A Study on Movie Recommendations using Collaborative Filtering

A Study on Movie Recommendations using Collaborative Filtering

Rahul Pradhan[1], Ashish Chandra Swami[2], Akash Saxena[3] and Vikram Rajpoot[4]

[1,4]Department of CEA, GLA University, Mathura
[2]S S Jain Subodh PG College, Jaipur
[3]Compucom Institute of Technology and Management, Jaipur

E-mail: [1]rahul.pradhan@gla.ac.in, [2]ashishchandraswami@gmail.com, [3]akash.saxena@gmail.com, [4]vikram.rajpoot@gla.ac.in

Abstract. The Recommendation system plays a major role nowadays, which is used for many applications. We know that the online content and service providers have a huge amount of content so the problem which arises is which data is required for whom so the problem of providing apposite content frequently. This paper represents the overview and approaches of techniques generated in a recommendation system. One of them is Collaborative Filtering which we will discuss about.

1. Introduction

The evolution of technology introduced the endless platform from basic to advance such as Machine Learning, Data Mining, and the Internet of Things, etc. As information technology is developing day by day, we can grab information from the internet effortlessly, but it becomes hard to use all of them appropriately. [1] To solve some of these problems big data was introduced. The big data is becoming a latest research interest in the cyber world. Big data has many characteristic which are volume, velocity, variety, value and veracity. [2] Big data is used to manage the data efficiency by pulling out some patterns and knowledge in an efficient way. But there are many users which have different taste so in this huge amount of data we have to give the user the useful content and helps them to decide the better one. To do these activities we have to rely on some automated machine so that it helps to reduce the effort of an individual. Here comes Recommendation System comes into a picture. Currently, Recommendation Systems are highly popular in our daily life like movies, news, books, shopping items, and music. Furthermore, the major companies like YouTube, Amazon, LinkedIn, Netflix use recommendation systems in social and e-commerce sites, etc. [3] It is used many technologies and algorithms to make the suggestion for an individual according to their requirement more precise and accurate. Recommendation Systems are classified in mainly three categories which are Content-based systems, the Collaborative Filtering system and the Hybrid recommendation system. Content-based systems works on the basis of the label or genre of an item. If a user watched a movie so it recommends similar movies based on director, a genre, and many more aspects. The main theory behind the collaborative filtering is that if users 'A' and 'B' have rated correspondingly in the past, then there will be an assumption that they will rate correspondingly in the future. Based

on the abstract of 'closeness', 'past', and 'rating', the basic algorithm can take several forms. [4] Hybrid is the combination of Content Based and Collaborative Filtering System. However, all these systems are not a fragment of precise, and research is going on to enhance the real-time performance of these systems. [5] We are using Movie Lens data set for our research.

2. Related work

Many researchers have introduced many researches in the area of constructing recommendation systems. [6] Recommendation systems are majorly used various applications to recommend items that a customer is probably to be concerned with customer preferences. Here we are only taking collaborative filtering only.

2.1. Collaborative filtering

Collaborative Filtering is a technique which collects information in the form of rating and check similarities from the other users. For example, consider A and B are two users are watching movies online on a specific platform , both of them watched the A1 movie and then they watched the B2 movie(on different time).Now the user B has watched movie A2, so this movie will also be to A who hadn't watch that movie yet. There are mainly two types of collaborative filtering techniques Memory Based approach and Model Based approach. In Memory Based approach, the recommendation is done on the basis of past activities, whereas in the model based approach, some prediction will take place on the basis of ratings, page viewing time, clicks and so on. The following diagram represents the working of the collaborative filtering. Figure 1 shows an example of collaborative filtering. There are two approaches which are used

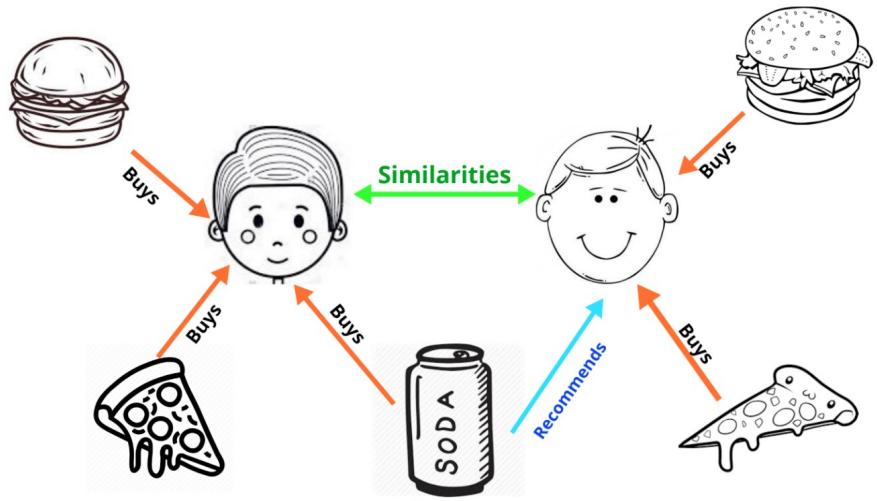

Figure 1. Collaborative Filtering Mechanism.

to solve using collaborative filtering user approach and item approach.

2.1.1. User based collaborative filtering User based collaborative filtering method is used to anticipate items to the user that are items of preference earlier for other users who are close to the user. [7] For example, let user 'A' and user 'B' have an indistinguishable preference manner. If a user A likes an item 1 user based collaborative filtering can recommend item 1 to user 2. It needs exact rating scores of items rated by users between users. Figure 2 shows a pictorial representation of the given example of a user based collaborative filtering.

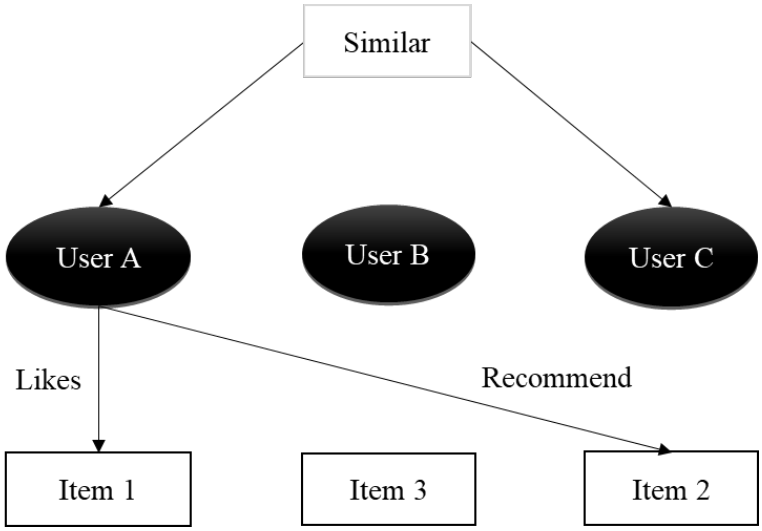

Figure 2. Explaination of user based collaborative filtering.

2.1.2. Item based collaborative filtering The item based collaborative filtering method is used to calculate items by identifying the closeness between an item and more items that are related with the user before. Let's consider item 1 and item 2 are similar. If a user likes item 1, item based collaborative filtering can suggest Item 2 to the User. It requires a kit of items that the user has rated before to anticipate the closeness between other items and a particular that item and then, it gives a suggestion in terms of that item by merging the user's earlier choices based on these item similarities [8]. In an item based collaborative filtering, a user's preferred data can be acquired in many ways like giving the rating to an item explicitly within a certain range or using various factors like click rate, users watch time, etc. Figure 3 shows a pictorial representation of the given example on item based collaborative filtering.

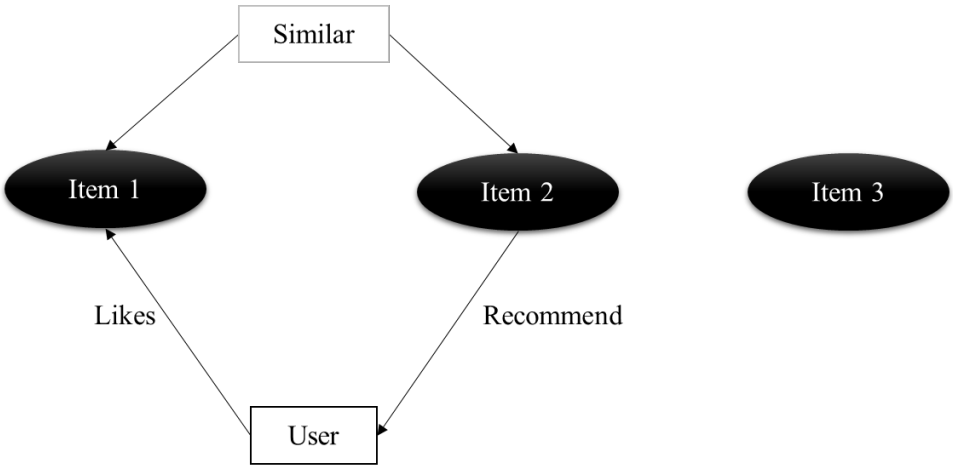

Figure 3. Explaination of item based collaborative filtering.

2.2. User similarity computation
The similarity between users is computed by estimating the value of the items estimated of the items estimated by two users. Each user uses N dimension vector to represent item score, for

example, to calculate of similarity of U1 and U3, first determine the set of films that they all resulted as M1,M2,M4,M5 and relative outcomes of these films. The outcome of vector U1 is 1, 3, 4, 2, and the outcome of vector u3 is 2, 4, 1, 5. The similarity of U1 and U3 is computed by the similarity formula [9]. Table 1 shows the table of above explanation.

Table 1. Calculating user's similarity.

U	M1	M2	M3	M4	M5
U1	1	3	3	4	2
U2	3	1	4		
U3	2	4		1	5
U4	2		2		

The similarity of u and u' is referred as $sim(u, u')$, the frequently used method of predicting user similarity are Cosine Similarity and Pearson Correlation similarity.

2.2.1. Cosine similarity This method computes the resemblance between two users by taking the cosine of the angle between the two vectors.

$$\begin{aligned} sim(x,y) &= cos(\overrightarrow{x}, \overrightarrow{y}) \\ &= \frac{\overrightarrow{x}.\overrightarrow{y}}{\|\overrightarrow{x}\|.\|\overrightarrow{y}\|} \\ &= \frac{\sum_{s\in s_{xy}} r_{x,s} r_{y,s}}{\sqrt{\sum_{s\in s_{x,y}} (r_{x,s})^2}.\sqrt{\sum_{s\in s_{x,y}} (r_{y,s})^2}} \end{aligned} \tag{1}$$

Among them,$r_{x,s}$ and $r_{y,s}$ are the outcome of goods s acquired by user x and y respectively, s_{xy} is the set of movies that user x and user y both acquired on. Mathematically $s_{x,y} = \{s \in items | r_{x,s} \neq \epsilon \cap r_{y,s} \neq \epsilon\}$.

2.2.2. Correlation -based similarity In this method, the resemblance between two items is computed by enumerating Pearson- r correlation. To make the correlation calculation precise we must first separate the co-rated cases. [10] Here $\overline{r_x}$ and $\overline{r_y}$, are the mean.

$$sim(x,y) = \frac{\sum_{s\in s_{xy}} (r_{x,s} - (\overline{r_x}))(r_{y,s} - (\overline{r_y}))}{\sqrt{\sum_{s\in s_{x,y}} (r_{x,s} - \overline{r_x})^2} \sqrt{\sum_{s\in s_{x,y}} (r_{y,s} - \overline{r_y})^2}} \tag{2}$$

2.2.3. Euclidean Distance Euclidean distance is the distance between two coordinates. The given equation here looks like:

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{3}$$

Here x and y coordinates can be users and items. However, this is not a very accurate way to find the similarity between two users or items.

2.3. Pro and cons

2.3.1. Pros

- It delivers a relevant content.
- It can bring traffic to your site.
- It reduces overload.
- It reduces workload and overhead.

2.3.2. Cons (challenges faced)

- Cold start problem / Handling Unknown Users
  - When the system doesn't have much information of a user, then the system face a problem what should recommend to that user.
- Data Sparsity
  - Same as above cons. If we don't have any information about the user then how do we suggest the movies to that user?
- Scalability
  - When the number of user's data and number of items data becomes huge then we have to make sure that these data will be displayed by taking minimum time.

3. Proposed Approach

Figure 4 will tell you about the basic process involved in the recommendation system. Recommendation system works on basically on two things product details and user details. We have to collect them from the system or from the database and make decisions on the basis of ratings if a similar items were found then it will generate recommendation system otherwise no recommendation system will be generated.

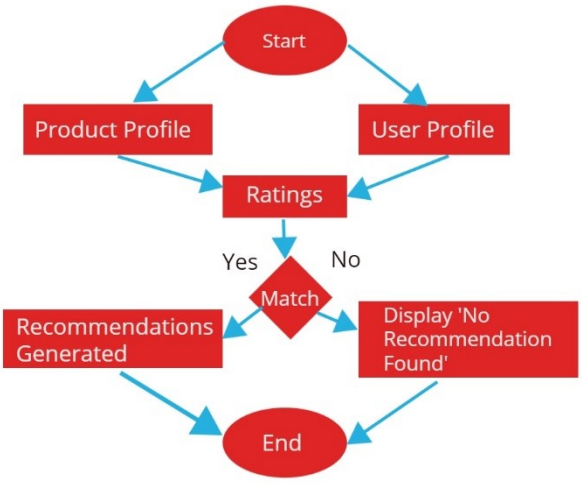

Figure 4. Flow chart for recommendation system employing collabrative filtering.

4. Experimental Setup

We are using Movie Lens Dataset [11] for our research. In this research we are applying item based collaborative filtering. The reason behind this is because user taste may change with respect to time but item doesn't change it remains same. There are certain stages to make our recommendation system efficiently to respond.

- Data Loading
  - To load the data and display accordingly we have to perform some operation like merging the two file in the dataset.
- Data Slicing
  - Here we are removing unnecessary column and data.
- Data Cleaning
  - In the real world data if we make a table of ratings in the recommendation system we find that most of the user are not rating the movies and are mostly inactive. The same cases are with movies either users don't watch or it's get too old. To make our computation more accurate we will remove such users and movies from our research. Now to predict similarity we have two methods either we can use correlation method or cosine method. In our research we had used correlation method.

Now to predict similarity we have two methods either we can use correlation method or cosine method. In our research we had used correlation method.

5. Results

Now let's suppose there are users who rated certain movies, for example, a certain user had rated the following movie as shown in Table 2:

Table 2. Movie Ratings.

Movie Name	**User Rating**
(500) Days of Summer (2009)	5
Alice in wonderland (2010)	3
Aliens (1986)	1
2001: A Space Odyssey (1968)	2

If we look carefully the user has rated romantic movies high and action movies low based this rating user will be recommended romantic movies. Here are the list of recommendation along with predictions (see Table 3). If the prediction goes to negative then that movie won't be recommended. Figure 5 shows the graph between movies and the prediction made by the recommendation system. As you can see here except few starting movies (which is rated good by user) all movies which are highly recommended from top to bottom (bottom to top on the graph 5).

Table 3. Predicted Movie Ratings

Movie Name	Predicted User Rating
(500) Days of Summer (2009)	2.584556
Alice in Wonderland (2010)	1.395229
Silver Linings Playbook (2012)	1.254800
Yes Man (2008)	1.116264
Adventure land (2009)	1.112235
Marley & Me (2008)	1.108381
About Time (2013)	1.102192
Crazy, Stupid, Love. (2011)	1.088757
50/50 (2011)	1.086517
Help, The (2011)	1.075963
Up in the Air (2009)	1.053037
Holiday, The (2006)	1.034470
Friends with Benefits (2011)	1.030875
Notebook, The (2004)	1.025880
Easy A (2010)	1.015771
Secret Life of Walter Mitty, The (2013)	0.997979
Perks of Being a Wallflower, The (2012)	0.967425
Toy Story 3 (2010)	0.963276
Ugly Truth, The (2009)	0.959079
Harry Potter and the Half-Blood Prince (2009)	0.954180

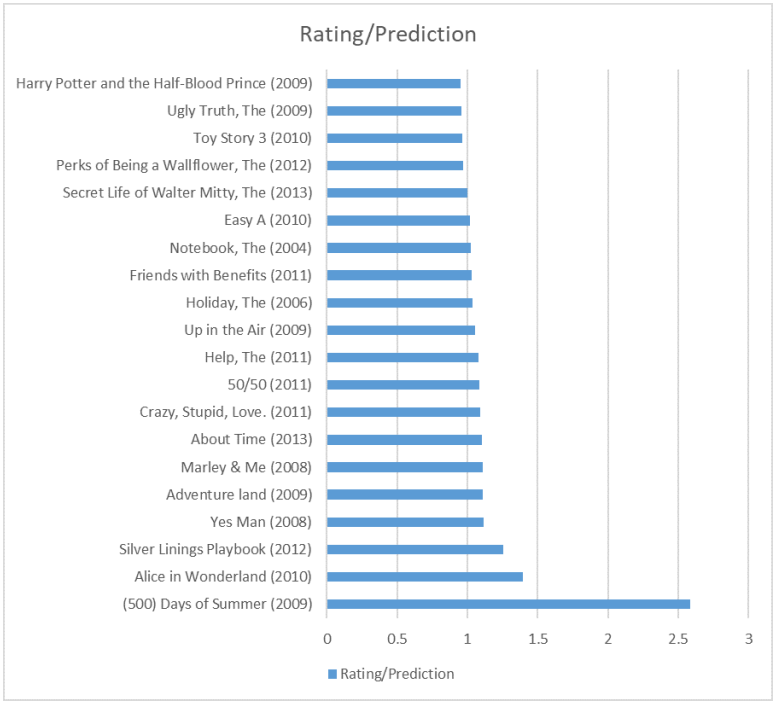

Figure 5. Graph between movies and the recommendation.

6. Conclusion

The Recommendation system is a very powerful technology which helps people to find what they like. These systems have certain limitations not recommending efficiently to the users. If we take Collaborative Filtering as an example, then we find it is most successful and powerful algorithm, [12] even though it is best but this algorithm has some high runtime and faces some issues like data Sparsity which can be removed by using a Hybrid movie recommendation system. But each algorithms have its strength and weakness. Our proposed approach can handle quite a big amount of data effectively. In future, we will work on its weakness and on its user interface.

References

[1] Ko S K, Choi S M, Eom H S, Cha J W, Cho H, Kim L and Han Y S 2011 *Symposium on Human Interface* (Springer) pp 558–566
[2] Aljunid M F and Manjaiah D 2019 *Data Management, Analytics and Innovation* (Springer) pp 283–295
[3] Wang Z, Yu X, Feng N and Wang Z 2014 *Journal of Visual Languages & Computing* **25** 667–675
[4] Goel D and Batra D 2009 *Department of Electrical and Computer Engineering, Carniege Mellon University* 1–7
[5] Raval N and Khedkar V 2019 *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH* **8** 2507–2512
[6] Wei J, He J, Chen K, Zhou Y and Tang Z 2016 *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)* (IEEE) pp 874–877
[7] Lee Y 2015
[8] Gong S 2010 *JSW* **5** 745–752
[9] Cui B B 2017 *ITM web of conferences* vol 12 (EDP Sciences) p 04008
[10] Sarwar B, Karypis G, Konstan J and Riedl J 2001 *Proceedings of the 10th international conference on World Wide Web* pp 285–295
[11] MovieLens 2018 Introduction-to-Machine-Learning https://github.com/codeheroku/Introduction-to-Machine-Learning/tree/master/CollaborativeFiltering/dataset (Accessed on 01/31/2021)
[12] Phorasim P and Yu L 2017 *International Journal of Advanced Computer Research* **7** 52