Missed Data Prediction and Movie Recommendation Using Hybrid Linear Regression Based Pearson Correlation Coefficient Method

V Durga Devi[1], P Jayapriya[2], R Priya[3]
Department of Computer Application[1], Department of Computer Science,[2], Department of Computer Application[3]
Vistas1, Ethiraj College For Women[2], Vistas3
Email: durga_vetri33@yahoo.com[1], jayapriya1972@gmail.com2, priyaa.research@gmail.com[3]

Abstract- The main goals of a web data clustering algorithm are to produce appropriate clusters for the end user, to assign the available data to the most relevant cluster, to respond the end user instantly. In this paper, we propose a method namely hybrid linear regression based Pearson correlation coefficient method to cluster movie data. The proposed algorithm is the combination of correlation and linear regression method. The advantage of this new technique is fast operation on dataset containing items and provides facilities to avoid unnecessary scans to the database.
.

Index Terms- Movie recommendation; data prediction; similarity measure; collaborative filtering.

1. INTRODUCTION

The exponential growth of the data on www is a challenge to Search Engines. User needs to use information retrieval tools. In any information retrieval system ranking plays a main Role. Most of the Search engines return million of pages for a given query, It is highly impossible for a user to preview all the returned results, ranking is helpful in web searching. Based on content and connectivity, ranking is divided into two categories. Content based ranking is depends on content of web page, Connectivity ranking based on link analysis technique[1]. Optimization techniques play an important role in the field of science and engineering. Over the last five decades, numerous algorithms have been developed to solve complex optimization algorithms. Since more and more present-day problems turn out to be nonlinear, multimodal, discontinuous, or dynamic in nature, derivative-free, no exact solution methods attract ever-increasing attention. Evolutionary biology or swarm behaviors inspired most of these methods [2].

The evolution of the Internet into a global information infrastructure has lead to an explosion in the amount of available information. If one were to look at the Web as a distributed, heterogeneous information base, Web personalization amounts to creating a system that responds to user queries based on information about him/her. As a trivial example, a biologist querying on cricket in all likelihood wants something other than what a sports enthusiast would. The explosion of the Web has brought the personalization problem front and center, not just because of the amount of time wasted in searching for information but also because of the massive traffic surge this has generated in the Internet backbone [3].

There are many online recommendation systems, like Netflix.com and Amazon’s similar-product recommendations, which show top results of a data mining back end. They generate thousands of recommendations, but are constrained to show only the top 3 or 4 out of them[4]. Other sites like Jinni and Pandora have more interactive and visually appealing results but these use datasets specifying the genome of the song/film which are painstakingly made by hand. Many a times these results are good, but they do not give any feedback as to why a movie is being recommended.

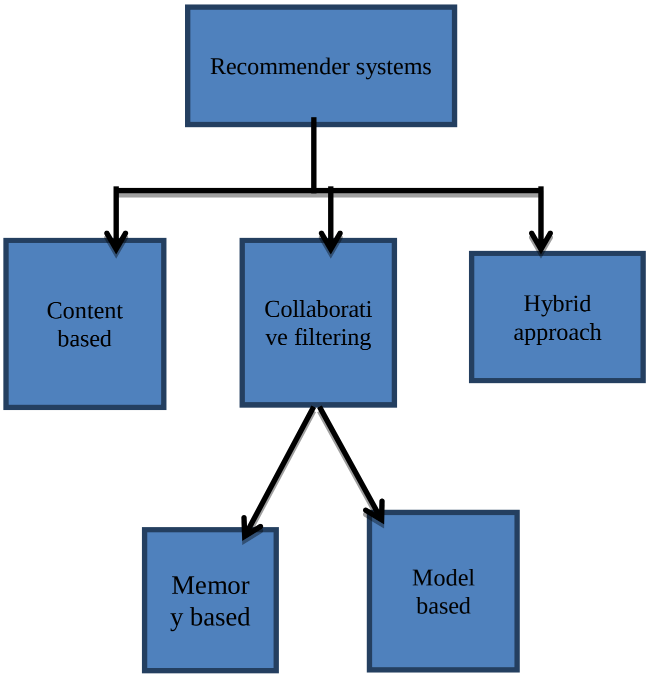

Figure- 1Basic types of recommendation systems

Many recommender systems, algorithms or methods have been presented so far. Initially, the majority of research effort was spent on the collaborative systems and explicit user feedback. Although collaborative recommender systems are generally trusted to be more accurate, they suffer from three well known problems: cold start, new object and new user problem[5]. New user / object problem is a situation, where recommending algorithm is incapable of making relevant prediction because of insufficient feedback about current user / object. The cold start problem refers to a situation short after deployment of recommender system, where the system cannot provide any relevant recommendation, because of insufficient data generally. Using attributes of objects and hence content based or hybrid recommender systems can speed up learning curve and reduce the cold start problem. Moreover, content-based recommender systems can compute similarity of a new object based on its features effectively eliminating the new object problem[6].

The exploration of large data sets is a difficult problem. Though clever visualization techniques partly help to solve the problem, still sometimes it is easier for humans to look at computed patterns and a good visualization of the data to find more intricate patterns in the data. Thus, involving the user in the data mining process by preprocessing the data using simple statistical techniques and then presenting the results to the user, visually, is a better idea. Integration of interactive visualization and data mining algorithms combines the intuitive power of the human mind with computational capabilities of the computer, thus improving the quality and speed of the data mining process[7].

The organization of paper is as follows: Section 1 starts with introduction about movie recommendation types, problems in existing process and growth of search engines. Section 2 shows the detailed survey of existing technologies in recommendation systems. Our proposed methodology is explained in Section-3 with some equations .Section-4 explains the performance analysis with graphical representation and comparison with existing techniques. Finally paper is concluded with section 5.

2. LITERATURE SURVEY

Sarwar et al. (2001) [8] divide collaborative filtering into two categories: memory based collaborative filtering algorithms and model-based collaborative filtering algorithms. Memory-based algorithms use all available user-item data to generate a prediction. Based on all data it determines the most related users, similar to the target user. These neighbours are similar because they have statistically common interests. To determine these so-called neighbours, several statistical techniques are used. Finally, the top n most similar items are recommended for the target user. The memory-based collaborative filtering algorithms are also called user-based collaborative filtering algorithms. The advantage of user-based collaborative filtering is the sparsity and scalability. Many recommender systems use data with lots of users and items, but with relatively few number of actual ratings. User-based collaborative filtering only uses necessary data, which reduces the run time.

Model-based collaborative filtering first builds a model of user ratings only. To do this, it uses several machine learning techniques, such as clustering, rule-based and Bayesian network approaches. Each of the machine learning techniques uses its own approach. The clustering model formulates collaborative filtering as a classification problem, while the Bayesian network model treats it as a probabilistic model and the rule-based model as an association-rule model. The model-based collaborative filtering algorithms are also called item-based collaborative filtering algorithms.

Lops et al. (2011) [9] stated that the recommendation process of a content-based recommender system basically consists of matching the attributes of a user profile against the attributes of a content object. The outcome of this process is just the level of the user’s interest in an object. It is crucial for a content-based model that the user profile is accurate.

SongJie Gong and Zhejiang [10], proposes a 'personalized recommendation systems' is widely utilized in e-commerce websites to provide recommendations to its users. This approach is employed to provide recommendations in this project which makes the prediction smoother. In this approach, item clustering is done using the two techniques Pearson correlation technique and Adjusted cosine similarity technique to find the similarity between the items. Then, users are clustered depending on alikeness between the user targeted and cluster center. Users are grouped into clusters based on their likes and dislikes for an item and every cluster has a center. The authors state that the proposed method is more accurate than the traditional method in generating recommendations.

Rahul Katarya et al., (2017) [11] proposed a novel recommender system has been discussed which makes use of k-means clustering by adopting cuckoo search optimization algorithm applied on the Movielens dataset. The approach had been explained systematically, and the subsequent results have been discussed. It is also compared with existing approaches, and the results have been analyzed and interpreted. Evaluation metrics such as mean absolute error (MAE), standard deviation (SD), root mean square error (RMSE) and t-value for the movie recommender system delivers better results as our approach offers lesser value of the mean absolute error, standard deviation, and root mean square error. The experiment results obtained on Movie lens dataset stipulate that the proposed approach may provide high performance regarding reliability, efficiency and delivers accurate personalized movie recommendations when compared with existing methods.

EmrahInan et al., (2018) proposed recommender system combines content information of movie features (cast, director, genre, etc.) with a collaborative filtering approach. The similarity scores of movie features are supplemented by a goal programming model in the content-based approach. Pearson correlation is selected as a collaborative filtering algorithm that predicts movies to satisfy user tastes considering the content-based similarity scores. MovieLens dataset is used for experimental setup and Mean Absolute Error is measured for the comparison of approaches [12]

ShreyaAgrawal et al., (2017) proposed a Hybrid approach by combining content based filtering andcollaborative filtering, using Support Vector Machine as a classifier and genetic algorithm is presented in the proposed methodology and comparative results have been shown which depicts that the proposed approach shows an improvement in the accuracy, quality and scalability of the movie recommendation system than the pure approaches in three different datasets. Hybrid approach helps to get the advantages from both the approaches as well as tries to eliminate the drawbacks of both method [13].

SakshiBansal et al., (2016) proposed some techniques to predict genre of movies based on user's posted movie tweets and recommending movies to users' according to predicted genre. For this purpose, the pre-processed twitter extracted movie tweets using tokenization, porter stemming, stop word removal and use Word-Net dictionary for synonym matching. Further, the Latent Semantic Indexing technique need to apply which in turn involves Singular Value Decomposition on this pre-processed data and predicts genre on the basis of IMDb movie genre categorization [14].

3. RESEARCH METHODOLOGY

3.1 *Clustering phase*

The steps involved in Fuzzy Bat clustering are as follows:

1. Initialize the population size, velocity, frequencies, pulse emission rate, Loudness and the maximum number of iterations.
2. Create an initial population of bats randomly.
3. The membership matrix is initialized with random values between the ranges 0 - 1.
4. Compute the cluster centers for each bat using (1).
5. Estimate the fitness value of each bat and find the current best solution using (2) & (3).
6. For each initial bat solution update the Velocity matrix and the location vector using (5), (6) & (7).
7. Evaluate the new bat solutions using fitness Function and find the current global best as initial cluster Centre and apply Fuzzy C Means Clustering to cluster the users into different groups.

The k-means clustering algorithm can be enhanced by the use of a kernel function; by using an appropriate nonlinear mapping from the original (input) space to a higher dimensional feature space, one can extract clusters that are non-linearly separable in input space. Furthermore, we can generalize the kernel k-means algorithm by introducing a weight for each point a, denoted by w(a). As we shall see later, this generalization is powerful and encompasses the normalized cut of a graph. Let us denote clusters by πj , and a partitioning of points as {πj} k j=1. Using the non-linear function φ, the objective function of weighted kernel k-means is defined as:

$$\mathrm{D}(\{\pi j\}\ \mathrm{k\ j{=}1}) = \sum_{k=1}^{n} \mathrm{w(a)k€(a)} - \mathrm{mj} \qquad (1)$$

$$\mathrm{Mj} = \frac{\sum_{b=0}^{n} w(\mathrm{b})€(\mathrm{b})}{\sum_{b=0}^{n} w(b)} \qquad (2)$$

Note that mj is the "best" cluster representative, Since (3)

$$\mathrm{mj = argminz\ X\ a{\in}\pi j\ w(a)k\varphi(a) - zk}$$

e computed using kernel function κ , and are contained in the kernel matrix K. All computation is in the form of such inner products, hence we can replace all inner products by entries of the kernel matrix.

3.2 ***Missing data prediction phase***

The hybrid linear regression based Pearson correlation coefficient method is used for user and item similarity calculations. The hybrid linear regression based Pearson correlation coefficient takes the factor of the differences in user rating styles into account. However, the hybrid linear regression based Pearson correlation coefficient overestimates the similarities of users who happen to have rated a few items identically, but may not have similar overall preferences. Thus, we adopt the solution of linear regression weighting factor in order to devalue the similarity weights that are based on a small number of coated items.

The variable whose value is influenced is called dependent variable and the variable which influences the values is defined as independent variable. In regression analysis independent variable is also known as regress or predictor while the dependent variable is known as regressed or explained variable. Simple linear regression analysis is a technique used for estimating the unknown value of a dependent variable from the known value of independent variable. In other words, X and Y are two related variables, then linear regression techniques helps to estimate the value of Y for a given value of X. Similarly, estimate the value of X for given value of Y.

3.3 *Similarity indexing phase:*

A similarity measure is a relation between a pair of objects and a scalar number. Common intervals used to mapping the similarity are [-1, 1] or [0, 1], where 1 indicates the maximum of similarity. The similarity is defined as average of attributes similarities. Attribute similarity is defined according to attribute type. Similarity of numeric attributes is defined as their difference normalized by maximal allowed distance.

$$\text{Sim (x,y,maxidistance)} = \text{maxi}(0, \frac{\text{maxidistance}-/\text{x}-\text{y}/}{maxisistance})$$

For string attributes (movie name) the similarity is defined as inverse of relative distance

$$\text{Sim(x,y)} = 1- \frac{length(x,y)}{\max\text{imum}(length(x),length(y)} \quad (5)$$

$$\text{Sim(x,y)} = /\text{x n y}/*/\text{x U y}/ \quad (6)$$

The annotation wrapper can prepared with the label, prefix, suffix and unit index datas collected from the search result records. Moreover, the annotation is used to link the web page with another web page for which the best ranking is made. So, one can make use of the annotation wrapped along with web search result and thus enabling the user with more convenient way to gain knowledge from their required area.

4. EXPERIMENTAL SETUP

To test the recommendation performance, we use a Movie lens dataset . It was collected by the secondary data such as online, newspapers and tweets, social media for the research work in the field of recommender system. The dataset includes 943 users, 1682 movies and 1,00,000 ratings. In this dataset, there are three type of information's are available such as demographic information of users, information about movies and rating score of that movies. The performance of the proposed hybrid linear regression based Pearson correlation coefficient recommender system is evaluated with existing random forest algorithm.

4.1 *Performance analysis*

- Mean Absolute Error (MAE) is a measure that represents the positive and negative deviations between the predicted and the observed values [7].

$$\text{MAE(y; ^y)} = \frac{1}{n}\sum_{i=1}^{n}(y(i) - y * (i)) \quad (8)$$

- Mean Squared Error (MSE) measures the accuracy of machine learning algorithms. Mathematically, it is defined as

$$\text{MSE(y; ^y)} = \frac{1}{n}\sum_{i=1}^{n}(y(i) - y * (i))\text{^}2 \quad (9)$$

- Root Mean Squared Error (RMSE) or root-mean-square deviation (RMSD) measures the average squares of the errors. Mathematically, it is defined as:

$$\text{RMSE} = \sqrt{\sum \frac{(Y-^{Y})\text{^}2}{n}} \quad (10)$$

parameters	**hybrid linear regression based Pearson correlation coefficient**	**Random forest**
Mean Absolute Error (MAE)	3.76	0.561
Mean Squared Error (MSE)	2.987	1.325
Root Mean Squared Error (RMSE)	3.12	1.43
Accuracy	79.32 (79%)	6.84 (68%)

5. CONCLUSION

In this paper, we propose an effective missing data prediction algorithm for collaborative filtering. By judging whether a user (an item) has other similar users (items), our approach determines whether to predict the missing data and how to predict the missing data by using information of users, items or both. The procedures of replacing missing data with a value of zero or the same value can have an adverse effect by generating outliers and noise in the data. In addition, to remove attributes that contain missing data as this will negatively affect the size of dataset and performance of prediction model.

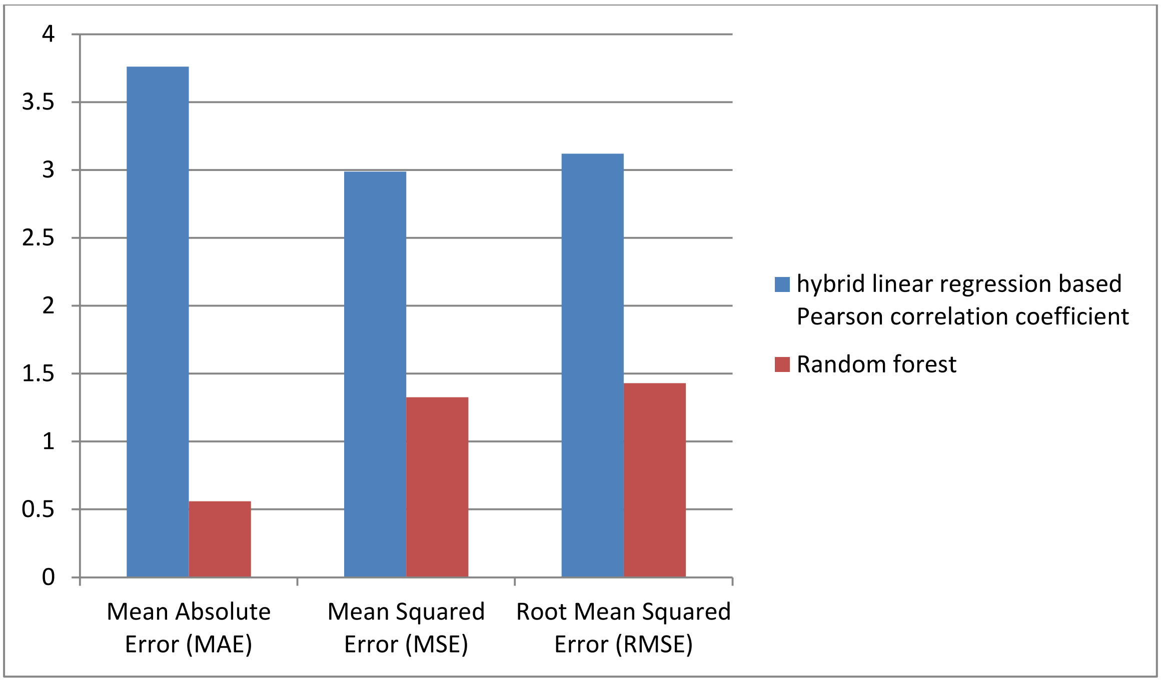

Figure-2 error calculation

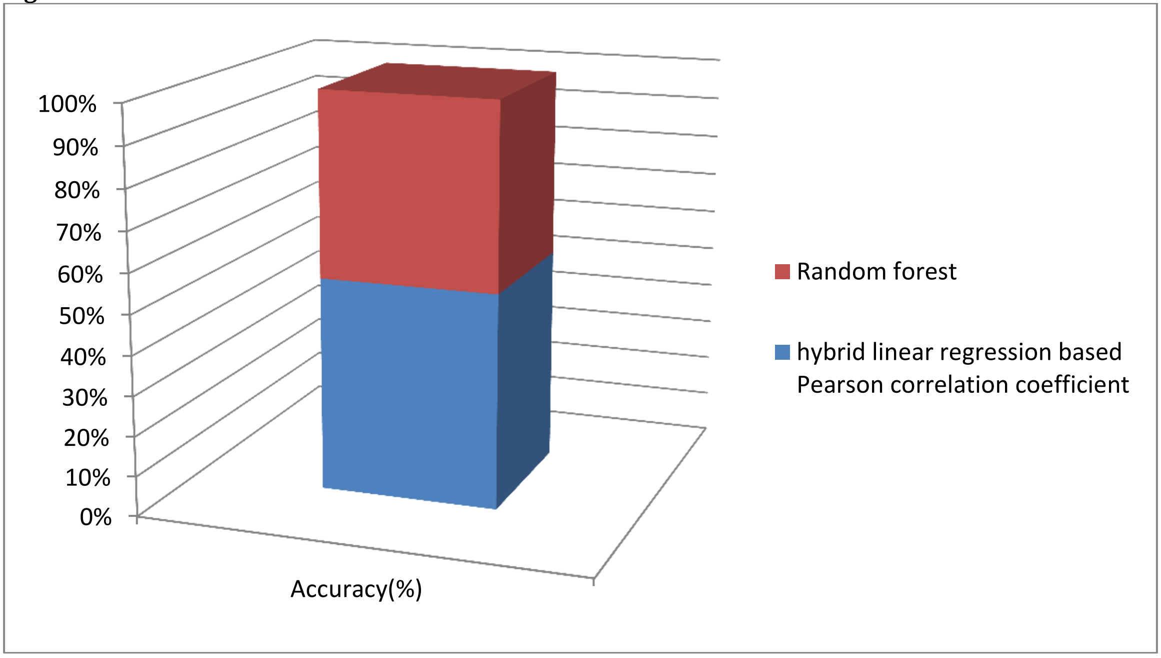

Figure-3 Accuracy calculation

REFERENCE

[1] Algorithm and Proposed Parallel K-means clustering for Large Data Sets.” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.

[2] Akkaya, Kemal, FatihSenel, and Brian McLaughlan. "Clustering of wireless sensor and actor networks based on sensor distribution and connectivity.” Journal of Parallel and Distributed Computing 69, no. 6 (2009): 573-587.

[3] Shafeeq, Ahamed, and K. S. Hareesha. (2012)"Dynamic clustering of data with modified k-means algorithm." In Proceedings of the 2012 conference on information and computer networks, pp. 221-225.

[4] Yang, Q. and Zhang, H. (2003), Web-Log Mining for predictive Caching, IEEE Trans. Knowledge and Data Eng., 15(4), 1050- 1053.

[5] Ristoski, P.; Mencia, E.L. &Paulheim, H. (2014) : A Hybrid Multi-Strategy Recommender System Using Linked Open Data, In ESWC.

[6] Peska, L.; Vojtas, P. (2013) : Enhancing Recommender Systems with Linked Open Data. In FQAS 2013, Springer, LNCS 8132, 483-494

[7] Daniel A Keim. (2002)"Information Visualization and Visual Data Mining".In IEEE Transactions on Visualization & Computer Graphics, Vol 7, JanMar.

[8] Sarwar et al. (2001), “Item-Based Collaborative Filtering Recommendation Algorithms”.

[9]. Lops et al. (2014), “Content-based Recommender Systems: State of the Art and Trend” Proc. 14th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, ACM Press, pp. 426-434.

[10] SongJie Gong, Zhejiang, (2010) “A Collaborative Filtering Recommendation Algorithm based on User Clustering and Item Clustering”. Business Technology Institute.

[11] Rahul Katarya , Om PrakashVerma, (2017) ” An effective collaborative movie recommender system with cuckoo search” Egyptian Informatics Journal 18 105–112

[12] EmrahInana,∗ , FatihTekbacakb , CemalettinOzturk, (2018) ” Moreopt: A Goal Programming based Movie Recommender System” Advances in Soft Computing and Machine Learning in Image Processing, Springer, , pp. 477–495.

[13] ShreyaAgrawal, Pooja Jain, (2017) ”An Improved Approach for Movie Recommendation System”International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)Volume 3, Issue 4

[14] SakshiBansal, Chetna Gupta, AnujaArora,(2016) ”User Tweets based Genre Prediction and Movie Recommendation using LSI and SVD” Eighth International Conference on (pp. 354-359). IEEE.

(A.1)