INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 05 ISSUE: 08 | AUG - 2021 ISSN: 2582-3930

Collaborative filtering recommendation system based on clustering

Shinde Santosh J.
Dept. of Computer Science and Engg.
Ashokrao Mane Group of Institutions, Kolhapur
Maharashtra, India

Prof. Chougule Pradip.A.
Dept. of Computer Science and Engg.
Ashokrao Mane Group of Institutions, Kolhapur
Maharashtra, India

Abstract— **Providing or recommending appropriate content based on the quality of experience is the most important and challenging problem in the recommendation system. Since collaborative filtering (CF) is one of the most prominent and popular technologies used in recommendation systems, we propose a new cluster-based CF (CBCF) method that uses only user ratings Incentive/Punish User (IPU) model, in which it is therefore easy to implement. Our goal is to design such a simple clustering-based method without further prior information, while improving the accuracy of recommendations. To be precise, the purpose of CBCF with an IPU model is to improve recommendation performance by carefully utilizing different preferences between users, such as accuracy, recall, and F1 score. Specifically, we formulated a constrained optimization problem. Our goal is to maximize the recall rate (or equivalent F1 score) for a given accuracy. For this reason, users are divided into several clusters based on actual scoring data and Pearson correlation coefficient. After that, we reward/punish each project based on the preference trend of users in the same cluster. Our experimental results show that for a given accuracy, without clustering in terms of recall or F1 score, the baseline CF scheme has a significant performance improvement. [1].**

Keywords- Clustering, collaborative filtering, incentivized/penalized user model, Pearson correlation coefficient, recommender system.

I. INTRODUCTION

Due to the large number of videos, audios, essays, art, etc. created online and offline, it may become increasingly difficult for people to find their favorite content effectively. For example, the United States produces and publishes hundreds of feature films and hundreds of thousands of books every year. However, a person will read up to 10,000 books in his life, and then he/she must choose his/her favorite book from them. On the one hand, by helping people choose appropriate content according to their personal preferences, recommendation systems have been developed and used in different fields (for example, the film industry, the music industry, etc.) [1].

In particular, online commerce industries such as Amazon and Netflix have successfully explored how to increase customer loyalty. For example, Amazon.com and Netflix provide personalized products through their own recommendation systems [2], [3], which generate large sales. Although a variety of recommendation systems have been developed, such as content-based personalized recommendation and knowledge-based recommendation, collaborative filtering (CF) is one of the most prominent and popular technologies for recommendation systems [4], [5]. The CF method is usually divided into memory-based CF and model-based CF. In model-based CF, the training data set is used to develop a model for predicting user preferences. Different machine learning techniques such as Bayesian networks, clustering and rule-based methods can also be used to build models. The main content we implemented here is to cluster items based on user ratings. The next step is to develop an incentive/punish user model to provide users with more accurate product recommendations.

II. RELATED WORK

1) Zibin Zheng, Jieming Zhu, and Michael R. Lyu. "Service-generated Big Data and Big Data-as-a-Service: An Overview"

With the prevalence of service computing and cloud computing, more and more services have emerged on the Internet, generating a large amount of data, such as tracking logs, QoS information, service relationships, etc. The massive amount of service generated data has become too large and complex. Effective processing by traditional methods. How to

store, manage and create value from service-oriented big data has become an important research question. On the other hand, as the amount of data continues to increase, there is an urgent need for a single infrastructure that provides common functions for managing and analyzing big data generated by different types of services. To meet this challenge, this article outlines big data generated by services and big data as a service. First, use the big data generated by the three services to improve system performance. Then, adopt big data as a service, including big data infrastructure as a service, big data platform as a service, and big data analysis software as a service to provide general big data-related services (for example, access to the big data and data analysis results generated by the service) To users to improve efficiency and reduce costs.

2) Zibin Zheng, Hao Ma, Michael R. Lyu, and Irwin King "QoS-Aware Web Service Recommendation by Collaborative Filtering"

With the increasing emergence and adoption of Web services on the World Wide Web, Quality of Service (QoS) has become more and more important for describing the non-functional characteristics of Web services. This paper proposes a collaborative filtering method that uses the past experience of service users to predict the QoS value of Web services and make Web service recommendations. The paper proposes a user collaboration mechanism to collect past Web service QoS information from different service users. Then, based on the collected QoS data, a collaborative filtering method is designed to predict the Web service QoS value. Finally, a prototype named WSRec was implemented in Java language and deployed on the Internet for actual experiments. In order to study the QoS value prediction accuracy of our method, we collected 1.5 million Web service invocation results of 100 real-world Web services in 22 countries from 150 service users in 24 countries. Experimental results show that our algorithm has better prediction accuracy than other methods. Our web service QoS data set has been publicly released for future research.

3) T. Niknam, E. Taherian Fard, N. Pourjafarian, and A. Rousta."An efficient hybrid algorithm based on modified imperialist competitive algorithm and K-means for data clustering"

Clustering technology has received attention in many research fields, such as engineering, medicine, biology, and data mining. The purpose of clustering is to collect data points. K-means algorithm is one of the most commonly used clustering techniques. However, the result of K-means depends on the initial state and converges to the local optimum. In order to overcome the local optimal barriers, a lot of research has been done on clustering. This paper proposes an efficient hybrid evolutionary optimization algorithm based on the combination of modified empire competition algorithm (MICA) and K-means (K), called K-MICA, which is used to optimally cluster N objects into K clusters. Then tested the new hybrid KICA algorithm on multiple data sets, and compared its performance with MICA, ACO, PSO, simulated annealing (SA), genetic algorithm (GA), tag search (TS), honeybee mating optimization (HBMO) A comparison and K-means were made. The simulation results show that the proposed evolutionary optimization algorithm is robust and suitable for processing data clustering.

4) X. Li and T. Murata. "Using multidimensional clustering based collaborative filtering approach improving recommendation diversity."

Li et al. It is recommended to incorporate multi-dimensional clustering into the collaborative filtering recommendation model. In the first stage, the proposed algorithm is used to collect and cluster background data in the form of user and project profiles. Then delete the inferior clusters with similar characteristics, and further select suitable clusters based on cluster pruning. In the third stage, project predictions are made by weighted averaging the deviations from the mean of neighbors. This method may increase the diversity of recommendations while maintaining the accuracy of JSPM NTC, Dept. Of Comp. Engg. 2015-16 7

5) Z. Zhou, M. Sellami, W. Gaaloul, M. Barhamgi, and B. Defude. "Data providing services clustering and management for facilitating service discovery and replacement"

Zhou et al. By considering the composite relationship between input, output and semantic relationship, the data provision (DP) service is represented by a vector. Use refined fuzzy C-means algorithm to cluster the vectors. By merging similar services into the same cluster, the capabilities of service search engines have been significantly improved, especially in large Internet-based service repositories. However, in this approach, it is assumed that there is a domain ontology to promote semantic interoperability. In addition, this method is not suitable for some services that lack parameters.

6) M.C. Pham, Y.Cao, R.Klamma, and M.Jarke. "A clustering approach for collaborative filtering recommendation using social network analysis."

Pham et al. It is proposed to use network clustering technology on users' social networks to identify their neighbors, and then use the traditional CF algorithm to generate recommendations. This work depends on the social relationship between users.

7) **R. D. Simon, X. Tengke, and W. Shengrui. "Combining collaborative filtering and clustering for implicit recommendation system."**

Simon et al. A high-dimensional, parameter-free, split hierarchical clustering algorithm is used. The algorithm only needs implicit feedback on past user purchases to discover the relationships among users. According to the clustering results, high-interest products are recommended to users. However, implicit feedback does not always provide certain information about user preferences.

III. PROBLEM STATEMENT

The existing system has the following three important problems:

1. New user/item cold start problem:

The performance of these systems is affected by new user/project cold start issues. When a new user or a new item

enters the system, the score and content information are used to predict and recommend the user or item.

2. Data sparsity:

This problem occurs when the user-item matrix is very sparse, that is, users only rate a small number of items, thus reducing the accuracy of the recommendation. In most of these systems, the percentage of ratings assigned by users is very small compared to the percentage of ratings that the system must predict; therefore, prediction accuracy is praised and memory-based CF methods cannot scale well, and ultimately prediction calculations will be extended. Dimensionality reduction, clustering, and item-based collaborative filtering are more common methods to alleviate this challenge. In this case, the system will be affected.

3. Higher minimum absolute error value:

The existing system also has a higher minimum absolute error value because it cannot provide accurate data.

Existing systems have cold start and data sparse problems. Because the newly added uses and products/items will not be considered in the existing system. To overcome this problem, we combine a variety of collaborative filtering, such as adaptive collaboration (knowledge-based) and item-based collaborative filtering. Therefore, it is recommended that newly added products and users can be considered in the recommendation system, and the minimum absolute error value should be reduced.

IV. PROBLEM SOLUTION

Since collaborative filtering (CF) is one of the most prominent and popular technologies used in recommendation systems, we propose a new cluster-based CF (CBCF) method that uses only user ratings Incentive/Punish User (IPU) model, in which it is therefore easy to implement.

Our goal is to design such a simple clustering-based method without further prior information, while improving the accuracy of recommendations. To be precise, the purpose of CBCF with an IPU model is to improve recommendation performance by carefully utilizing different preferences between users, such as accuracy, recall, and F1 score.

Specifically, we formulated a constrained optimization problem. Our goal is to maximize the recall rate (or equivalent F1 score) for a given accuracy. For this reason, users are divided into several clusters based on actual scoring data and Pearson correlation coefficient. After that, we reward/punish each project based on the preference trend of users in the same cluster.

Our experimental results show that for a given accuracy, without clustering in terms of recall or F1 score, the baseline CF scheme has a significant performance improvement.

V. WORKING MODEL

System Architecture

The development of the Internet has provided people with more ways to interact, and at the same time, it has provided them with a place where they can find information about almost everything. A recommender system can be thought of as a way of combining collection aspects to help people find the information they need or what they are interested in. Recommendation systems are used in various online applications ranging from e-commerce to search engines. There are a variety of techniques that can be used to implement a recommendation system, and each technique has its advantages and disadvantages.

The development of recommendation systems is driven by e-commerce, but other applications are also available for them to use, such as search results and news portal customization. The implementation of CBCF IPUM technology is to overcome some shortcomings of traditional technology. The flaws include performance aspects, but also trust, security, and privacy issues.

The proposed work consists of two phases:

1) Clustering stage:

Clustering is a preprocessing step that divides big data into manageable parts. A cluster contains some similar services, just like a club contains some like-minded users. This is another reason besides the abbreviation. We call this method ClubCF (Collaborative Filtering Based on Clustering). Since the number of services in the cluster is far less than the total number of services, the calculation time of the CF algorithm can be significantly reduced. In addition, since the ratings of similar services in the cluster are more relevant than the ratings of different services, the accuracy of recommendation based on user ratings can be improved.

Some standard partitioning methods (such as K-means) have several limitations:

The result strongly depends on the choice of the number of clusters K, the correct value of K is initially unknown;

Implement incentive/punish user model:

The main contributions of our work are summarized as follows.

• Proposed an easy-to-implement CBCF method using the IPU model to further enhance UX-related performance.

• To design our CBCF method, we first formulated a constrained optimization problem. Our goal is to maximize the recall rate (or equivalent F1 score) for a given accuracy.

• We digitally find the amount of incentives/penalties to be given to each project based on the preference trends of users in the same cluster.

• We have evaluated the performance of the proposed method through a large number of experiments, and proved that the F1 score of the CBCF method using the IPU model is improved compared with the baseline CF method without clustering, and the recall rate of a given (fixed) precision can be significantly improved. This is an increase of about 50%.

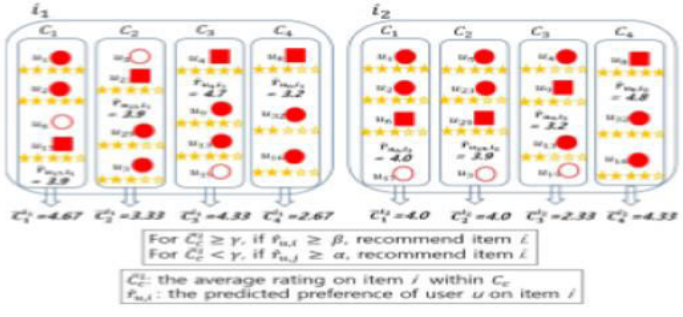

Figure 1. Proposed Clustered Based Collaborative Filtering Method with IPU Model

Figure 1 shows an example of the proposed cluster-based collaborative filtering method with an incentive/penalty user model, which assumes two projects and four clusters. Here, the colored square item and the colored circle item represent test data and training data, respectively. In the figure, i1 and i2 are two items, which can be clustered according to the preferences or ratings given by thousands of users. So each project is divided into n clusters. Then the average score of item i in the cluster will be calculated. Then we are considering some thresholds, rewarding products/items above the threshold, and penalizing products/items below the threshold. Then recommend highly accurate items to users.

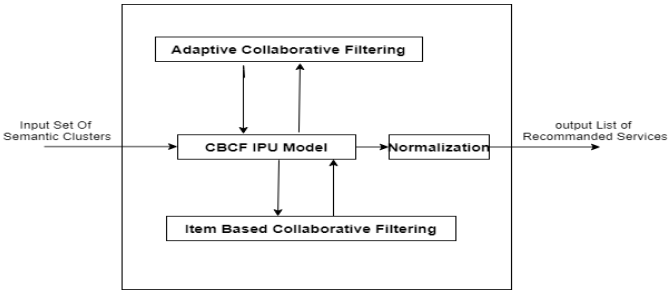

Figure 2 Clustering Based Collaborative Filtering IPU Model

Figure 2 shows the basic architecture of the proposed system. In this case, there are some input cluster sets. The CBCF IPU model will be implemented on this cluster. In the existing system, only project-based collaborative filtering can be realized, but the system has data sparseness and cold start problems. Therefore, in order to avoid the shortcomings in the proposed system model, adaptive collaborative filtering will overcome this shortcoming. The system provides us with an output list of recommended services.

Architecture:

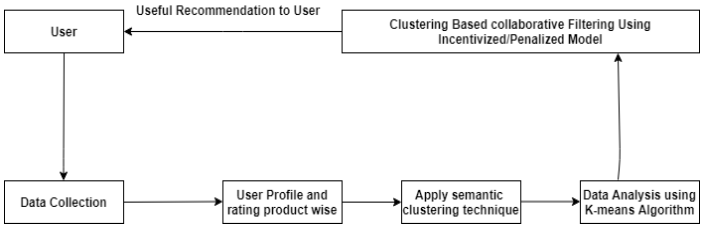

Figure 3. Architecture of CBCF IPU Model

Figure 3 shows the architecture design of the CBCF IPU model. In this architecture, data collection is the first step, collecting data from customers or users based on their interests. It can then be clustered based on the user profile provided by the user and the product wise rating. Then you can use the K-means algorithm for data analysis. It can then be used as input to the suggested system model. The system works on the input and provides users with accurate predictions or recommendations.

VI. RESULT AND DISCUSSION

Table 1 : Error table

Cluster	Size	Actual rating for Item base system	Actual rating with Preferences for Hybrid system	Prediction for Item Based	Prediction for Hybrid Based	MAE for Item system	MAE for Hybrid system
Jewellery & Toy's	155	10	5	21	7	0.070967	0.012903
Perfume & Sun glasses	69	7	8	30	12	0.333333	0.057971
Cloth's	95	5	2	12	3	0.073684	0.010526
Electronics	157	109	11	39	16	0.445859	0.031847
Grouped Cluster	270	25	13	81	19	0.207407	0.022222
All Cluster's	541	51	27	159	40	0.199630	0.024029

Factors that responsible for improving the performance of system are as follows

1. In this system Adaptive Collaborative Filtering is used along with Traditional Item Based System.
2. So MAE value has been improved considerably
3. If we consider execution time for running Adaptive Collaborative Filtering is more as compared to Traditional Item Based System.
4. But Accuracy has improved tremendously

VII. LIMITATION OF SYSTEM

This particular system requires the maximum number of products and users, because if the amount of data is larger, the system will give more accurate recommendations.

VIII. CONCLUSION

A hybrid recommendation system method for big data applications is proposed to generate meaningful recommendations for user sets of items or products that may be of interest. Before applying CF technology, services were merged into some clusters through the AHC algorithm. Then calculate the score similarity between services in the same cluster. Since the number of services in the cluster is far less than the number of services in the entire system, the cost of online computing time is lower.

In addition, since the ratings of services in the same cluster are more correlated with each other than ratings in other clusters, predictions based on the ratings of services in the same cluster will be more accurate than all ratings. Similar or different services in all clusters. The proposed method overcomes the limitations of existing systems, such as data sparsity, scalability, accuracy, and cold start problems. Provide

customers with recommendations by improving the accuracy of recommendations in big data applications.

IX. FUTURE ENHANCEMENT

The proposed system has multiple large-sized clusters. This will affect the performance of the system. For feature research, if we divide this large cluster into a small cluster, the system may give users more accurate recommendations.

X. ACKNOWLEDGMENT

We are thankful to those people who help us a lot in making of this paper. This will help us to grow both academically and professionally.

REFERENCES

[1] Z. Zheng, J. Zhu, and M. R. Lyu,"Service-generated big data and big data as-aservice: An overview,"in Proc. IEEE Int. Congr. Big Data, Oct. 2013,pp. 403-410.

[2] Z. Zheng, H. Ma, M. R. Lyu, and I. King,"QoS-aware web service recommendation by collaborative filtering," IEEE Trans. Services Comput.,vol. 4, no. 2, pp. 140-152, Feb. 2011.

[3] T. Niknam, E. Taherian Fard, N. Pourjafarian, and A. Rousta,"An efficient algorithm based on modified imperialist competitive algorithm and K-means for data clustering," Eng. Appl. Artif. Intell., vol. 24, no. 2, pp. 306-317, Mar. 2011.

[4] X.Li andT. Murata,"Using multidimensional clustering based collaborative filtering approach improving recommendation diversity," in Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol., Dec. 2012, pp. 169-174.

[5] Z. Zhou, M. Sellami, W. Gaaloul, M. Barhamgi, and B. Defude,,"Data providing services clustering and management for facilitating service discovery and replacement,"IEEE Trans. Autom. Sci. Eng., vol. 10, no. 4,pp. 1-16, Oct. 2013.

[6] M. C. Pham, Y. Cao, R. Klamma, and M. Jarke," A clustering approach for collaborative filtering recommendation using social network analysis," J. Univ. Comput. Sci., vol. 17, no. 4, pp. 583-604, Apr. 2011.

[7] CONG TRAN, JANG-YOUNG KIM3, WON-YONG SHIN, AND SANG-WOOK KIM, "Clustering-Based Collaborative Filtering Using an Incentivized/Penalized User Model", IEEE Acces s 2019

[8] B. Yin, Y. Yang, and W. Liu, ''Exploring social activeness and dynamic interest in community-based recommender system,'' in Proc. 23rd Int. Conf. World Wide Web, Seoul, South Korea, 2014, pp. 771–776.

[9] J. Chen, H. Wang, and Z. Yan, ''Evolutionary heterogeneous clustering for rating prediction based on user collaborative filtering,'' Swarm Evol. Comput., vol. 38, pp. 35–41, Feb. 2018.

[10] U. Liji, Y. Chai, and J. Chen, ''Improved personalized recommendation based on user attributes clustering and score matrix filling,'' Comput. Standards Interfaces, vol. 57, pp. 59–67, Mar. 2018.

[11] J. Chen, L. Wei, Uliji, and L. Zhang, ''Dynamic evolutionary clustering approach based on time weight and latent attributes for collaborative filtering recommendation,'' Chaos Solitons Fractals, vol. 114, pp. 8–18, Sep. 2018.

[12] H. Jazayeriy, S. Mohammadi, and S. Shamshirband, ''A fast recommender system for cold user using categorized items,'' Math. Comput.Appl., vol. 23, no. 1, p. 1, Jan. 2018.

[13] Sarita Dhankhar, and Neha, " Hybrid Recommender System under Temporal Vector." International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 7, July 2014.