Item-Based Energy Clustering Recommendation

Tu Cam Thi Tran[1], Lan Phuong Phan[2], and Hiep Xuan Huynh[2](✉)

[1] Vinh Long University of Technology Education, Vinh Long province, Vietnam
tuttc@vlute.edu.vn
[2] Can Tho University, Can Tho City, Vietnam
{pplan,hxhiep}@ctu.edu.vn

Abstract. Previous recommendation systems have focused on algorithms to make the recommendations based on the individual items. However, in many areas, the introduction about a cluster of the items based on the general characteristics of the item is more important than just focusing on the individual items. In this paper, we have proposed a new approach for the recommendation system, the proposed method uses the energy distance to group the items with similar properties or characteristics into a cluster, then based on the item clusters to give the most suitable recommendations for the users. In addition, the methods based on error (MAE_(c)) and accuracy (Precison_(c)-Recall_(c)) are also selected to evaluate the reliability of the new proposed model on two popular datasets Jester5k and MovieLens100k. Besides, the proposed model is also compared with two item-based collaborative filtering models using the Cosine and Pearson measures in "rrecsys" package and three item-based collaborative filtering models using the Matching, Euclidean and Karypis measures in "recommenderlab" package. The experimental results have shown that the proposed model is better than the compared models.

Keywords: Item-based · Energy distance · Clustering recommendation · Recommendation system · Item clusters

1 Introduction

A recommendation system [1] is a decision support system that provides recommendations by predicting user preferences. In addition, a group recommendation system [2, 3, 5] analyzes the interests of the group members and makes a final decision, which will be accepted by all members. The Energy [4, 16] measures the distance between the distributions of random vectors. The dimensions of those vectors are not certainly equal. Energy distance is also widely applied in research [16] such as: testing independence by distance covariance, goodness-of-fit, generalizations of clustering algorithms, change point analysis, feature selection, etc.

Previously studied energy-based recommendation systems [3, 4] were mainly focused on giving the recommendations based on individual items or the user group-based recommendation system [3]. In many areas, introduction about a cluster of the items [12] (e.g. movie cluster, home appliances cluster, etc. These clusters are based on

P. Cong Vinh and N. Thanh Tung (Eds.): ICCASA 2023, LNICST 579, pp. 115–125, 2024.
https://doi.org/10.1007/978-3-031-58878-5_8

general characteristics of the item to group), this is more important than just focusing on the individual items [11]. However, the relationship between the item clusters using energy distance in the recommendation system has not yet been considered in all these studies.

In this article, we propose a new recommendation model that considers relationships among the item clusters using energy distance. This approach is made on the basis of determining the energy relationship between the item cluster in pairs. In addition, we used an accuracy-based evaluation method (Precision_(c), Recall_(c)) and using error-based evaluation methods (MAE_(c) to evaluate the proposed model, and compared it with two item-based collaborative filtering models using the Cosine and Pearson measures in "rrecsys" package and three item-based collaborative filtering models using the Matching, Euclidean and Karypis measures in "recommenderlab" package.

The structure of the article is organized as follows. Section 2 presents related work, presenting: the clustering recommendation based on the items, the energy approach. Sect 3 shows the methods to be used for evaluating the clustering recommendation models (the accuracy-based evaluation method, the error-based evaluation method). Sect 4 depicts the proposed model that uses the relationship among the item clusters and the energy distance. Sect 5 shows the experiment results on the "recommenderlab" and "rrecsys" package for both the Jester5k datasets and MovieLens100k datasets. Section 6 is the conclusion.

2 Related Work

2.1 The Clustering Recommendation Based on the Items

The clustering recommendation systems [3, 8, 9] are clustered based on the types of cluster to which the system recommends. Clusters can be mainly based on the interactions among the members of the cluster.

Table 1. Ratings matrix for the item cluster.

Cluster	Cluster_members	Rating $r(i_j, u_k)$			
		u_1	u_2	…	u_k
$c1$	i_1	?	$r(ij, u_k)$	…	?
	i_2	$r(i_j, u_k)$	?	…	$r(i_j, u_k)$
	i_3	$r(i_j, u_k)$	?	…	?
$c2$	i_4	$r(i_j, u_k)$	?	…	?
	i_5	$r(i_j, u_k)$	$r(ij, u_k)$	…	$r(ij, u_k)$
	i_6	$r(i_j, u_k)$	$r(i_j, u_k)$	…	$r(i_j, u_k)$
$c3$	i_7	?	?	…	?
	i_8	$r(i_j, u_k)$	?	…	?
	i_9	$r(i_j, u_k)$	?	…	?
cn	…	…	…	…	…
	i_j	$r(i_j, u_k)$	?	…	$r(ij, u_k)$

The clustering is a process used to partition the data into a cluster. The data with the same characteristics, or the similar properties, they will be grouped into a cluster. The number of the clusters will be less than the number of individual items of the original data. The characteristic of the clustering method is to reduce the amount of data to be compared, to save time for the recommendation. For example, **Table 1** presented the ratings matrix for the item cluster.

2.2 The Energy Approach

Energy distance.
Energy distance [7, 16] is a statistical distance between the observed variables. The concept is based on the notion of Newton's gravitational potential energy, which is a function of the distance between two bodies in a gravitational space. Energy distance is applied to random vectors, where these random vectors have an unlimited size. Let $I_1 = I_{11}, I_{12}, \ldots, I_{1n}$ and $I_2 = I_{21}, I_{22}, \ldots, I_{2m}$ be independent random vectors in Euclidean space. The energy distance between I_1 and I_2 is define as:

$$\varepsilon_{n,m}(I_1, I_2) = 2\mathrm{E}|I_1 - I_2|_d - \mathrm{E}\left|I_1 - I_1{}'\right|_d - \mathrm{E}\left|I_2 - I_2{}'\right|_d \tag{1}$$

In (1), a random variable $I_1\prime$ (or $I_2\prime$) represents a copy, which is independent and distributed like I_1 (or I_2).

The potential energy (shortly, energy) of the independent random variables I_1 and I_2 is defined by distance function ε as the follow:

Where:

$$\mathrm{E}|I_1 - I_2|_d = \frac{1}{\mathrm{nm}} \sum\nolimits_{\mathrm{i}=1}^{\mathrm{n}} \sum\nolimits_{\mathrm{j}=1}^{\mathrm{m}} \left|\mathrm{I}_{1\mathrm{i}} - \mathrm{I}_{2\mathrm{j}}\right| \tag{2}$$

$$\mathrm{E}\left|I_1 - I_1{}'\right|_d = \frac{1}{n^2} \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{n} \left|I_{1\mathrm{i}} - I_{1\mathrm{j}}\right| \tag{3}$$

$$\mathrm{E}\left|I_2 - I_2{}'\right|_d = \frac{1}{m^2} \sum\nolimits_{i=1}^{m} \sum\nolimits_{j=1}^{m} \left|I_{2\mathrm{i}} - I_{2\mathrm{j}}\right| \tag{4}$$

Advantages of energy distance include: Energy distance is very easy to compute, it is consistent and require no distributional assumptions other than finite first moments.

The energy between the clusters.
The energy [6, 16] between the clusters was calculated from the data in the rating matrix (U x I x R), where each row is a multivariate observation.

Energy distance is used to measure the statistical distance between two clusters, and to search for the best partition between clusters.

The energy distance between two clusters C_i, C_j of size n_i, m_j is the energy distance $e(C_i, C_j)$, defined by:

$$e\left(C_i, C_j\right) = \frac{n_i m_j}{n_i + m_j}[2G_{ij} - G_{ii} - G_{jj}] \tag{5}$$

where

$$G_{ij} = \frac{1}{n_i m_j} \sum_{\mathrm{p}=1}^{n_i} \sum_{\mathrm{q}=1}^{m_j} \|I_{\mathrm{ip}} - I_{\mathrm{jq}}\|^{\alpha} \tag{6}$$

In (5), C_i is the cluster i, C_j is the cluster j; I_{ip} denotes the p-th observation in the i-th cluster, The exponent alpha should be in the interval (0,2], $\|.\|$ denotes Euclidean norm.

2.3 Evaluation Approach for the Clustering Recommendation

To evaluate the clustering recommendation system [13, 14], the accuracy-based evaluation method (the precision_(c), recall_(c)), and the error-based evaluation method (MAE_(c)), both are used.

The accuracy-based evaluation method.
Precision [13] is the fraction of the number of relevant recommended items (true positives) in relation to the total number of recommended items.

$$precision\_(c) = \frac{\left|predicted_k(c) \cap relevant(c)\right|}{k} \tag{7}$$

Recall [13] is the fraction of the number of relevant recommended items in relation to the number of all relevant items.

$$recall\_(c) = \frac{\left|predicted_k(c) \cap relevant(c)\right|}{relevant(c)} \tag{8}$$

where, k is the length of the list of recommended items and c is cluster.
$predicted_k(c)$ shows a list of k items recommended to cluster c.
$relevant(c)$ denotes all items relevant for c.

The error-based evaluation method.
The formula of Mean Absolute Error cluster (MAE_(c)) [13] between the clusters is shown in (5)

$$MAE\_(c) = \frac{\sum_{r(c,j)\in R_c} \left|r(c,j) - \hat{r}(c,j)\right|}{|R_c|} \tag{9}$$

The rating of cluster c, r (c, j) is calculated in (12).

$$r(c,j) = \frac{\sum_{i\in c} r(i,j)}{|c|} \tag{10}$$

The prediction of cluster g, $\hat{r}(g, j)$ is calculated in (13).

$$\hat{r}(c,j) = \frac{\sum_{i\in c} \hat{r}(i,j)}{|c|} \tag{11}$$

With c is the number of clusters; R_c shows the set of ratings of cluster c collected the ratings of item; $r(c, j)$ is the real rating of cluster c at the item j; $\hat{r}(c, j)$ is the predicted rating of cluster c at the item j.

3 The Model of the Item-Based Clustering Recommendation

Figure 1 shows a overview of the item-based clustering recommendation model using the energy approach. This recommendation model is described as follows:

– Input (U x I x R)

+ $U = \{u_1, u_2, \ldots, u_n\}, u_k \in U, k = 1..n$, (including n objects).
+ $I = \{i_1, i_2, \ldots, i_m\}, i_j \in I, j = 1..m$, (including m attributes).
+ The rating matrix R, with each R_{ui} is a value of R:

$$R_{ui} = \begin{cases} r_{ui} \ \ if \ \ the \ user \ u \ rates \ the \ item \ i \\ \emptyset \ if \ \ the \ user \ u \ dose \ not \ rate \ the \ item \ i \end{cases}$$

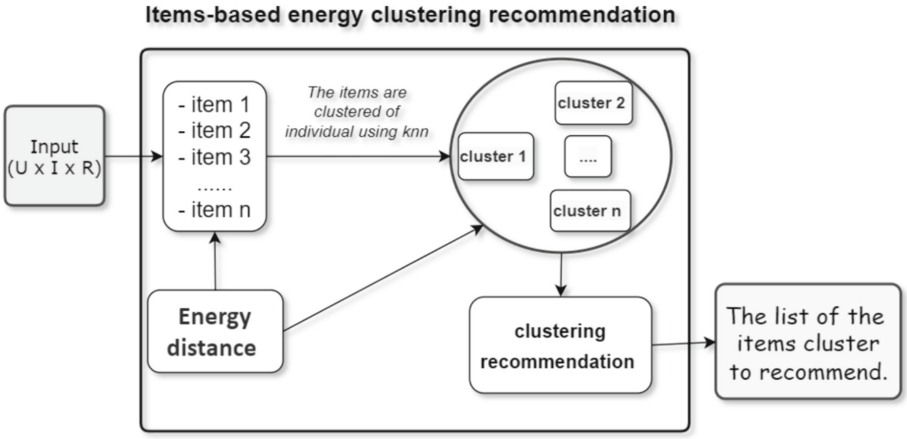

Fig. 1 Item-based clustering recommendation model with energy.

- The energy distance is used to calculate energy between the items and between the clusters. The formula of energy distance are presented in (1) and (5).

- The clustering recommendation is used to predict the missing ratings of the cluster.

- Output is the list of the item clusters with the highest ratings used to recommend Top-N items.

4 Item-Based Clustering Recommendation Algorithm

The energy-based clustering recommendation algorithm is presented in follow:

Algorithm. Energy-based clustering model

Input: The Data Matrix (U x I x R);

Output: Recommending Top-N items;

Begin

[1]: The energy between an item and an item is calculated in I.

<Matrix[i][j] = Energy [i_i x i_j]>;

[2]: The missing rating values of the R matrix is predicted by the knn method using the energy calculated in step 1.

[3]: The items are divided into clusters by using the energy distance.

<Clustering_List_C[i][j] = Energy [Clustering _C[i]],[Clustering _C[j]]>;

[4]: The ratings of the item cluster are predicted based on the average method of cluster

[**5**]: The predicted ratings of the item cluster are sorted with DESC.

< Sort (Clustering_List_C[i])>;

[6]: The n items are recommended with the highest predicted ratings.

< Print (Top-C[i]>);

End.

5 Experiment

5.1 Dataset

The data is selected for the experimental results includes two popular sets, that are: Jester5k dataset [17] and MovieLens Dataset [18]. These two data sets are presented in **Table 2**.

Table 2. Ratings matrix for the item cluster.

Name	*Rating*	*Rating value*	*Date*	*Note*
Jester5k dataset	5000 x 100 (5000 users and 100 jokes)	From -10 to + 10	April 1999 and May 2003	All chosen users have rated 36 or more jokes
MovieLens Dataset	100,000 (943 users with 1682 movies)	From 1 to 5	Released 4/1998	Each user constraint rated at least 20 movies

5.2 Tool

In the article, the proposed model is built by R language (name EIB - Energy Items_cluster Based is items based clustering collaborative filtering recommendation system using energy distance). This model is compared with two models including: PIB (Pearson Items Based using the Pearson measure) and CIB (Cosine Items Based using the Cosine measure), the PIB and CIB models are available in the "rrecsys" package [14].

Besides, the proposed model is also built in "recommenderlab" package [15] with "energy" package [16] (name IBCF_energy_cluster - Items Based Collaborative Filtering using energy distance to cluster). Compared three models including: IBCF_matching (Item-based collaborative filtering model using matching measure), IBCF_euclidean (Item-based collaborative filtering model using euclidean measure), IBCF_karypis (Item-based collaborative filtering model using the karypis measure), these three models are available in the "recommenderlab" package.

5.3 Scenario 1: Item-Based Clustering by "rrecsys" Package

The experiment result with Jester5k dataset.

This scenario evaluates the error with the MAE_c value of the proposed models EIB (Energy_cluster Items Based using the energy distance) and PIB (Pearson Items Based using the Pearson measure) available in the "rrecsys" package on Jester5k dataset with k nearest items (knn) is 5, 10, 20, 30, 40, 50.

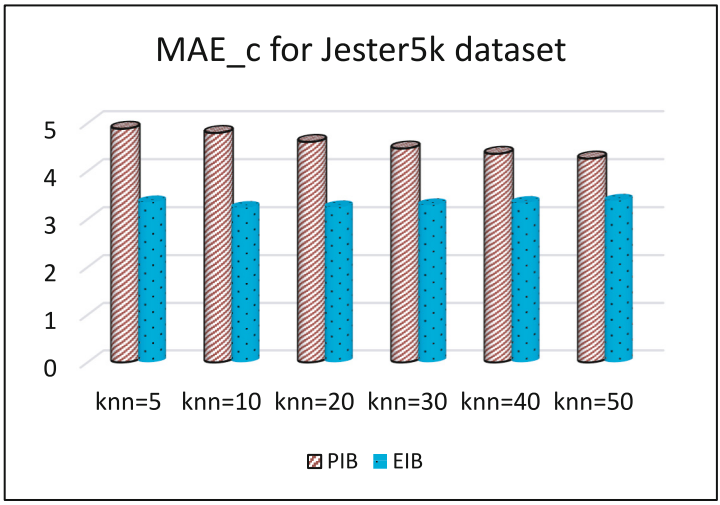

Fig. 2 The MAE_c error value for two models PIB and EIB with Jester5k.

The results in Fig. 2 present: when k nearest items (knn) is 5, 10, 20, 30, 40, 50, the MAE_c error values of EIB are always lower than the MAE_c error values of PIB.

The experiment result with Movielense dataset.

This experiment result showed the MAE_c error value of the proposed models EIB (Energy Items_cluster Based), PIB (Pearson Items Based) and CIB (Cosine Items Based) (the PIB and CIB models are available in the "rrecsys" package) on Movielense dataset with k nearest items (knn) is 30, 40, 50, 60.

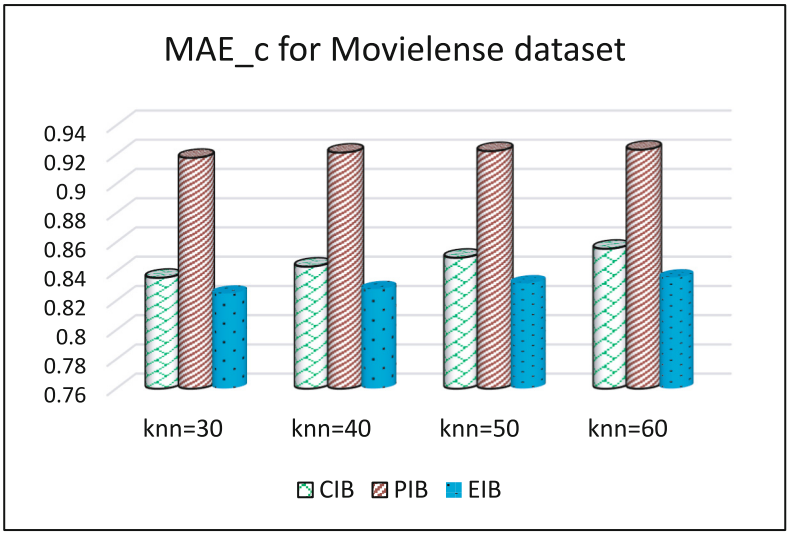

Fig. 3 The MAE_c error for three models EIB, PIB and CIB with Movielense.

Figure 3 Presents that the MAE_c error values of EIB are always lower than the MAE_c error values of PIB and CIB, when k nearest items (knn) is 30, 40, 50, 60.

5.4 Scenario 2: Energy-Based Clustering by "Recommenderlab" Package

This scenario presents item_based clustering recommendation system with energy distance on the Jester5k dataset using two evaluation approach that are the Precision_(c) and Recall_(c). The proposed model (IBCF_energy_cluster) is built with the item-based clustering collaborative filtering model using the energy approach. This model is compared with other models such as: IBCF_matching, IBCF_euclidean, IBCF_karypis, with k nearest items (knn) is 25, 30, 35, 40.

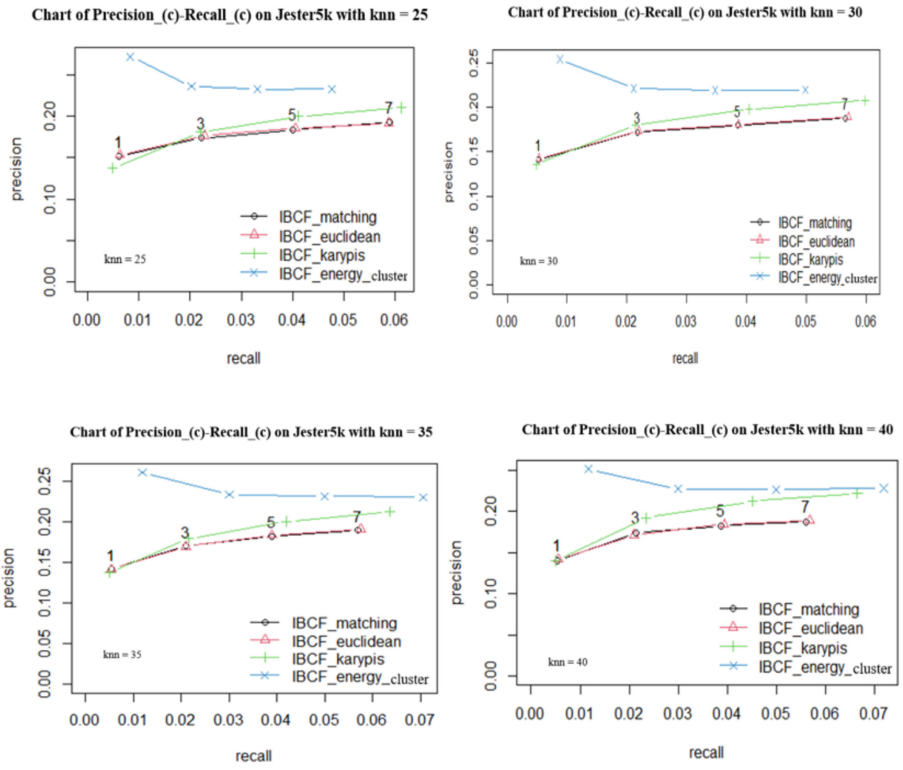

Fig. 4 Figures for chart of Precision_(c) - Recall_(c) with knn = 25, 30, 35, 40 on Jester5k.

Figure 4 presents the Precision_(c) - Recall_(c) values of four models. In which the Precision_(c) - Recall_(c) value of the IBCF_energy_cluster model are always higher than the Precision_(c) - Recall_(c) value of the IBCF_matching, IBCF_euclidean and IBCF_karypis, when k nearest items (knn) is 25, 30, 35, 40.

The experiment result of Fig. 5 presents the ROC curve of four models. In which the ROC curve of the IBCF_energy_cluster model are always higher than the ROC curve of the IBCF_matching, IBCF_euclidean and IBCF_karypis, when k nearest items (knn) is 25, 30, 35, 40.

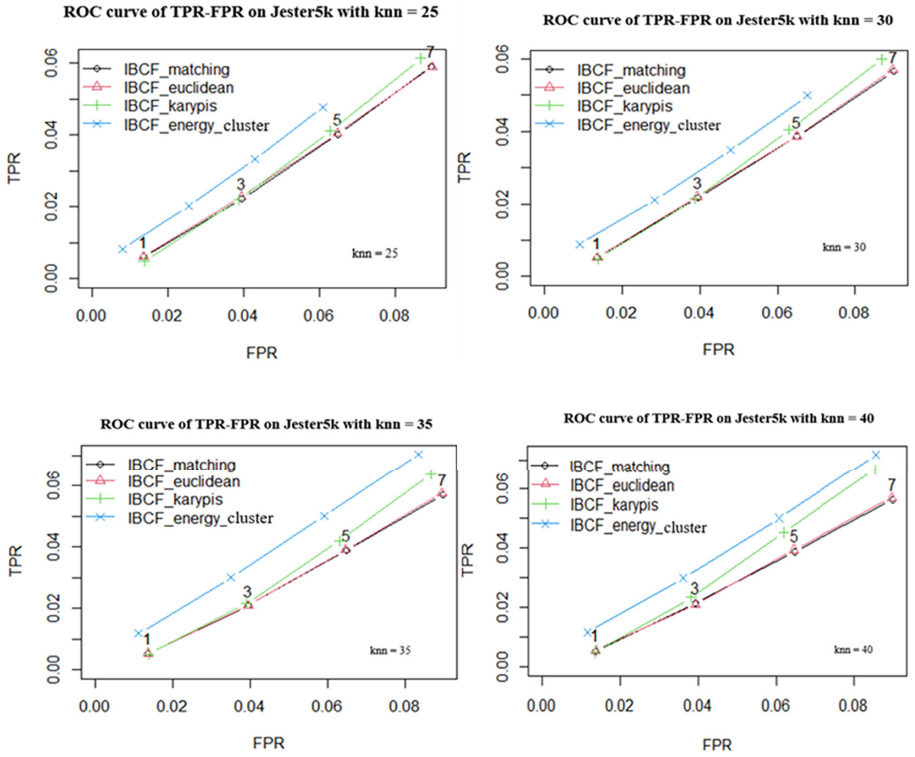

Fig. 5 Figures for ROC curve of TPR-FPR with knn = 25, 30, 35, 40 on Jester5k.

6 Conclusion

In this paper, a clustering algorithm is proposed by us in the item-based collaborative filtering recommendation model using a new energy method to predict the missing ratings of individuals, after predicting the missing ratings of the cluster. Finally, recommend the most relevant items to the user who needs the recommendation based on the predicted rating of the item cluster. The proposed item-based clustering recommendation model was evaluated on both Jester5k and MovieLens by using the MAE_(c) error and the Precision_(c) - Recall_(c) precision value. In general, the cluster proposed model based on the energy distance gives a smaller error than the Peasion and Cosine-based comparison model for both datasets in the "rrecsys" package; and the accuracy of the proposed model is higher than the accuracy of the models using the matching, euclidean, and karypis measures in the "recommenderlab" package. Therefore, the item-based clustering recommendation model using the energy distance presents the feasibility of applying potential energy to cluster in the recommendation problems.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems. a survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, vol 17, 734–749 (2005)
2. Boratto, L B., Carta, S.: State-of-the-art in group recommendation and new approaches for automatic identification of groups", In G. A. Alessandro Soro, Eloisa Vargiu and G. Paddeu, editors, Information Retrieval and Mining in Distributed Environments. Springer Verlag. In press, (2010)
3. Tran, T.C.T., Phan, L.P., Huynh, H.X. (2023). A Group Clustering Recommendation Approach Based on Energy Distance. In: Dinh, T.N., Li, M. (eds) Computational Data and Social Networks . CSoNet 2022. Lecture Notes in Computer Science, vol 13831. Springer, Cham. https://doi.org/10.1007/978-3-031-26303-3_9. (2022)
4. Tran, T, C, T., Phan P. L., Huynh, X, H.: Energy-based collaborative filtering recommendation, International Journal of Advanced Computer Science and Applications(IJACSA), 13(7), 557–562. (2022)
5. Boratto, L B., Carta, S., Satta, M.: Groups Identification and Individual Recommendations in Group Recommendation Algorithms, pp. 27–34. CEUR Workshop Proceedings, (2010)
6. Li, S., Rizzo, L, M.: K-groups: A Generalization of K-means Clustering (2017). ArXiv e-print 1711.04359. https://arxiv.org/abs/1711.04359
7. Li, S.: K-groups: A Generalization of K-means by Energy Distance, Ph.D. thesis, Bowling Green State University (2015)
8. Ntoutsi, I., Stefanidis, K., Norvag, K., Kriegel, HP: gRecs: a Group Recommendation System Based on User Clustering. In: Lee, Sg., Peng, Z., Zhou, X., Moon, YS., Unland, R., Yoo, J. (eds) Database Systems for Advanced Applications. DASFAA 2012. LNCS, vol 7239. Springer, Berlin, Heidelberg (2012)
9. Felfernig, A., Boratto, L., Stettinger, M., Tkalčič, M.: Algorithms for Group Recommendation. In: Group Recommender Systems. SECE, pp. 27–58. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75067-5_2
10. Dara, S., Chowdary, C.R., Kumar, C.: A survey on group recommender systems. J. Intell. Inf. Syst. **54**, 271–295 (2020)
11. Sarwar, B.M., Karypis, G., Konstan, J. A., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web (WWW '01). Association for Computing Machinery, New York, NY, USA, pp. 285–295. (2001)
12. Li, C., Ma, L.: Item-based Collaborative Filtering Algorithm Based on Group Weighted Rating, 2020 13th International Symposium on Computational Intelligence and Design (ISCID), pp. 114–117. Hangzhou, China (2020)
13. Felfernig, A., Boratto, L., Stettinger, M., Tkalčič, M.: Evaluating Group Recommender Systems. In: Group Recommender Systems. SECE, pp. 59–71. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75067-5_3
14. Çoba, L., Zanker, M., Symeonidis, P.: Environment for Evaluating Recommender Systems, https://rdrr.io/cran/rrecsys/. Repository CRAN, (2019)
15. Hahsler, M.: recommenderlab, A Framework for Developing and Testing Recommendation Algorithm (2015)
16. Rizzo, M., Székely, G.: Energy distance. Wiley Interdisciplinary Reviews: Computational Statistics **8**(1), 27–38 (2016)
17. https://rdrr.io/cran/recommenderlab/man/Jester5k.html. Accessed on 01 Feb 2021
18. https://rdrr.io/cran/recommenderlab/man/MovieLense.html