Journal of Scientific & Industrial Research
Vol. 80, February 2021, pp. 159-165

A New Approach for Movie Recommender System using K-means Clustering and PCA

Vikash Yadav[1]*, Rati Shukla[2], Aprna Tripathi[3] and Anamika Maurya[4]

*[1]ABES Engineering College, Ghaziabad, Uttar Pradesh, India
[2]GIS Cell, MNNIT Prayagraj, Allahabad, Uttar Pradesh, India
[3]VIT Bhopal, Madhya Pradesh, India
[4]Harcourt Butler Technical University, Kanpur, Uttar Pradesh, India

Received 07 August 2020; revised 15 October 2020; accepted 09 December 2020

Recommendation systems are refining mechanism to envisage the ratings for items and users, to recommend likes mainly from the big data. Our proposed recommendation system gives a mechanism to users to classify with the same interest. This recommender system becomes core to recommend the e-commerce and various websites applications based on similar likes. This central idea of our work is to develop movie recommender system with the help of clustering using K-means clustering technique and data pre-processing using Principal Component Analysis (PCA). In this proposed work, new recommendation technique has been presented using K-means clustering, PCA and sampling with the help of MovieLens dataset. Our proposed method and its subsequent results have been discussed and collation with other existing methods using evaluation metrics like Dunn Index, average similarity and computational time has been also explained and prove that our technique is best among other techniques. The results achieve from the MovieLens dataset is able to prove high efficiency and accuracy of our proposed work. Our proposed method is able to achieve the MAE of 0.67, which is better than other methods.

Keywords: Average Similarity, Collaborative filtering, Local Multidirectional Score Pattern, MovieLens, Root Mean Squared Error

Introduction

Recommendation systems are widely used in almost every field of intelligent systems and e-commerce applications. The recommendation system has the ability to provide reliable and accurate information to the particular user.[1–5] Recommendation system accumulates the related information and suggestions to the users. Example such as, if new user login to paytm website, then its recommender system tries to find the choices and behaviour of the user, further if it is not able to produce enough suggestions, then this kind of situation is called cold-start situation.[6–8] In that condition, system is unable to recommend new comer's suggestion and preferences, however as fast as user goes through the website and browsing the paytm pages, then he received so many recommendations as per its choices and likes. Various business applications like online shopping, web intelligence, tourism and entertainment utilizes the mechanism of the recommender systems for the growth of their business and the development of this mechanism becomes challenging task for the researchers. The accuracy and efficiency is still going towards in its full-fledged state. The recommendation systems are well depends on filtering like collaborative, content-based, context-based, social and hybrid based. Collaborative filtering (CF) is very popular method used in the today's era of recommendation systems.[9–13] The preference suggestion for users and group of users is given by the ratings of the user in the collaborative filtering method. For example, e-commerce website, user has suggested the ratings of various products like juicer- mixer-grinder, LED TV, mobile, then the recommendation system will provide the suggestion for the given products. The decision in the in the collaborative filtering is based on previous choices of users.[14,15] If a user wants to purchase mobile phone, then recommendation system will advise the same related product to the user, RS has learn the preferences of the particular user. The recommendation systems which are used to provide human related preferences such as age, country, sex, DOB are known as demographic-based recommendation systems.[16–18] The combination of demographic, collaborative and context with various

*Author for Correspondence
E-mail: vikash.yadav@abes.ac.in

computational intelligence and machine learning methods are used in the recommendation systems now days. The challenging task of recommendation systems is always its accuracy and efficiency, because so many algorithms from the domain like machine learning, data mining and artificial intelligence are working effectively. The data mining, machine learning and Artificial intelligence are most efficient and effective technique to enhance accuracy and efficiency of recommendation systems. The techniques such as Geometric Features, Support Vector Machines, and Local Multidirectional Score Pattern (LMSP) descriptor have shown best result in the field of image processing, HCI and medical imaging. These techniques too can be put in intelligent recommendation systems to increase the accuracy, efficiency and reliability. Our proposed recommendation system used the method of sampling which was taken from the MovieLens dataset, freely available. Later on, PCA has been applied to the sampled data to reduce the dimensions of the dataset. Generally, PCA has been used to reduce the dimensions of the image. The reduced data is always a best option to train and test the model. That is why, it is commonly used in image processing and medical imaging. Further, k-means clustering techniques have been employed to database for classification of users by making use of preferences taken from other users. The major findings of our paper are as follows:

(1) We proposed novel collaborative-based recommender systems by using new algorithm.
(2) We applied sampling, PCA and K-means clustering algorithm to increase the accuracy and therefore decrease the error.
(3) We able to increase the performance of K-means clustering by incorporating sampling and PCA.
(4) Our proposed method is evaluated using standard and free MovieLens dataset.
(5) We used SD, RMSE, t-value, MAE, Dunn Index, average similarity and computational time to analyse the performance of the system.
(6) Our proposed method has better MAE of 0.67 and better simulation time, which indicates the better performance when compared with other existing methods.

Related Works

It has been found that more than 1 billion people surfing the website these days. Among these internet users, most of them are used to involve in the one or more online groups either directly or indirectly to the e-commerce websites.[19] The virtual spaces available in the internet are used to produce communities for the purpose of similar types of user's interaction and information sharing. The user from the entire globe can join the community and community compete each other to attract the user to join. The quantity of end users in community indicates success of the community. This success can be achieved by sharing and recommendation of more useful information to the user.[20] The sharing and recommendation of the useful information is the capability of functionality of organisation. Therefore, a need of well-structured management which is capable of sharing useful information to the management so that it can decrease efforts and increase the sharing capability of the community.

A recommender system is applied to achieve useful information of items, information and assistance to the user by combining preference from the other users, from different authorities and from user attributes.[21] The recommender systems are of two types: collaborative and content-based. The product is selected in collaborative filtering generally built on the relation between current users and past users of the product and services. The items are suggested in content-based filtering based on relation among user and its preferences of the product and services.[22] The main difference between both the techniques is that the similarity of preferences between users has been considered in collaborative method while in content-based, similarity of preferences between user and item is considered. Besides these two filtering technique, there is another filtering technique which is the combination of content-based and collaborative filtering.

Content-based Filtering

This technique maps available items with the user's preferences and items rating in the past to evaluate best suggestions for the user.[23] This method is different from information retrieval (IR) and information filtering (IF). In IR, all information is retrieved with the help of keywords specify by the user. The information is suggested automatically in the content based filtering. In this technique useful data is stored first either it is taken explicitly or implicitly from the items profile. After analysis of stored data, the system is made suggestion to the user similar to their preferences. Content-based filtering is the subset of the information retrieval.[24]

Yih *et al*. proposed a system which represents items and user's preferences in the form of vector.[25] Further these vectors are also used to represent user's profile and these vectors are processed to get the similarities between items. Arnold *et al*. introduced a framework which used the information retrieval technique in their content-based filtering called Boolean search indexes.[26] They combine keywords using Boolean operators in their recommendation procedure. Lee *et al*. introduced a technique based on another information retrieval method called probabilistic method. They used the probabilistic method to calculate the probability of the document to check whether it is met with the user's need or not.[27] Siddiqui *et al*. proposed a system based on natural language processor which can used to retrieve data from natural sentences.[28]

Many recommendation systems mostly used content-based filtering to suggest different relevant information to the user. Trivedi *et al*. proposed a system called Stuble Upon which was used to suggest user during web browsing. They usually tracked the user likings and browsing history to suggest the appropriate pages to the user.[29] This type of filtering is very popular in music recommender systems. Petersen *et al*. introduced a recommender system to the music world called Last.Fm. It was built from ratings given by user to particular music.[30] Liu *et al*. introduced the recommendation of news called Google news based on both content-based and collaborative filtering for recommendation of the users. They identify user click behaviour to recommend the likings of the articles to the users.[31]

Collaborative-based Filtering

The collaborative-based filtering is based on users having similar likings while content-based is based on likings of similar items. The recommendation for the active users is provided on the basis of similar users and their interests.[32] A group has been formed with the users of same interests. The filtration of information can be done in any source. The relations between user and item are found to be more complex and vast. This feature makes it more powerful as it is able to distinguish between good and bad documents. They are more accurate than other filtering techniques because it takes the benefits of real user rankings instead of machine based rankings.[33] In a music recommendation system, user going to listen classical songs with bad audio quality and user wind up with no more classical songs. The second user also not happy with classical songs but likes few classical songs then this suggestion go to the first user because they both like classical songs. This creates new group and community, which is not possible with the content-based filtering.

Proposed Method

In the present section, we have presented recommendation system with sampling, PCA and K-means clustering technique. The k-means clustering is very popular approach under unsupervised learning. The PCA is used in our research because it works efficiently and produce good results as compare to other existing methods.

The diagram of proposed recommendation system is shown in Fig. 1. The population from the MovieLens dataset is taken. The sampling is very popular technique used to extract some relevant data from the population. Sampling is used because working on full data is too expensive. The sampling used in this research is random sampling without replacement. The concept behind random sampling is that every item has equal probability and without replacement means when an item is withdrawn, it is removed from the population. In Recommender system, dataset having features with huge dimensional space but also very vast information in that space. The features of every object have less significance in huge feature space. The distance between data points and density are less significant in clustering of huge data space. In a matrix of millions of rows and columns, most of them are zeros. To reduce this meaningless data, we use PCA for reduction of dimensions. PCA is a method to calculate patterns in big datasets. PCA analysis has been done in a 2d dimensional space using Gaussian. The two principal components are obtained after centralising of data. The

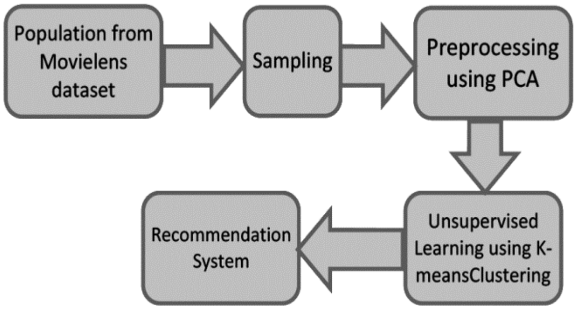
Population from Movielens dataset
Sampling
Preprocessing using PCA
Unsupervised Learning using K-meansClustering
Recommendation System

Fig. 1—Block Diagram of Proposed System

first components contain 90% of energy, so removing second component means that only 10% of total information removed from the dataset. Clustering is an unsupervised learning method. The main purpose of technique is to allocate an item to the group so that the items in the group are more identical than that of other groups. This method is also able to produce more meaningful groups in the dataset. In this work, similarity is calculated using Euclidean distance given by:

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \quad \ldots (1)$$

The main aim of clustering method is to minimize intra-cluster length and maximize inter-cluster length. K-means clustering is the most popular partitioned technique. The function partition the given N dataset into k disjoint subsets S_j which contains N_j data points as close as possible to the other points on the basis of similarity measures. Each partition contains N_j data points for the given C_j centroids. In order to make good partition the distance between all the data points with the centroids should be minimized. The objective of our research is to minimized E from Eq. 2 given below.

$$E = \sum_{1}^{k}\sum_{n\in s_j} d(X_n\,, C_j) \quad \ldots (2)$$

where

X_n = nth vector, C_j = Centroid, d = distance measure

The interchange of items takes place between clusters until the value of E can’t decreases further.

Evaluation Metrics

(1) ***Mean Absolute Error (MAE)***: The MAE for movies is calculated by[34]:

$$\text{MAE} = \frac{\sum|P_{IJ} - T_{IJ}|}{X} \quad \ldots (3)$$

where X →total number of movies, P_{IJ}→prediction of I user on J movies and T_{IJ}→movies true prediction.

(2) ***Root Mean Squared Error (RMSE)***: RMSE is very good measurement tools to check the accuracy of the given results. It is calculated by actual and predicted value of system. RMSE can be given by[34]:

$$\text{RMSE} = (\sqrt{(\sum(Predicted - Actual)(Predicted - \ Actual)/N)} \quad \ldots (4)$$

RMSE is the difference between observed and predicted value of the model. RMSE is calculated by square of the difference between predicted and actual value taken square root divided by N.

(3) ***Standard Deviation***: The quantity of clusters is inversely proportional to the SD values. If the SD value increases, the quantity of clusters also decreases. This represents that any changes made in SD value also effects clusters.[34] The SD can be calculated by:

$$\text{S.D.} = \sum_{i}^{k}{=}_{1}\left\{\sum_{j=1}^{\mathrm{i}}{}_{1}\left(\sum_{i=1}^{m}\sqrt{(Expected - Mean\ )X\ (Expected - Mean\ /\ m)}\right)\right\}$$

$$/\text{elements ini} \quad \ldots (5)$$

(4) ***T-value***: The calculation of t-value using MovieLens dataset is given by[34]:

$$\text{T-value} = \sum_{i=1}^{k}\sum_{j=1}^{k}\left(\frac{\overline{x_\iota} - \overline{x_J}}{\sqrt{\frac{SD_i{}^2}{No.of\ element\ in\ i} + \frac{SD_j{}^2}{No,of\ element\ in\ j}}}\right) \quad \ldots (6)$$

The t-value totally depends on the standard deviation and means of the different clusters.

(5) ***Dunn Index***: Dunn index is used to assess the potential of the clusters based on its internal clusters. Higher value of dunn index represents the superior potential of the clusters. Dunn index is given by[35]:

$$\text{Dunn index(U)} = \min_{1\le i\le c}\left\{\min_{1\le i\le c}\left\{\frac{\partial(X_i\,,X_j)}{\max\limits_{1\le i\le c}\{\triangle(x_k)\}}\right\}\right\} \quad \ldots (7)$$

(6) ***Average Similarity***: This indicates similarity factor among two clusters. Cosine similarity factor is calculated by taking cosine between two non-zero vectors inclined at an angle θ is given by[35]:

$$A.B = ||A||\ .\ ||B||\ .\cos(\theta) \quad \ldots (8)$$

Cosine similarity between two vectors is given by using dot product and their magnitude is given by:

$$\text{Similarity} = \cos(\theta) = \frac{A\,.B}{||A||\,.||B||}$$

$$= \frac{\sum_{i=1}^{n} A_i . B_i}{\sqrt{\sum_{i=1}^{n} A_i{}^2} . \sqrt{\sum_{i=1}^{n} B_i{}^2}} \quad \ldots (9)$$

(7) ***Computational Speed***: It is total run time of algorithm to produce the output.[35] If start time is S and end

time is E in the simulation then the computational time is given by:

$$PT = E-S \qquad \ldots (10)$$

Results and Discussion

In this present section, we present rigorous analysis of our experimental results obtained from proposed recommender systems using MovieLens dataset.[36] We run all the experiments in the computer which has a configuration of I5 processor, 6GB RAM, and python 3.8.1 environment.

We considered the MovieLens database freely available online for the analysis of results of our proposed recommender systems, developed by Minnesota University under the project of GroupLens research. This database contains 1 lakhs ratings (1–5) and having 943 users rated 1682 movies. The constraint on each user rated to rate minimum 20 movies. We used some evaluation metrics in order to evaluate performance our recommender system. We calculated SD, RMSE, t-value, MAE, Dunn Index, average similarity and computational time to understand the performance our recommender system.

In Table 1, we collate our proposed work results with the existing work. The mean absolute error of our proposed work is 0.67, which is better than the GAKM cluster, KM-PSO-FCM, PCA-GAKM, PCA-SOM and PCA shown in Fig. 2. The results of all methods are not good when compare with our proposed work. The results of other methods such as RMSE, SD and t-value are not better than our proposed method.

In Table 2, methods described in previous table are also comparing with respect to dunn index and average similarity. The value of dunn index and average similarity is 0.34318 and 0.96 which is better than the other existing methods as clearly shown in Fig. 3. The results in the table indicate the better performance of our proposed work.

In Table 3, comparison with respect to processing speed is depicted. The result is far better than the other existing methods irrespective of PCA-GAKM and PCA. In these two cases, processing speed is low from our proposed work (Fig. 4). The results indicate that our proposed movie recommender system is better to give recommendations to the user and group of user.

Table 1 — Comparison between different metrics with our proposed method

Method	RMSE	MAE	SD	t-value
PCA-SOM	3.46029	1.96	0.33554	7.92395
KM-PSO-FCM	1.30643	0.74	0.12658	2.98925
GAKM Cluster	1.37705	0.78	0.13343	3.15101
PCA	3.51325	1.99	0.34042	8.03919
PCA-GAKM	1.62421	0.78	0.15737	3.71637
Proposed Method	0.67	1.16520	0.11290	2.66619

Table 2 — Comparison of Average similarity and Dunn Index measures in existing methods with our proposed method

Method	No. of Clusters	Dunn Index	Average Similarity
KM-PSO-FCM	4	0.30609	0.97
PCA-SOM	4	0.11547	0.95
PCA-GAKM	4	0.24620	0.96
GAKM Cluster	4	0.29038	0.97
PCA	4	0.11381	0.95
Proposed Method	4	0.34318	0.96

Table 3 — Comparison of processing speeds between different methods with our proposed method

Method	Processing Speed in sec
PCA-SOM	153.78
KM-PSO-FCM	143.31
GAKM Cluster	329.43
PCA	27.65
PCA-GAKM	49.05
Proposed Method	53.35

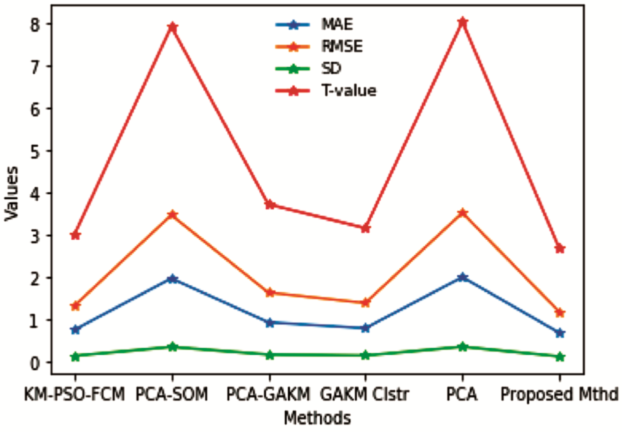

Fig. 2—Comparison of different methods with our proposed method

Conclusions and Future Work

Recommender systems are very popular in organisation to recommend items to user and group of users from the billions of items in the field of e-commerce and other organisations. They are also able to predict the behaviour of modern user to launch new items and to analyse existing items. We are also able to proposed efficient, reliable and effective collaborative movie recommendation system. Our results also showed that performance of our proposed

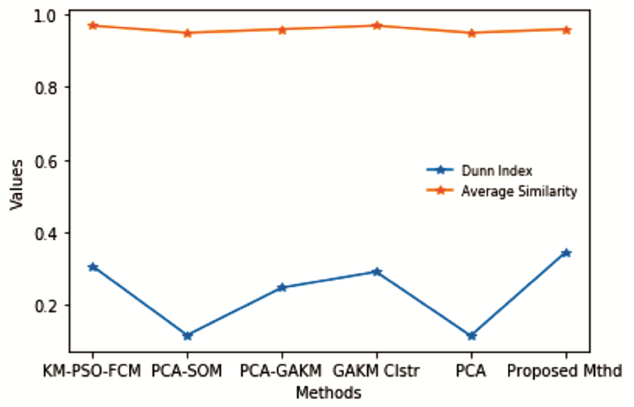

Fig. 3 — Comparison of Dunn index and average similarity with our method

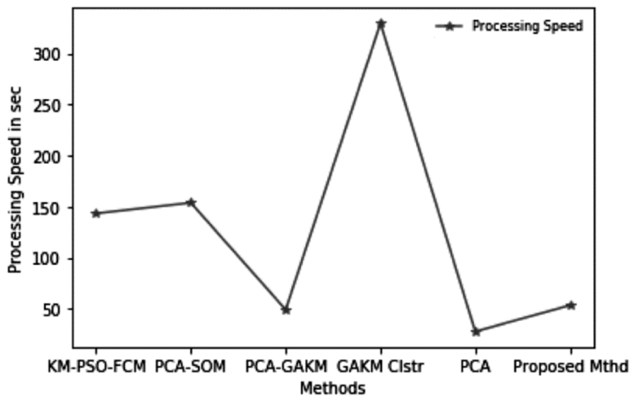

Fig. 4 — Comparison of Processing speeds with our proposed method

recommender system works better than other existing methods. The values from our proposed system for standard deviation (SD), root mean square error (RMSE), mean absolute error (MAE), t-value, Dunn Index, average similarity and computational time is 0.11290, 1.16520, 0.67, 2.66619, 0.34318, 0.96 and 53.35 respectively. Our proposed work with respect to processing time is also better than the existing methods. In future, we will incorporate some more feature extraction, supervised learning and deep learning techniques to improve accuracy of recommender systems. We will also apply this proposed work to big data and try to improve its privacy and efficiency.

References

1 Katarya R & Verma O P, A Collaborative Recommender System enhanced with particle swarm optimization technique, *Multimed Tools Appl*, **75(15)** (2016) 9225–9239.

2 Ortega F, Hernando A, Bobadilla J & Kang J H, Recommending items to group of users using Matrix Factorization based Collaborative Filtering, *Inf Sci*, **345** (2016) 313–324.

3 Bobadilla J, Ortega F, Hernando A & Gutie´rrez A, Recommender Systems Survey, *Knowl-Based Syst*, **46** (2013) 109–132.

4 Katarya R & Verma O P, Recent developments in affective recommender systems, *Phys A Stat Mech Appl*, **461** (2016) 182–190.

5 Katarya R, Jain I, Hasija H, An Interactive Interface for Instilling Trust and providing Diverse Recommendations, *IEEE Int Conf Comput Commun Technol (ICCCT)* (2014) 17–22.

6 Ji K & Shen H, Jointly Modelling Content, Social Network and Ratings for Explainable and Cold-start Recommendation, *Neurocomputing*, **218** (2016) 1–17.

7 Zhao W X, Li S, He Y, Chang E Y, Wen J-R & Li X, Connecting social media to e-commerce: cold-start product recommendation using microblogging information, *IEEE Trans Knowl Data Eng*, **28** (2016) 1147–1159.

8 Son L H, Dealing with the new user cold-start problem in recommender systems: A comparative review, *Inf Syst*, **58** (2016) 87–104.

9 Da Silva E Q, Camilo-Junior C G, Pascoal L M L & Rosa T C, An evolutionary approach for combining results of recommender systems techniques based on collaborative filtering, *Expert Syst Appl*, **53** (2016) 204–218.

10 Zhou Q, "Supervised Approach for Detecting Average over Popular Items Attack in Collaborative Recommender Systems", *IET Inform Secur*, **10(3)** (2016) 134–141.

11 Yang Z, Xu L & Cai Z, Re-scale AdaBoost for Attack Detection in Collaborative Filtering Recommender Systems, *Knowl-Based Syst*, (2016) 1–32.

12 Liang X, Xia Z, Pang L, Zhang L & Zhang H, Measure prediction capability of data for collaborative filtering, *Knowl Inf Syst*, **49(3)** (2016) 975–1004.

13 Hernando A, Bobadilla J & Ortega F, Anon Negative Matrix Factorization for Collaborative Filtering Recommender Systems based on a Bayesian Probabilistic Model, *Knowl-Based Syst* **97** (2016)188–202.

14 Xu Y & Yin J, Engineering Applications of artificial intelligence collaborative recommendation with user generated content, *Eng Appl Artif Intell*, **45** (2015) 281–294.

15 Puglisi S, Parra-Arnau J, Forne´ J & Rebollo-Monedero D, On Content-based Recommendation and user privacy in social-taggingsystems, *Comput Stand Inter*, **41** (2015) 17–27.

16 Zhao W X, Li S, He Y, Wen J-R & Li X, Exploring demographic information in social media for product recommendation, *Knowl Inf Syst*, **49(1)** (2015) 61–89.

17 Katarya R & Verma O P, "Restaurant recommender system based on psychographic and demographic factors in mobile environment", *IEEE Inte Conf Green Comput Internet things* (2015) 907–912.

18 Al-Shamri M Y H, User profiling approaches for demographic recommender systems, *Knowl-Based Syst*, **100**, 1–13.

19 Albors J, Ramos J C & Hervasa J L, New learning network paradigms: Communities of objectives, crowd sourcing, wikis and open source, *Int J Inf Manag*, **28(3)** (2008) 194–202.

20 Arguello J, Butler B, Joyce E, Kraut R, Ling K S, Ros C & Wang X, Talk to me: Foundations for successful individual–group interaction in online communities, in *Proc SIGCHI*

Conf Human Factors in Comput Syst Montreal, Quebec, Canada,(2006) 959–968.

21 Frias-Martinez E, Magoulas G, Chen S Y & Macredie R, Automated user modeling for personalized digital libraries, *Int J Inf Manag*, **26(3)** (2006) 234–248.

22 Su X & Khoshgoftaar T M, A Survey of Collaborative Filtering Techniques, *Artif Int*, **2009** (2009) 1–19.

23 Ferman A M, Errico J H, Beek P V, Sezan M I, Content-based filtering and personalization using structured metadata, in *Proc 2nd ACM/IEEE-CS joint conf Digit libr,* 2002, 393–393.

24 Jannach D, Zanker M, Felfernig A & Friedrich G, Recommender systems: an introduction, Cambridge University Press, 2010.

25 Yih W T, Learning term-weighting functions for similarity measures, in *Proc Conf Empirical Method Nat Language Proces*, **2** (2009) 793–802.

26 Arnold J F & Voss L L, US Patent No. 6,745,161, Washington, DC: U.S. Patent and Trademark Office, 2004.

27 Lee C & Lee G G, Probabilistic Information Retrieval Model for a Dependency Structured indexing system, *Inf Process Manage*, **41(2)** (2005) 161–175.

28 Siddiqui T J & Tiwary U S, Utilizing local context for effective information retrieval, *Int J Inf Tech Decis* **7(1)** (2008) 5–21.

29 Trivedi A, Rai P, Du Vall S & Daumé III H, Exploiting tag and wordcorrelations for improved webpage clustering, in *Proc 2nd Int Workshop on Search and Mining user-generated contents* (2010) 3–12.

30 Petersen M K, Hansen L K, Emotional Nodes among lines of lyrics, in *Automatic Face & Gesture Recognition and Workshops,* (2011) 821–826.

31 Liu J, Dolan P, Pedersen E R, Personalized News Recommendation based onclick behaviour, *In Proceedings of the 15th international conference on Intelligent userinterfaces,* (2010) 31–40.

32 Adomavicius G, Tuzhilin A, Toward the next generation of recommender systems: A Survey of the state-of-the-art and possible extensions, *IEEE T Knowl Data En*, **17(6)** (2005) 734–749.

33 Ricci F, Rokach L & Shapira B, Introduction to recommender systems handbook (2011)1–35, Springer US.

34 Rukmi, A M & Iqbal I M, Using k-means++ algorithm for researchers clustering, in *AIP Conf Proc,* AIP Publishing: Melville, NY, USA, **1867** (2017) 020052-1–020052-5.

35 Putri D C G, Leu J-S & Seda P, Design of an Unsupervised Machine Learning-Based Movie Recommender System, *Symmetry*, **12(2)** (2020) 1–27.

36 http://grouplens.org/datasets/movielens/.