International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 11, pp: 299-303 www.ijaem.net ISSN: 2395-5252

Movie Recommendation by Using Jaccard Distance

Sahil Narang[1], Harshita Malhotra[2], Renu Chaudhary[3], Vanshika Gandhi[4], Shourya Pokhriyal[5]

B.TECH IT[1] HMRITM
B.TECH IT[2] HMRITM
PROFESSOR IT DEPT[3] HMRITM
B.TECH IT[4] HMRITM *B.TECH IT[5] HMRITM*

Date of Submission: 05-12-2020 Date of Acceptance: 20-12-2020

ABSTRACT: Movie recommendation using Jaccard distance is a relatively new approach in the field of Python. As one of its first implementations, Jaccard distance has gained a lot of attention. Together with Utility Matrix it makes movie recommendation based on similar taste easy and reliable. Using this, the recommendation is made based on the behaviour of multiple users and not on the properties of the content that is being consumed.

KEYWORDS: Recommendation, Jaccard distance, Python.

I. INTRODUCTION

Recommendation system is widely being used today. From Netflix to Amazon, all popular websites use recommendation system to recommend movies and products by keeping track of the user activities and storing the data to make similar recommendation by identifying the common pattern and taste of each and every user. The need and the demand for recommendation system is increasing day by day because there are abundant resources on the internet today and it becomes tedious for the users to search from a large collection and then select the item or the movie that is similar to their taste, so that is where the recommendation system comes into play by extending a helping hand to make our life easy by filtering out the content based on our preference. It provides users with personalized content and services and is beneficial to both service providers and the users as it reduces the cost of finding and selecting items in an online environment. Therefore, we need to use accurate and efficient recommendation techniques within a system that will provide dependable and relevant recommendation for users. The best approach to follow for the recommendation system is to consider the preferences of others whose taste is similar to that of yours and recommend movies that they have watched which can be done by identifying the behaviour of multiple users and using collaborative filtering with Jaccard distance to filter out the content. An important part of collaborative filtering is to use Utility Matrix in order to recognize users who have similar preferences and there after using Jaccard distance as the metrics to compare rating provided by viewers and figure out if they have similar taste. This way viewers with similar taste can come closer.

II. FUNDAMENTAL OF PYTHON

Conceived in the late 1980s, Python is general purpose ,dynamically typed and world‘s fastest growing programming language. One of the core principles of python is that it provides syntax which is easy to use and execute which makes it one of the most popular programming language across the globe. It is so versatile that it allows web development,game development ,software development and even provides great functionality to deal with mathematics, statistics and many scientific functions. Therefore,it is used by not just software engineers but even mathematicians ,data analysts, scientists ,accountants ,network engineers and even beginners for a variety of different tasks such as Data Analysis and Visualization,Artificial Intelligence and Machine Learning , Automation, Web/Mobile/Desktop Apps development,Software testing or even hacking . It is an interpreted, open source,high level language and supports procedural ,functional and object oriented programming.

Data is an important aspect when it comes to programming.Python provides various ways of storing the data in the right way and managing it, so it becomes easy to access and perform operations on.This is done with the help of Data Structures.

Python has implicit or built in support for data structures that anybody can use and also has

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 11, pp: 299-303 www.ijaem.net ISSN: 2395-5252

powerful functions.The built in data structure is comprised of lists ,dictionaries, tuples and sets.On the other side,python also supports data structures that are written by the user. The most common user defined data structures are stack ,queue, graphs,trees,linked list and hashmap. Python supports arrays(using numpy) which can be used to build many data structures.

Python Data Structures

List-Lists is used to store data in sequential manner.It can store heterogenous data types .List is mutable in nature,which means, it can have different kind of data and one can even change the data whenever they want to.

Tuples- Similar to lists,tuples also store data in sequential manner but are immutable in nature.Thus their performance are faster as compared to lists.

Dictionary- Dictionary is data structure that holds key value pairs,where key is an attribute which has some specific values assigned to it. Dictionaries are mutable in nature.

Sets- Set is a data structure which is a collection of unordered elements. All elements in a set are unique,which means they are present just once.These values can be of any data type but are not indexed.And because of this we cannot perform indexing operations on sets such as slicing. Sets are used where the order of data doesn‘t matter but unique data elemnts are needed

Python sets are highly useful to perform mathematical operations such as union,intersection etc.

III. FUNDAMENTAL MATHEMATICAL MODELS OF COLLABORATIVE FILTERING

Mathematical formulas are presented in this section, which are performed in the proposed model. Primarily, neighborhood-based prediction technique has been described and then various traditional similarity measures are presented and shown their limitations in recommendations generation

Neighborhood-based prediction technique

Let us consider $U = \{u_1, u_2, u_3, \ldots u_m\}$ as a user set, $I = \{i_1, i_2, i_3, \ldots i_n\}$ as an item set and $R = \{R(1,1), R(1,2), R(m,n)\}$ as a set of user-item rating pairs where, $R_{(u,i)}$ means rating of the user u for item i . Generally $R_{(u,i)} \in W = \{w_1, w_2, w_3, \ldots w_N\}$ set of discrete rating scores where a higher rating indicates the user strongly likes that item. Similarity indices between users or items are calculated from the user-item rating matrix to group similar users and items for appropriate recommendations. The fundamental role of a user-user similarity index is to classify users who have given similar ratings to a set of items in the past. Likewise, an itemitem similarity index is calculated by taking into account the set of items that are co-rated by a set of users. The prediction is then generated for all unrated items by forming similarity indices of targeted users with k-nearest neighbors. The unknown rating is predicted based on user-user similarity as

$$r_{ij} = \bar{r}_i + \frac{\sum_k Similaries(u_i, u_k)(r_{kj} - \bar{r}_k)}{number\ of\ ratings}$$

Where, $R^*_{(u,i)}$ is the predicted rating of the targeted user u for item i and $R_{(u)}$ is the mean rating of user u . $S_{(u)}$ is the number of top similar users who have also rated item i . $R_{(v,i)}$ is rating of the nearest neighbor v for item i . $R_{(v)}$ is the mean rating of the nearest neighbor v . $Sim_{(u,v)}$, is the similarity index between the user u and its nearest neighbor v . The absolute sign for similarity value is used in the denominator to avoid the negative correlation between the targeted user and nearest neighbors. The motivation behind $(R_{(v,i)} - R_{(v)})$ is that if the neighbor v has rated the item above average, then it will add to the average rating of the user u Similarly, the logic behind multiplying $Sim_{(u,v)}$ with $(R_{(v,i)} - R_{(v)})$ is that if the similarity between the user and its neighbor v is very high, then the user‘s rating prediction will be heavily influenced by the neighbor v and vice versa.

The traditional similarity measures

Similarity measure in the recommender system is the statistical measure of how two users and items are related to each other. There are several traditional similarity metrics such as Cosine (COS), Pearson’s Correlation (COR), Constrained Pearson’s Correlation (CPC), Mean Squared Difference (MSD), Jaccard, JMSD etc. [2].

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 11, pp: 299-303 www.ijaem.net ISSN: 2395-5252

Cosine Similarity (COS):

$$Cosine\ Similarity: Sim(u_i, u_k) = \frac{r_i \cdot r_k}{|r_i||r_k|} = \frac{\sum_{j=1}^{m} r_{ij} r_{kj}}{\sqrt{\sum_{j=1}^{m} r_{ij}^2 \sum_{j=1}^{m} r_{kj}^2}}$$

Cosine similarity measures the angle between two rated vectors where the smaller angle indicates greater similarity and higher angle show lesser similarity [3].

Where $R_{(u,i)}$ is the rating of the item i given by user u and $I_{(u, v)}$ is the number of co-rated items of users u and v . The range of cosine similarity is 0 to 1, where higher value signifies the closest similarity between users u and v .

Mean Square Distance (MSD):

Similarly, Mean Square Distance (MSD) between two users is calculated by the ratio of sum square of the difference of ratings on co-rated items and the cardinality of co-rated items. The Mean square Similarity is then calculated by subtracting MSD from 1.

IV. LIMITATIONS OF TRADITIONAL SIMILARITYMEASURES

Though different similarity measures are adopted by different online companies for their recommendation system, still a lot of loopholes have been noticed in various situations, which leads inaccurate prediction.

1. Let $U_1 = (2,0,3,0)$ and $U_2 = (5, 2, 0, 2)$ are the rating vectors of two users where only one co-rated item presents. It is noticed that the Pearson's correlation coefficient cannot be determined as because denominator becomes zero. Likewise, cosine similarity yields 100% similarity regardless of an actual matching.
2. Further, let $U_1 = (2,1,3, 2)$ and $U_2 = (1, 2, 2,3)$ be the rating array of two users. Even though both the users are highly similar, the Pearson's correlation coefficient generates zero similarity indexes.
3. In another situation, let $U_1 = (2, 2, 0,1)$ and $U_2 = (4, 4, 0, 2)$ be ratings of two users. In this case, the cosine similarity index computes a similarity value of 1 which is an insignificant similarity valuation. Cosine similarity always yields a very high similarity (i.e. 1) when ratings are multiple of each other because in that scenario geometrically they overlap each other on the same straight line

V. PROPOSED SIMILARITY MODEL

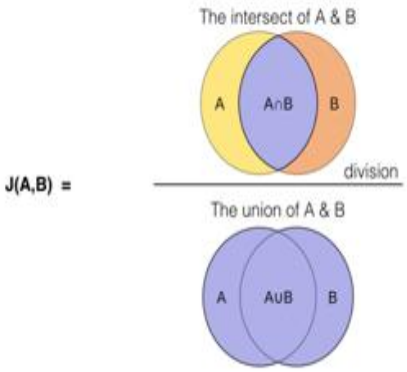
Jaccard Distance

The Jaccard Index, also known as intersection over union is used for measuring the similarity between finite sample sets. It is defined as the size of intersection divided by size of union of the union of sample sets[1] :

$$J(A, B) = \frac{|A \wedge B|}{|A \cup B|} = \frac{|A \wedge B|}{|A| + |B| - |A \wedge B|}$$

(If A and B are both empty , define J(A,B) = 1)

$$0 \leq J(A,B) \leq 1$$



According to venn diagrams,The Jaccard distance, measures the dissimilarity between the sample sets, is complementary to the jaccard index means we can find the jaccard distance by subtracting the jaccard index from 1, or we can divide the difference of size of the union and intersection of the finite sets by the size of the union[1].

$$dj(A,B) = 1 - J(A, B) = \frac{|A \cup B| - |A \wedge B|}{|A \cup B|}$$

.Examples:

There are 2 finite sets A and B,

A = {0,2,5,6}

B = {0,2,3,4,6,7,8,9}

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 11, pp: 299-303 www.ijaem.net ISSN: 2395-5252

J (A, B) = | A ^ B | / | A U B | = 3 / 9 = 0.33 (Jaccard Index)
It means set A and set B are 33.33% similar.

dj (A, B) = 1 - 0.33 = 0.67 (Jaccard distance)

Illustration with examples to formulate new similarity model

If there is a large dataset we can convert it in dataframe by using the python's Pandas library and by converting it in the dataframe we can use any column or any row as a set and hence perform the multiple operations like union, intersection , can find the similarity between the 2 columns, dissimilarity between 2 columns and by using the jaccard formula we can predict or recommend anything like if there is maximum similarity between 2 columns we can use any one column to reduce the redundancy in the dataset and also we can use this for recommendation systems.

In shopping websites we can use jaccard distance, if a customer buys something like shoes and other customer buys something like slippers then we can recommend both customers more shoes and slippers because their similarity score will be maximum because they fall under same category 'footwear'.

In movie websites like we can use jaccard distance, if a person watch a movie of specific genere and of greater than 3.5 rating movies and other person watches the movie of same genere like the other person but less than 3.5 ratings we can recommend them both the common movies on basis of similarity score, it would be easy to recommend and to watch the movies of your types.There are n number of examples in which we can use jaccard distance formula to find similarities and dissimilarities for better , easy and fast thinking and selecting the things.

VI. EXPERIMENTS AND RESULTS

Movie recommendation

A recommendation system is an automated system to filter some entities. These entities can be product, movies, songs or anything. Recommendation systems are used from all over on a daily basis from Netflix to Amazon Prime to YouTube. For example we watch a movie and then later we get a recommendation for a different movie based on our previous viewing. Many companies use recommendation system because they want to understand their customers well and show them information that is relevant to them while also sharing new items that they could be interested in. There are basically two main techniques used for recommendation system namely, content-based filtering and collaborative filtering. Content based filtering uses item features to recommend new items based on what the user has liked in past. Based on that data, user profile is generated which is then used to make suggestions to user. Collaborative filtering makes choices that user makes when buying, watching or enjoying. It then makes connection with other users of similar interest to produce a prediction. One popular example of collaborative filtering is Netflix. Everything based on their site is driven by their customer's selection which if made frequently enough get turned into recommendations. Netflix orders these recommendation in such a way that the highest ranking items are shown on the top [4].

There is a new way of recommending the movies that we have discovered, that is the Jaccard distance. Jaccard distance can be used with large datasets and give accurate recommendations. Jaccard distance is a mathematical formula that is used to find similarities between two sets of data. Jaccard distance is calculated by dividing the intersection of two sets (A and B) by union of the two sets (A and B). Let us understand this using an example. Let us assume that we have four sets A,B,C,D. Now suppose we have to recommend something to set A. We will apply the Jaccard distance formula between set A and the remaining sets, that is between set A and set B, set A and set C, set A and set D. Now we have three results/scores. The result/score that comes out to be the highest will be chosen. For example if the highest score is between set A and set B, then set A will get recommendations based on what set B likes or has watched.

Jaccard distance formula is efficient, reliable and will give fast results. It will provide us the best results as it is based on a pure mathematical formula. Jaccard distance formula can used for recommending others things also like songs, products and videos. With this method we will be able to provide the user what he/she actually likes [5].

VII. CONCLUSION

Recommendation system is vital for a smooth shopping as well as a great movie experience. They are at the heart of the internet economy which keeps the user hooked to social media as well as online shopping and entertainment platforms. It follows an approach to quantify similarities between users and recommends products or movies based on the common taste. It observes the viewer's pattern from every action starting from the searches to the wishlist along with the past experiences and stores it to make accurate

recommendation that the user might like and would certainly end up buying or watching. Calculating and employing Jaccard distance makes our recommendation system reliable and easy to use by quantifying the similarities at one go with fast processing. Jaccard distance takes into account the number of products rated by both the users that are being compared, but not the actual value of the rating itself. This approach can be used to predict and to see what others especially those who have a similar taste to us have bought or consumed.

SOME OF THE ADVANAGES FROM THE ABOVE RESULTS

a) It makes the recommendation system more reliable.
b) Efficiency increases
c) Execution is faster
d) Gives accurate results

REFERENCES

[1]. https://en.wikipedia.org/wiki/Jaccard_index
[2]. B.K. Patra, R. Launonen, V. Ollikainen, S. Nandi, A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data, Knowledge-Based Syst. 82 (2015) 163–177. doi:https://doi.org/10.1016/j.knosys.2015.03.001
[3]. H.J. Ahn, A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem, Inf. Sci. (Ny). 178 (2008) 37–51. doi:https://doi.org/10.1016/ j.ins.2007.07.024
[4]. https://www.sciencedirect.com/
[5]. https://neo4j.com/