Research Result

Using Cosine Similarity to Build a Movie Recommendation

Aryan Sahu[1], Saurabh Sharma[2]

[1,2]*Baderia Global Institute of Engineering and Management, INDIA*

ABSTRACT

A recommendation system is a system which provides suggestions for various resources, such as books, films, songs, etc. Film recommendations generally forecast what films a user likes based on the attributes in previously loved films. These recommendation systems benefit organisations, which collect data from big quantities of clients and seek to deliver the best available ideas effectively. Many aspects can be taken into account in the construction of a film recommendation system, such as film genres, actors involved in it or even the film director. The programmes can propose films based on one or two or more qualities. The suggestion algorithm is based on the type of genres that the user could prefer to look at in this article. The technique is content-based filtering with genre correlation. The system's dataset is the Movie Lens dataset. The tool for data analysis is Python.

KEYWORDS

content-based filtering, cosine similarity, recommendations, movie

1. INTRODUCTION

The amount of data transactions occurring per minute has drastically expanded in this age of the Internet. The vast volume of data has expanded drastically with the number of Internet users. However, not all the data on the Internet is used or offers users with good results. Data in such enormous volumes can turn out to be inconsistent and are discarded without proper processing. In such circumstances, consumers have to search several times before they eventually get what they were looking for. Researchers have developed recommendation systems to overcome this challenge. A recommendation system provides consumers with relevant information, taking their previous choices into consideration. Data are filtered and specially adapted to the needs of the user. With more and more data available on the Internet, recommendation systems have grown popular since they can provide information efficiently in a short period of time. Recommender systems in numerous fields such as music, films, news and products in general have been developed. In today's day, a majority of organ networks utilise customer recommendation systems. There are a few to name, LinkedIn, Amazon and Netflix. LinkedIn recommends re-existing relationships between the millions of people that the user might know about on this platform. In this approach, the user does not have to actively search extensively for persons. Amazon recommendation systems work so that associated products can be purchased by customers. Amazon provides suggestions for any newcomers in previously favoured categories if a given user wants to buy books from the shopping web. Similarly, Netflix keeps account of the types of shows a client sees and delivers comparable re-comments. By the technique in which the systems of recommendations work, they may be grouped into three types — content-based, collaborative and hybrid. A content-based recommendation system looks at the past activity of the user and detects patterns to recommend comparable products. Collaborative filtering analyses and connects the experiences and ratings of the user with other users. Recommendations are made based on those that have the most similarities. Content-based and collaborative filtering have their own limits. To solve this, scientists devised a hybrid strategy that combines the benefits of both strategies. This paper proposes a content-based system of recommendations that uses gender correlation. The dataset utilised for this purpose is a Movie Lens dataset with 9126 films categorised by genres. There are 11 genres in all. The ratings were obtained from 671 users for these movements. Taking into account the films that are highly rated by users, films with similar genres are suggested.

2. BACKGROUND

Recommenders are widely categorised into three different types: collaborative systems, content-based filters and hybrid systems [3]. Collaborative systems employ inputs from different users and perform various input comparisons [3]. They construct models based on consumers' historical behaviour [1]. For example, the film recommendation algorithms use user ratings for various movies [2], and try to discover similar users, and propose movies that have been well-rated [3]. There are two approaches to collaborative filtering systems—memory-based approaches and model-based approaches [3]. Continuously analysis memory-based techniques user data

to produce recommendations [3]. By using user ratings, accuracy is gradually improved over time [3]. They are domain-independent and do not require a study of content [3]. Model based approaches construct a user behaviour model and then forecast future behaviour with parameters [3]. The adoption of techniques based on partitioning also leads to improved scalability and accuracy [3].

Content-based filtering systems analyse documents or preferences from a parallel user and try to model this data [3]. They employ the special interests of the user and try to match the user's profile with the attributes of the various content objects to be recommended [3]. The additional disadvantage is that sufficient data is required to create a reliable classifier [1]. Content based filtering systems are classified into three methods—wrapping, filtering and embedding [3]. Wrapper approaches partition the features into subgroups, do analysis on these subsets and then evaluate the most promising of these subsets [3]. Filter methods are used to evaluate the features on their content using heuristic methods [3]. Both are independent of the algorithms utilised. Conversely, embedded techniques are linked to the algorithm used—functional selection takes place during the training phase [3]. In order to optimise recommendation systems, and to decrease limitations in each of the two ways, hybrid systems integrate collaborative and content-based filtering algorithms [3]. It is therefore attempting to extend the advantages of one method to compensate for the other's disadvantages [3]. Three hybrid systems are available — weighted hybrid, blended hybrid, and cross-source hybrid [3]. In weighted hybrid systems, the value for each object is preserved to identify the weighted amount in relation to the different context sources [3]. These weights are different based on the preferences of a user [3]. Each source is utilised to rank the numerous things in mixed hybrid techniques, and the few top items in each rank list are selected [3]. Cross-source hybrid approaches advise items in several contexts [3]. The basis of these approaches is that the more sources an item is contained, the more essential the item [3]. By filtering emotions, Wakil et al. have tried to improve their recommendation system [4]. When a user sees a given sort of film, certain emotions are activated [4]. Likewise, a user's emotions can cause a certain style of film to be watched [4]. They acknowledged that conventional user profiles do not take the user's emotional status into consideration and developed an algorithm using emotionality [4]. It analyses a sequence of colours chosen by the user in order to ascertain the user's current emotional state [4]. Debnath et al. devised a hybrid weighting recommendation system [5]. They estimated the relevance of different characteristics for each user and so allocated weights to these features [5]. They found the weighted amount to estimate which items the user would be interested in [5].

3. RECOMMENDATION SYSTEM USING CONTENT-BASED FILTERING

The approach employed for the construction of the recommendation system is content-based filtering. As previously said, content-based filtering analyses prior behaviour of users and recommends similar products on the basis of the considered parameters. This is designed to recommend films for users based on genre similarity. If a user has rated a given movie high, the algorithm recommends more films with similar genres. The dataset utilised is separated into two pieces.

One section includes the list of films and the categories under which they were grouped. The other half of the data set includes the list of films which the user has rated on a scale of 1–5, with the highest 5. The first thing to do is to build a unified dataset of films, genres and their ratings for associating genres with ratings. The ratings were converted to binary values for simplicity. If a certain user's rating exceeds 3, he will obtain a value of 1, otherwise it will be given a value of −1. The genres are likewise binarily divided, with a consistent manner. Of the total of 11 genres, if a film has a certain genre, it obtains a value of 1. If the genre is not in the film, it gets a value of 0. The matrix for the user profile offers a combined effect of genres and ratings by calculating the point product of the genre and the rating matrix. Again, a binary representation is selected for the sake of uniformity. If the dot value is negative, it is assigned 0. 1 is assigned for a positive value. After the dot product matrix has been obtained for all the films, a similitude measure is created by calculating the lowest distance between the user and the others. The values that deviate least from the current user's preferences are those recommended by the system. The following is the algorithm used to develop the recommendation system:

Algorithm

Step 1. Build a dataset data frame with film identification as rows and genres as pipeline columns.

Step 2. List all the genres available in the dataset.

Step 3. Iterate over the data framework of the genre previously produced. If a genre is in a film, the 1 value is allocated to the matrix genre.

Step 4. Read the rating sheet and build a rating matrix that assigns 1 to films with rating more than 3 and −1 to films with ratings less than or equal to 3.

Step 5. Calculate the dot of the two matrices — the matrix genre and the matrix rating. This is the matrix resulting

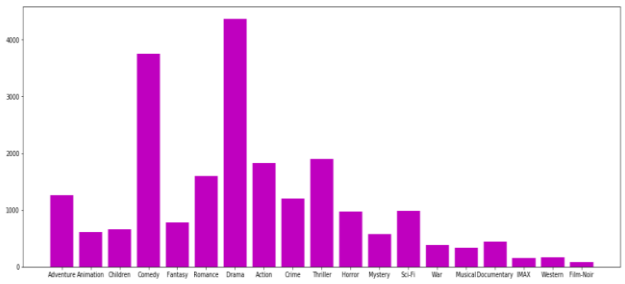
Step 6. Convert the matrix result to a binary format. Assign 0 to a negative point product value, else assign a value of 1.

Step 7. Calculate the distance Euclidian to the existing user from other users.

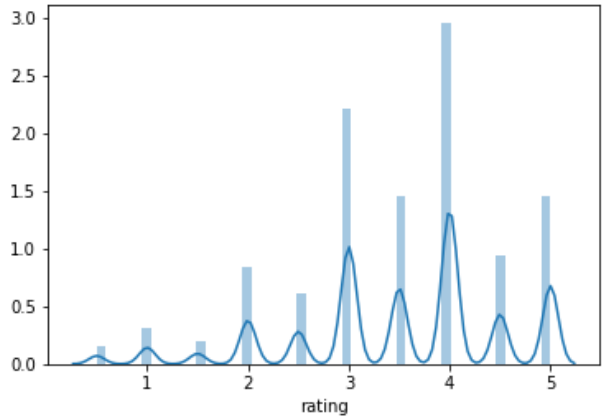
Step 8. Retain the minimum distance rows. These are the recommended films for the existing user.

4. RESULTS OF SIMULATION

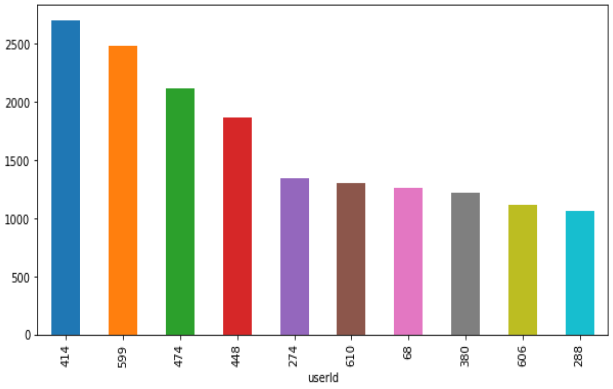
The matrix of the genre is built with films and genres separated by columns. The collection contains a total of 11 genres (Fig. 1). The rating matrix is transformed to a binary format for each user matching the film ID. Each user rated one or more films (Fig. 2). The result matrix is calculated using the genres matrix and rating matrix and is the point product of the previous two matrices. In Fig. 3, the result is processed further in a binary format. If the value of the dot product is higher than 0, 1 is otherwise assigned to this cell as 0.



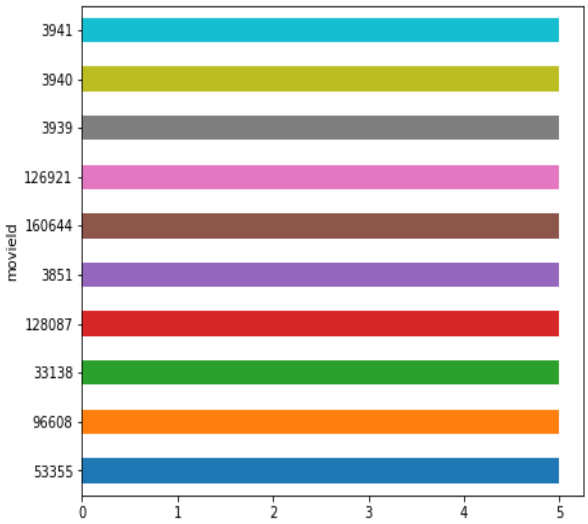
Most popular genres of movie released



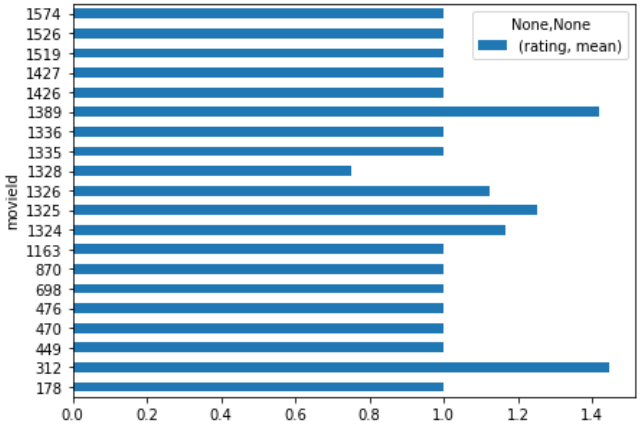
Distribution of users rating



Top 10 users who have rated most of the movies



Movies with low average rating



Model evaluation with KNN

Hit:93

Fault: 6.7

5. CONCLUSION AND FUTURE WORK

The recommendation system implemented in this work seeks to make film recommendations based on film genres. If a user scores a film of a particular genre highly, films of similar genres will be recommended. Recommendation systems are frequently used to search for accurate and relevant content in today's Web 2.0 era. Although basic recommendation systems encourage users based on a few characteristics, sophisticated parameters take several parameters into account. Intelligent consumer suggestions can be made by using machine learning in recommending systems. Given the potential of these technologies, their commercial worth is enormous. Several MNCs have used the advice system's potential to get clients to use their goods. This also has a major impact on data mining and web mining.

Mobile cloud computing (mcc) is capable of saving energy, improving user experience and application. All of the following frameworks have their own advantages and problems but are not up to the level to address all security, energy and user experience difficulties. Security issues are significant issues in mcc, they have to be more concentrated than other problems.

REFERENCES

[1] Ghuli, P., Ghosh, A., Shettar, R.: A collaborative filtering recommendation engine in a distributed environment. In: 2014 International Conference on Contemporary Computing and Informatics (IC3I). IEEE (2014)

[2] Zhao, L., et al.: Matrix factorization + for movie recommendation. In: IJCAI (2016)

[3] Bhatt, B.: A review paper on machine learning based recommendation system. Int. J. Eng. Dev. Res. (2014)

[4] Wakil, K., et al.: Improving web movie recommender system based on emotions. (IJACSA) Int. J. Adv. Comput. Sci. Appl. 6(2) (2015)

[5] Debnath, S., Ganguly, N., Mitra, P.: Feature weighting in content-based recommendation system using social network analysis. In: Proceedings of the 17th International Conference on World Wide Web. ACM (2008)