Abstract: Today’s generation is considered much faster than the previous one, the emergence of new technologies have made the time and an easy approach more preferable. In this scenario the role of recommendations in online as well as offline platform is more considered rather than self-testing. In this paper we are going to propose a movie recommender system (RS) in machine learning using Hybrid Collaborative Filtering which is going to recommend a movie that the user may like. This paper is mainly focused on the online movie recommendation platforms which are highly popular these days. In this paper we are using Collaborative Filtering (CF), K-Nearest Neighbor (KNN) algorithm, Cosine Similarity. The implementation is done in python programming language and the movie data is taken from MovieLens and user data is taken from IMDB 50000 Movie Dataset in the implementation part.

Keywords: Recommender System, Machine Learning, CF, KNN, Cosine similarity.

1. Introduction

The hot topic of the entire online marketing and one of the best technique to make online websites more user convenient is RS. It is used in different fields now-a-days, for recommending movies, books, images, music, clothes, accessories, online purchasing, online marketing etc.

RS is widely used in every online platform. For decreasing human effort and saving time and increasing the users interest towards the digital era, RS is playing a major role.

It is becoming an attention for various researchers and engineers. In 2006 when NetFlix held a competition to promote the advancement and research in the existing RS, it is widely becoming favorite topic among the computer graduates also.

A RS is a system which recommends suggestions and recommendations to the user based on the response of the mass users or the history of the existing user.

There are mainly three types of recommender system:
1. CF RS
2. Content based RS
3. Hybrid RS

Majority of the RS are made using CF because it is one of the most easiest and most basic technique to find recommendations. It also promotes the development of new techniques and methods.

Thus, our project is also based on the CF method with some addition on algorithms and techniques to make the predication more accurate and user friendly. The other algorithms and techniques are the KNN algorithm and cosine similarity.

The sets of data which is used in this proposed work is taken from MovieLens and the user rating is taken from IMDB 5000 Movie dataset for perfect predictions.

Also the concept of machine learning is also used in this paper as we know that all these methods and models are used to train the machine and to make it performable without much interaction of the user. Data mining techniques are also used so that the best of the best recommendation can be given.

2. Literature Survey

The RS is a system that analyzes user behaviour or opinions or wants as an input which will then be taken and the system will provide appropriate recommendations and suggestions to the users directly. The RS is one of the most important things in high level personalization development in E-Commerce. The main purpose of this RS is the development of a system supporting users in making their choices online and providing access to and quality of high level suggestions and recommendations for large number of users. The challenge of the RS is to provide low cost, personal, and quality recommendations in every aspects. Technically, we focus on what technique can use information effectively and efficiently, while psychologically, considering the process of designing user interactions. Therefore, the RS's cannot be said to be a simple theoretical concept.

It combined the KNN and gradient boosting method. Calculated the correlation between users and items with the KNN algorithm to filter out similar users and similar items. The gradient boosting with ensemble learning to predict users’ ratings.

In Hybrid Collaborative Filtering method combined the k-nearest neighbour (KNN) algorithm and gradient boosting method. It combined the KNN algorithm and gradient boosting method. Here, calculated the correlation between users and items with the KNN algorithm to filter out similar users and similar items. The gradient boosting with ensemble learning to predict users’ ratings.

In Gradient Boosting methods, the decision tree algorithm was used everywhere. The decision tree algorithm has many positive uses such as the low time complexity so that results the samples quickly. However, the single decision tree has some shorts. For example, it is easy to over-fitting for the single decision tree method. Although there are other measures to improve it such as pruning to reduce the extent of the over-fitting, the results are not great. There are a lot of things, for example, that the time is so low that you can give the recommendation very quickly.

For an item which has is rated by only one single user between them. Here let’s assume that Iuv= Iu - Iv represents a set of ratings that user u has rated on the items while user v has not rated on them. Use C1 to record the ratings of user u for the items and use C2 to record the ratings of user v the items. We can considerate the ratings of Iuv and Ivu to predict. The formula used to calculate the relation between users is defined as:

\[ m = \frac{\sum_{i \in I_u} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v) + \sum_{i \in I_v - I_u} (r_{vi} - \bar{v})(c - \bar{c})}{\sqrt{\sum_{i \in I_u} (r_{ui} - \bar{r}_u)^2 + \sum_{i \in I_v - I_u} (r_{vi} - \bar{v})^2}} \]  

(1)

\[ n = \sqrt{\sum_{i \in I_u} (r_{ui} - \bar{r}_u)^2 + \sum_{i \in I_v - I_u} (r_{vi} - \bar{v})^2} \]  

(2)

\[ \text{sim}(u, v) = \frac{m}{n} \]  

(3)

Here the cosine function is used to measure the correlation between similar items. For the item p and item q, the cosine function defines as following:

\[ g = \frac{\sum_{i \in I_p} (r_{pi} - \bar{r}_p)\sum_{i \in I_q} (r_{qi} - \bar{q}_q)\sum_{i \in I_p} (r_{pi} - \bar{r}_p)^2}{\sqrt{\sum_{i \in I_p} (r_{pi} - \bar{r}_p)^2}} \]  

(4)

\[ h = \sum_{i \in I_p} (r_{pi} - \bar{r}_p)(r_{qi} - \bar{q}_q) \]  

(5)

\[ \cos(p,q) = \frac{m}{n} \]  

(6)


They have designed a CF hybrid method that merges recommendations provided by different CF approaches based on a multi-class classification algorithm. This classification is performed based on the user rating behavior.

Experimental results demonstrate an improvement on the quality of predictions and recommendations in all the studied scenarios for both Movie Lens and Netflix datasets. The quality of the recommendations provided for a CF based RS to a user depends on both the goodness of the recommender and the features of that user.

Furthermore, the proposed method may reduce the impact of shilling attacks and profile injection attacks. These attacks introduce malicious ratings on a Recommender System to favor or disfavor the recommendation of specific items. How to introduce these malicious items is not trivial: RS implementation details must be known in order to hack it. Proposed method uses several Collaborative Filtering approaches to perform recommendations, so if one of them is attacked, the other ones will still provide reliable recommendations. Moreover, by classifying the users according to its more accurate CF approach, attacks will only affect a small group of users.

This work opens a novel research line in hybrid CF recommendations. The proposed model can be extended to incorporate more features to the users rating behaviour to improve the accuracy of the classification. Additionally, the proposed model can be enhanced by including more CF approaches that represent users not covered by the CF approaches used in this work.


Use of multi criteria for user data ranking to obtain appropriate recommendation results.

The recommendation system can obtain information relevant to user interest by analyzing ratings of users using the CF methods, based on the results of computation issued by the program with similar manual calculations.

The challenge of the RS is to provide low cost, personal, and quality recommendations both technically and psychologically. Technically, we focus on what methods can use information effectively and efficiently, while psychologically, considering the process of designing user interactions. Therefore, the RS cannot be said to be a simple theoretical concept of decision.


This survey introduces the various concepts related to machine learning and RS.

The introduction of various concepts related to machine learning and RS. In this work, various tools and techniques have
been used to build RS. Various algorithms such as K-Means Clustering, KNN, CF, Content Based Filtering have been described in detail. Further, after studying different types of machine learning algorithms then there is an illustration of how implementations and working of the proposed work are used for the implementation of the movie RS.

Finally, in this work for different cluster values, different values of Root Mean Squared Error (RMSE) are obtained. In this proposed work as the number of cluster's decreases, the value of RMSE also decreases. The results given by the proposed system are better than the old techniques on the basis of RMSE value.

RMSE can be calculated by:

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\]

3. Proposed Work

Here the focus lies on a hybrid recommender approach, which combines both content based and collaborative based approaches. It shows that many of the disadvantages of existing systems nullifies by combining them known concepts with new ones.

A hybrid recommender approach contains the use of multiple algorithm at the same time. Algorithm to be used is,

- **K-Nearest Neighbor (KNN)**

The proposed recommender system uses similarity score from cosine similarity which is based on user's liking recommending similar movie to the user.

Here the data require by the system is the data of the movies with the rating of the previous users who already rated that particular movie previously and that data is used for the prediction of the present user. After providing the data it goes to the process stage where the sorting of data is done according to different tastes of different users. After the sorting of data the proposed method is being imposed to the system which is Hybrid Filtering which further also contains the data mining to find the best result and then the result is recommended to the user.

A. Algorithm of KNN used in proposed work

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Load the data</td>
</tr>
<tr>
<td>2.</td>
<td>Initialize K to your chosen number of neighbors</td>
</tr>
<tr>
<td>3.</td>
<td>For each example in the data:</td>
</tr>
<tr>
<td>3.1</td>
<td>Calculate the distance between the query example and the current example from the data.</td>
</tr>
<tr>
<td>3.2</td>
<td>Add the distance and the index of the example to an ordered collection</td>
</tr>
<tr>
<td>4.</td>
<td>Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances</td>
</tr>
<tr>
<td>5.</td>
<td>Pick the first K entries from the sorted collection</td>
</tr>
<tr>
<td>6.</td>
<td>Get the labels of the selected K entries</td>
</tr>
<tr>
<td>7.</td>
<td>If regression, return the mean of the K labels</td>
</tr>
<tr>
<td>8.</td>
<td>If classification, return the mode of the K labels</td>
</tr>
</tbody>
</table>

4. Implementation

The implementation consists of many sub sections which are a sequence of steps to be followed to solve any Machine Learning problem. These are:

1. Collection of Data
2. Preparation of Data
3. Creation of Data
4. Training of Data
5. Testing of Data

1) **Collection of Data**

This is the first step of implementation in which the data is collected initially for the proposed system. The data set of movies is taken from the MovieLens which approximately consists around 100000 movies. It also contains feedback of previous users on the basis of which recommendation is being generated.

2) **Preparation of Data**

Second step is to make ready the data according to the requirement. This is the preprocessing step which represents the utility matrix which says which type of user rated which movie. User data and moviedata are separated into different data frames.

3) **Creation of Data**

The proposed algorithm is used in this step. The right number of clusters is choosed with the help of data sets. After that movies are divided into right number of clusters reference to CF.
4) **Training of Data**

   Here normalization of clusters are done after which the similarity of the user is measured with cosign similarity. Then the KNN prediction for the movie rating for top N users are executed.

5) **Testing of Data**

   Here prediction of movies and rating is done for the test user. This is need in proposed work to evaluate our model using some evaluation matrix.

   Thus, these are the proposed steps which is to be done for implementation of this proposed work. Afterwards, the system is ready to give recommendations to the user.

   The system is implemented in python programming language using KNN algorithm and the data set is taken from MovieLens and user data is taken from IMDB 5000 data sets.

   ![Fig. 2. Implementation model](image)

5. **Result of Proposed System**

   We tested our system by providing the first two alphabets of the movies we want to watch and we start getting suggestions of the top rated movies which is recommended by a very high number of peoples.

   Here we were searching for a movie name "AVENGERS" for which we typed the first two letters of the movie name and we start getting the recommendations.

   ![Fig. 3. Input given to recommender system](image)

   By providing the input we get some of the most nearest recommendations that the user may like.

   ![Fig. 4. Recommendations generated by the system](image)

Some more results of our proposed work.

**Input 1:**

![Fig. 5. Input given to recommender system](image)

**Output 1:**

![Fig. 6. Recommendations generated by the system](image)

Recommendation systems are being used in knowledge discovery techniques to the problem of making decisions on individual recommendation of data. Old CF techniques involve the amount of effort by the system increases with number of users. Hence, new RS's need to be developed to process high quality recommendations in a short span of time for wide scale network. Hybrid recommender system aims to be a very powerful alternative to search engines by not solely depend on keyword analysis, but by additionally using source analysis, author analysis, explicit ratings, implicit ratings, and citation analysis. Although some of the known methods have been known for past many years, they have not been applied in the reference of research paper and survey paper RS. Other approaches such as the 'intext distance similarity index' or CF annotations, classifications and links were developed. The combination of all approaches and methods is critical since each approach possesses disadvantages that can only be reduced by combining them. However, many problems remain unsolved, for instance regarding non-technical aspects like privacy concerns resulting from explicit and implicit ratings. Further, the analysis of various methods and techniques for calculating item to item similarities to identify relationship between those items and to produce an efficient recommendation for users to get the recommendations the as their wish. As a result, the hybrid approach is introduced which involves the use of multiple algorithm and methods at one time to give the exact similar and efficient result. Results showed that the method
outperformed than the other method. The same method is applied for the large real time rating dataset like MovieLens.

6. Conclusion

In this paper we have proposed a RS to provide recommendations to the user’s. The KNN algorithm along with CF is useful for the recommendations to the user. Also Cosine Similarity helped to make better the accuracy of the recommendations made by the proposed system.

We tested our system by providing the first two alphabets of the movies we want to watch and we start getting suggestions of the top rated movies which is recommended by a very high number of peoples. Thus the effort made by the user is very less and the user may not lose his interest in that particular site. Further there are many fields in which the proposed RS’s can be implemented like to recommend songs, books, venue, tourism, ecommerce sites, news etc.

References