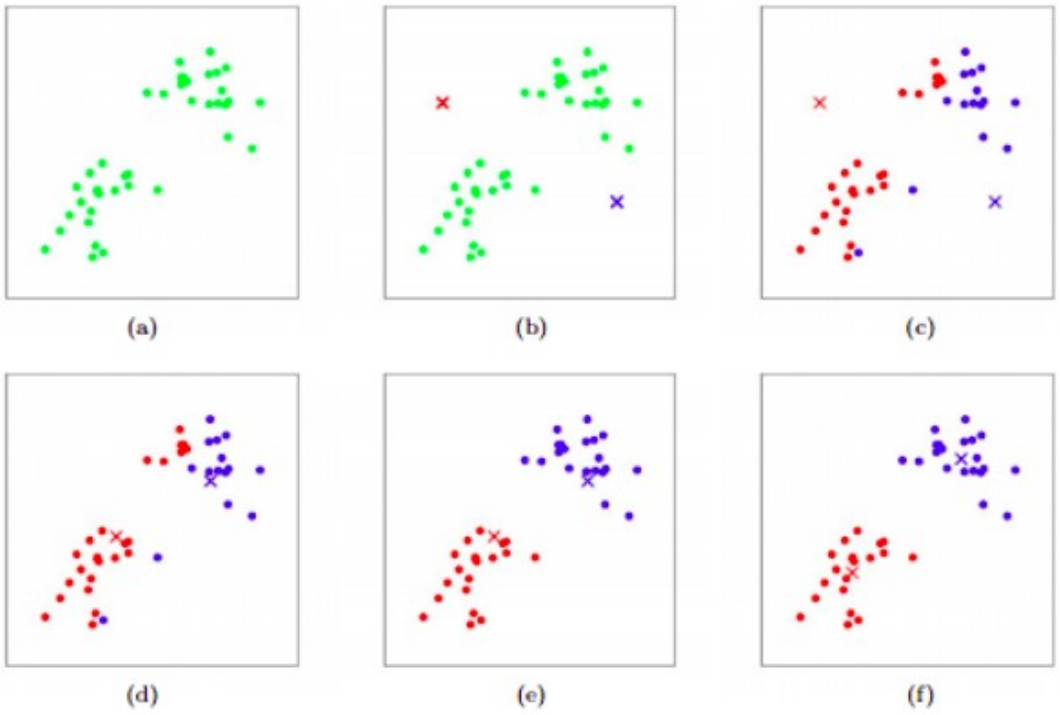

Figure 1 K centroids and re-classifying data points.

Improved Fuzzy Clustering Method FCM

In most fuzzy clustering methods, the choice of the first clustering centre is often determined at random. Nevertheless, the randomly produced clustering centre may not provide stellar clustering results since FCM clustering algorithms are prone to choose the initial clustering centre. To lessen the algorithm's reliance on the first centre point, this work suggests a technique for selecting the initial cluster centre that takes into account the greatest density and grey related degree principle [22,25].

The clustering centre, while doing dataset clustering analysis, has to be located in a region with a dense distribution of sample points. If there are more samples in the area around a certain sequence, then we say that sampling is dense. Consequently, the area with the densest data distribution may serve as the first clustering centre. Not only that, sample points in this region have very tiny distances between them, and the mean distance is also quite small. Most of the time, when a suitable distance mean threshold is provided, the distance and mean distance between the initial cluster centre in the dense area and other sample sites in the region are less than the threshold[23,26].

Following is an approach for selecting cluster centres that adheres to the highest density principle:

The first step is to use a formula to get the average distance value of all sample sequences; this value will serve as the threshold for the analysis.

Second Step: Make a circle with the radius of davg and the sample i($1 \leq i \leq n$) in the middle. Determine the shortest path, in geometric terms, between the first sample and the subsequent samples. If the value of d is less than davg, then the sample is positioned in the circle with its

centre highlighted. Find out how many samples fit within the circle where the sample is located (i) and put the result as nj.

In Step 3, the first initial clustering centre is represented by the sequence with the maximum density, which is produced as max(ni,1≤i=:n).

Enhanced fuzzy c-means clustering is the basis of a personalised recommendation tool. A data collection is first turned into a matrix, and then that matrix is sorted. First, we use an algorithm to fill up the data. Then, we use a fuzzy clustering technique to find the user's closest neighbour. Finally, we construct a collection of neighbours using a collaborative filtering approach [27,28].

On top of clustering, the aforementioned real-time approach additionally searches for the closest neighbour on the matrix. Although this approach may shrink the search area and fix the suggestion issue in real time, the accuracy of recommendations is diminished according to the aforementioned accuracy experiment. Enhancing suggestion efficiency while decreasing accuracy is not a good idea. In addition to the user rating matrix, the item characteristics are also at your disposal. There is a way to make the system more accurate by using this information [24,29].

There are a lot of factors that influence how well the recommendation system works. What follows are some factors that could influence how well fuzzy clustering algorithm recommendations pan out: Firstly, there is a direct correlation between the number of closest neighbours chosen for each suggestion and the recommendation accuracy [25,30,31]. Secondly, the recommendation accuracy varies across various degrees of sparsity. Lastly, there is a relationship between the number of recognised clustering centres and the recommendation accuracy. Accordingly, this work conducts tests from the aforementioned three angles to confirm the enhanced algorithm's efficacy, thereby ensuring the experiment's credibility.

(1)In order to test the effect of choosing a different number of closest neighbours on the recommendation algorithm's accuracy, we compare the recommendation algorithm's performance in various scenarios where we change the number of nearest neighbours before deciding on the number of cluster centres.

(2) To test how varying the number of cluster centres affects recommendation performance, we use a constant number of closest neighbours to choose which cluster centres to use in our trials.

(3) Observing the recommendation accuracy in the data sets with varying sparsities in the two groups of trials above should be enough to prove that data sparsity affects the recommendation outcomes. Based on an average division of the chosen e-commerce data set into five sections denoted as a, W, X, Y, and Z, this paper's simulation experiment is designed to be as accurate as possible. As a means of minimising error, each experiment uses one to serve as a test set and four as a training set for a total of five trials, with an average value taken as the final result. You can see the exact procedure for dividing the dataset in Table 1.

Table 1: Data set division

Number of experiments	Training set	Test set
1	uwxy	z
2	uwxz	y
3	uxyz	w

4	uwyz	x
5	wxyz	u

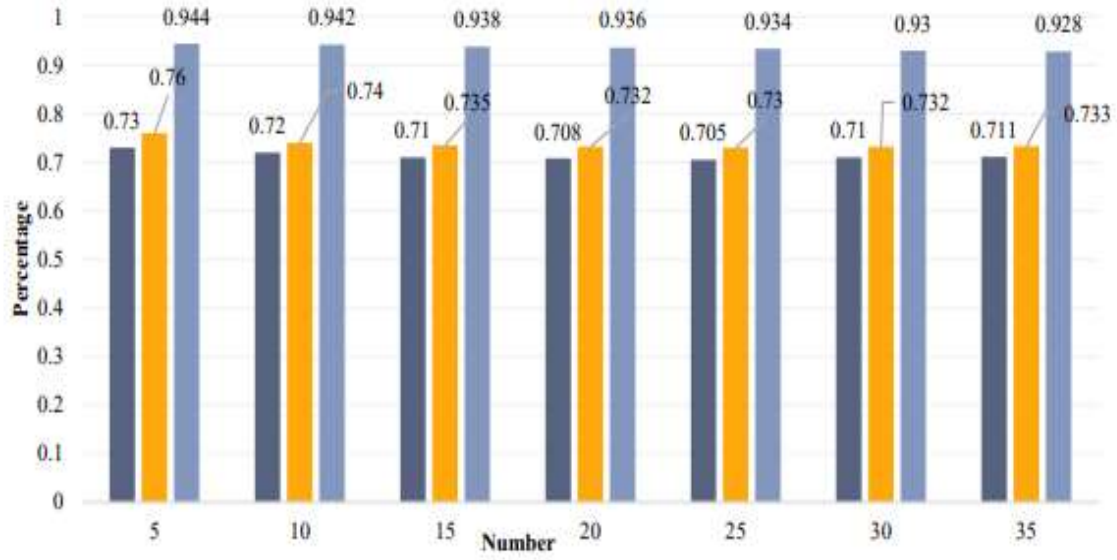

Figure 2 Algorithm test results

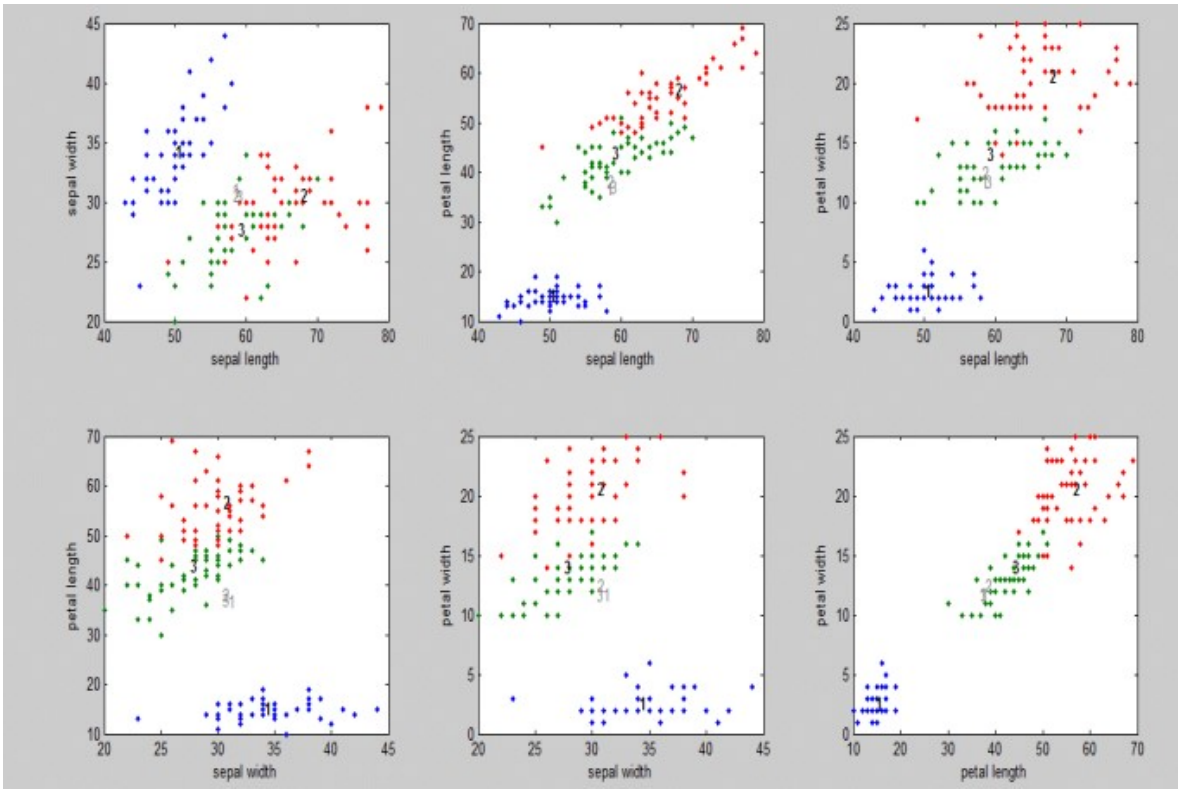

Figure 3 Fuzzy Clustering for Data - MATLAB & Simulink

Observing the comparison figure, it is evident that the quantity of cluster centres remains constant. As the number of closest neighbours increases, the fuzzy clustering (FC) value tends

to drop. However, there is a point at which the FC value tends to grow. The correlation between the recommendation accuracy and the number of closest neighbours may be seen. Hence, the advice you get will depend on how many cluster centres you choose. Recommendation algorithms vary in terms of accuracy, as can be seen in the comparative numbers provided. Therefore, it is clear that the accuracy of the recommendations will be impacted by the data sparsity.

CONCLUSION

Based on this, the planned project is a cross between Fuzzy Clustering and Temporal Clustering. However, the combined information gained from these two approaches surpasses all expectations and is essential for improving e-commerce websites. Since the online is becoming more important as a tool for businesses to gain an advantage and as a platform for sharing information, the significance of web usage mining is undeniable. Search engine optimisation (SEO) relies on analysing online access logs for patterns of user behaviour in order to provide a better experience for web surfers. Its primary goal is to group websites and their visitors into clusters for the purpose of studying user preferences and adapting the site and company to meet the evolving needs of end users. This method has been tested and found to work on web logs. Future research might expand the capabilities of web mining services and find further application for online use mining in the e-commerce sphere.

A major shortcoming of our research is that we were unable to find a satisfactory solution to the multiclass issue. By using our suggested techniques to the numerical study, we find that all of the binary-class datasets get compelling findings. Nevertheless, multiclass datasets show subpar performance. We need to do further research on multiclass datasets in the future. Another difference between our suggested approach and the conventional ones is the computational complexity. A more complicated programming code has to be performed, and the computing time needs to be increased. Our future study is to also provide an algorithm that is more efficient.

REFERENCES

1. Zhao Qing. Research on Personalized Recommendation Algorithm of chemical product e-commerce based on genetic fuzzy clustering [J]. Bonding, 2020, 44 (11): 4
2. Zhu Zhihui, Zhu Meifang. Research on e-commerce personalized recommendation algorithm based on genetic fuzzy clustering [J]. Journal of Jiujiang University: Natural Science Edition, 2019, 034 (001): 61-65 .
3. Li Qingxia, Wei Wenhong, Cai Zhaoquan. E-commerce personalized recommendation algorithm for hybrid user and project collaborative filtering [J]. Journal of Sun Yat sen University: Natural Science Edition, 2016, 55 (5): 6
4. Zhang Kaisheng, song Wenwei, Li Huizhen. Parallel recommendation algorithm based on fuzzy clustering [J]. Journal of Shaanxi University of Technology: Natural Science Edition, 2019 (4): 57-61
5. Zhao Hua, Lin Zheng, Fang AI, et al. A recommendation algorithm based on knowledge tree and its application in mobile e-commerce [J]. 2021 (2011-6): 54-58
6. Ji Xiaoyan, Li Yulong. Personalized recommendation algorithm based on improved clustering in Hadoop environment [J]. Journal of Lanzhou Jiaotong University, 2017, 036 (001): 70-76

7. Du Xi Xi, Liu Huafeng, Jing Liping. A superposition joint clustering recommendation model integrating social networks [J]. Journal of Shandong University: Engineering Edition, 2018, 48 (3): 7
8. Zhang Yinghui, Li Xue. Tourism recommendation algorithm based on fuzzy clustering [J]. Computer technology and development, 2016, 026 (012): 99-102
9. Gao Tianying, Zhang Zhigang, Li Guoyan, et al. A collaborative filtering recommendation algorithm based on user attribute and website type clustering [J]. Journal of Tianjin urban construction university, 2018, 24 (1): 6
10. Wang Xi, Wang Yanming, Wang, et al. An improved collaborative filtering recommendation algorithm [J]. Modern computer (Professional Edition), 2017, 14 (14): 10-15.
11. Dong Hui, Fang Xiao, Ma Jian, et al. Mobile e-commerce user item clustering collaborative filtering recommendation algorithm based on situational awareness [J]. Journal of Guangxi University for Nationalities (NATURAL SCIENCE EDITION), 2018, 24 (02): 67-74.
12. Shi Yingying, Ge Wancheng, Wang Liangyou, et al. Research on improvement of K-means clustering personalized recommendation algorithm [J]. Information and communication, 2016, No. 157 (01): 19-21.
13. Sun Kele, Deng Xianrong. An improved o2o e-commerce recommendation model based on gradient lifting regression algorithm [J]. Journal of Anhui University of architecture and Architecture: Natural Science Edition, 2016 .
14. Zhang Yinghui, Li Xue. Tourism recommendation algorithm based on fuzzy clustering [J]. Computer technology and development, 2016 (12): 99-102.
15. Wang Min, Ji Shaochun. Research on personalized recommendation of Digital Library Based on fuzzy clustering and fuzzy pattern recognition [J]. Modern information, 2016, 36 (04): 52-56.
16. Wang Xiaojun. Distributed hybrid collaborative filtering method in recommendation system [J]. Journal of Beijing University of Posts and telecommunications, 2016, 39 (002): 25-29.
17. Zhao Yan, Wang Yamin, Liu huailiang. Research on personalized Resource Recommendation Model Based on tag network clustering [J]. 2021 (2014-4): 179-183.
18. Wang Zhaokai, Li Yaxing, Feng Xupeng, et al. Personalized information recommendation based on deep belief network [J]. Computer Engineering, 2016, 42 (010): 201-206.
19. Hu Chaoju, sun Keni. Research on Personalized Recommendation Based on user fuzzy clustering [J]. Software guide, 2018, 017 (002): 31-34.
20. Li Haoyang, Fu Yunqing. Collaborative filtering recommendation algorithm based on tag clustering and project topic [J]. Computer science, 2018, v.45 (04): 247-251.
21. Wang Xiaojun, Fu Chao. Using fuzzy blocking to improve the scalability and accuracy of collaborative filtering [J]. Journal of Beijing University of Posts and telecommunications, 2017 (01): 78-82.
22. Zhang Yanju, Lu Chang. IFCM slope one collaborative filtering recommendation algorithm under missing data [J]. 2021 (2020-9): 185-188
23. Lu Q, Guo F. A novel e-commerce customer continuous purchase recommendation model research based on colony clustering. International Journal of Wireless & Mobile Computing, 2016, 11(4):309-317.

24. Hu Q Y, Zhao Z L, Wang C D, et al. An Item Orientated Recommendation Algorithm from the Multi-view Perspective. Neurocomputing, 2017, 269(dec. 20):261-272.
25. Liu X. An improved clustering-based collaborative filtering recommendation algorithm. Cluster Computing, 2017, 20(2):1281-1288.
26. Zheng G, Yu H, Xu W. Collaborative Filtering Recommendation Algorithm with Item Label Features. International Core Journal of Engineering, 2020, 6(1):160-170.
27. Cui L, Huang W, Qiao Y, et al. A novel context-aware recommendation algorithm with two-level SVD in social networks. Future Generation Computer Systems, 2017, 86(SEP.):1459-1470.
28. Feng W, Zhu Q, Zhuang J, et al. An expert recommendation algorithm based on Pearson correlation coefficient and FP-growth. Cluster Computing, 2019, 22(3):1-12.
29. Zhu H, Tian F, Wu K, et al. A multi-constraint learning path recommendation algorithm based on knowledge map. Knowledge-Based Systems, 2018, 143(MAR.1):102-114.
30. Yang F, Wang H, Fu J. Improvement of recommendation algorithm based on Collaborative Deep Learning and its Parallelization on Spark. Journal of Parallel and Distributed Computing, 2021, 148(2):58-68.
31. Zhou X, Su M, Feng G, et al. Intelligent Tourism Recommendation Algorithm based on Text Mining and MP Nerve Cell Model of Multivariate Transportation Modes. IEEE Access, 2020, PP (99):1-1.
32. Fang X, Wang J, Sheng D, et al. Recommendation algorithm combining ratings and comments. AEJ - Alexandria Engineering Journal, 2021, 60(6):5009-5018.
33. Akter S, Wamba S F. Big data analytics in E-commerce: a systematic review and agenda for future research. Electronic Markets, 2016, 26(2):173-194.
34. Fan Y, Ju J, Xiao M. Reputation premium and reputation management: Evidence from the largest e-commerce platform in China. International Journal of Industrial Organization, 2016, 46(May):63-76
35. Tan Libin, Tang dunbing, Chen Weifang, et al. Open design decision-making method with large-scale user participation [J]. Computer integrated manufacturing system, 2020, 26 (4): 9.
36. Mu Jun. community network data crawler algorithm based on association rule mining [J]. Microelectronics and computer, 2018, 035 (008): 105-108.
37. Liu Jingping, Li Ping. A fuzzy cognitive collaborative filtering algorithm [J]. Computer engineering and science, 2018, 040 (005): 898-905.
38. Zhou Chaojin, Wang Yuzhen. Research on personalized recommendation of agricultural products based on improved collaborative filtering algorithm [J]. Journal of Shaoyang University (NATURAL SCIENCE EDITION), 2017, 014 (006): 23-31
39. KK Rakesh, AS Aneeshkumar, Optimization of Fuzzy Logic-Based Genetic Algorithm Techniques in Wireless Sensor Networks Protocols, International Journal of Intelligent Systems and Applications in Engineering, vol-12, issue-14s, 548-556.
40. Sreela Sreedhar, Varghese Paul, AS Aneeshkumar, Solitude Conserve Attribute Cryptographic CP-ABFE Data Protocols in Fuzzy Cloud Service Provider, Indian Journal of Science and Technology, vol-8, issue-25, 1-5.

41. A.S. Aneeshkumar., C. Jothi Venkateswaran, A novel approach for Liver disorder Classification using Data Mining Techniques, Engineering and Scientific International Journal, vol-2, issue-1, 15-18.