

Product Recommendation System

Vismay Patel
Ahmedabad University
AU1841071

Shivam lakhtariya
Ahmedabad University
AU1841084

Nisarg Patel
Ahmedabad University
AU1841048

Priyanshi Shah
Ahmedabad University
AU1841009

Abstract— Recommendation systems enable users to access products that they may not be aware of. The two traditional recommendation techniques are content-based and collaborative filtering. While both methods have their own advantages, they also have certain disadvantages, some of which can be solved by combining both techniques to improve the quality of the recommendation. Broadly speaking, a recommendation system provides specific suggestions about items (products or actions) within a given domain, which may interest the given active user [1].

Different types of similarities are used in recommending the products to the user. In this paper we are going to discuss mainly about two types of similarities i.e cosine similarity and jaccard similarity and also about how k-means clustering can be used in recommending the products based on product ratings and search keywords.

Keywords- k-mean clustering, Correlation, Single value decomposition, cosine similarity , jaccard similarity, Term Frequency and Inverse Document Frequency (tf-idf).

I. INTRODUCTION

Product recommendation is generally a filtering system which seeks to predict, display and suggest the product to users that they would like to purchase. This type of system is utilized in a variety of fields such as news, research articles and many more.

Filtering attempts to discover user preferences, and to learn about them in order to anticipate their needs. Broadly speaking, a recommendation system provides specific suggestions about items (products or actions) within a given domain, which may interest the given active user.

Many different approaches to the recommendation system problems have been published [2–4], using methods from machine learning and approximation theory. Independent of the process used and based on how the recommendations are made, recommendation systems are usually classified [3] into the following categories: Collaborative filtering that try to identify groups of people with similar interest to that of the user and recommend items that they liked and Content-based recommendation systems which use content information to recommend items similar to those previously preferred by the user.

In filtering we use different similarity measures to recommend the products. There are many types of similarity measures such as manhattan distance, euclidean distance, pearson correlation, cosine similarity, jaccard similarity etc.

Cosine similarity measures the similarity between two vectors by calculating the cosine of the angle between them. It is useful when we don't want magnitude to skew the results. Most commonly used similarity measures are cosine similarity and pearson correlation.

Jaccard similarity is a statistical measure of similarity which is used to find the similarity and dissimilarity between two sets.

Clustering is an unsupervised learning technique. It can be used to obtain interesting patterns in the data. It is used to create features based on the input attribute. The main methods used for achieving clustering is K MEANS and hierarchical clustering. Clustering is used in market segmentation where we try to find the items that are similar to each other. K-means is used to partition the dataset in k pre-defined clusters where each data point belongs to only one group.

K-means clustering is commonly used because it can handle large datasets, adapts easily and is relatively simple to implement.

II. LITERATURE SURVEY

ACCORDING TO Y. S. THAKARE ET AL. [4], THE PERFORMANCE OF K-MEANS ALGORITHM WHICH IS EVALUATED WITH VARIOUS DATABASES SUCH AS IRIS, WINE, VOWEL, IONOSPHERE AND CRUDE OIL DATA SET AND VARIOUS DISTANCE METRICS. IT IS CONCLUDED THAT PERFORMANCE OF K-MEANS CLUSTERING IS DEPEND ON THE DATABASE USED AS WELL AS DISTANCE METRICS.

Recommender systems in ecommerce[5.] Electronic Commerce, cross-sell, up-sell, mass customization - The ideas of new applications in the field of recommendation systems in e-commerce sites.

Soumi Ghosh et al. [6] proposed a comparative discussion of two clustering algorithms namely centroid based K-Means and representative object based Fuzzy C-Means clustering algorithms. This discussion is on the basis of performance evaluation of the efficiency of clustering output by applying these algorithms. The result of this comparative study is that FCM produces closer results to the K-means but still computation time is more than k-means due to involvement of the fuzzy measure calculations.

Shi Na et al. [7] Proposed the analysis of shortcomings of the standard k-means algorithm. As k-means algorithm has to calculate the distance between each data object and all cluster centers in each iteration. This repetitive process affects the efficiency of clustering algorithms.

III. IMPLEMENTATION

A. We used python to implement user-built k-means clustering.

B. We had implemented a recommendation process based on features which is product Id and ratings. In this, based on the input of the product Id, it fetches the rating value and based on that rating value it displays the similar product Id having the same rating as that product. The step by step approach is as followed:-

At first we imported the relevant libraries such as sklearn, minmaxscaler and many more.

Further important features which are product Id and ratings were fetched from the dataset on which we had performed clustering in order to recommend users a best product.

We had then plotted the graph of ratings vs product ID just to have a clear idea what kind of ratings a particular product had.

The K means clustering was then applied to both of the features in order to assign a cluster accordingly. Here we have taken the number of clusters as 5 because the ratings values range from 1 to 5 and at maximum we can only have five clusters to put them in.

After that as a result we have built a function named show recommendations which suggests users similar products based on their input parameter.

B. First we had to convert the text file of the data to a readable format, so we converted that text file to excel file using python script and then read the file using the inbuilt function and printed the output in an appropriate format. We had to recommend the product based on the ratings given by other users so we printed a matrix between the users and the products purchased by the users. The spaces in the matrix were filled by the values of the ratings. The data may have null values so we replaced the null values to zeros in order to make it computable. Then we transposed the matrix as we wanted to perform item-item collaborative filtering. Then we found the cosine similarities between the products and printed out the similarity matrix. Then based on the product purchased by the user and the rating given by the user to that particular product we recommend the products based on the similarity index of that product with all the other products.

D. We had also implemented the clustering algorithm which recommends similar kinds of products to the users based on their search. Initially our dataset was not having a product description feature and since it was necessary to have it to perform clustering we modified our dataset and added a product description feature.

We implemented k-means clustering without the use of inbuilt functions. The problem with clustering is that they do not specify the ground truth value so we have to figure it out ourselves. So we first start by specifying the number of centroids using the elbow method, number of samples and the number of iterations required. After that we scattered plot our samples for better understanding. We then initialise the centroids by randomly picking the points from our samples.

After that we made a function to calculate the euclidean distance between the samples and the centroids. By the use of this function we were able to assign the samples to each of the clusters. The sample whose distance is closest to one of the centroids of the clusters is assigned to that particular cluster. After the clusters are made we recalculate the centroids again for more accuracy within a cluster.


This is done by using the formula given below:

$$x_c = \frac{\sum_i^n A_i y_{c,i}}{\sum_i^n A_i}$$


Where A_i is the mass of the point and y_i is the coordinate of that sample.

After recalculating the centroids we compare the distances of each sample with both the new centroid as well the old centroid. The centroid which minimises the sum of distances is kept and is updated as the new centroid.

B. PRODUCT RECOMMENDATION BASED ON PRODUCT ID AND RATINGS



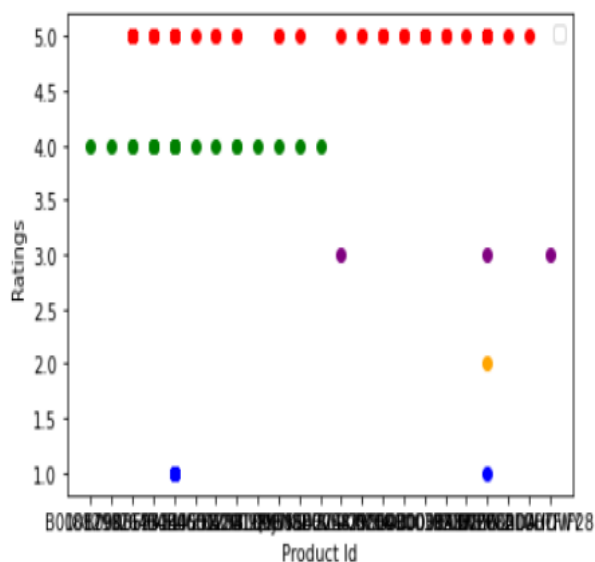
```
show_recommendations("B000179R3I")
```



	Product_Id	Ratings
0	B000179R3I	4
3	1882931173	4
12	B000058A81	4
20	B000058A81	4
63	826414346	4
64	826414346	4
65	826414346	4
69	826414346	4
78	595344550	4
89	595344550	4
95	595344550	4
97	595344550	4
105	595344550	4
116	B000IZ8AZO	4
120	B0001Z3TLQ	4

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

A. CLUSTERING ON PRODUCTID AND RATINGS



C. JACCARD SIMILARITY

	productID	7106823	7128355	20794207	26204207	60539453	60539461	60550546	60958596
	productID								
	7106823	1.000000	0.999699	0.999850	0.999850	0.999549	0.999774	0.999774	0.999699
	7128355	0.999699	1.000000	0.999699	0.999699	0.999398	0.999624	0.999624	0.999624
	20794207	0.999850	0.999699	1.000000	0.999850	0.999549	0.999774	0.999774	0.999699
	26204207	0.999850	0.999699	0.999850	1.000000	0.999549	0.999774	0.999774	0.999699
	60539453	0.999549	0.999398	0.999549	0.999549	1.000000	0.999474	0.999474	0.999398
	
B0084BM6UO	0.991202	0.991051	0.991202	0.991202	0.990901	0.991126	0.991126	0.991051	
B0087LZ3WO	0.999474	0.999323	0.999474	0.999474	0.999173	0.999398	0.999398	0.999323	
B008U11166	0.999474	0.999323	0.999474	0.999474	0.999173	0.999398	0.999398	0.999323	
B008ZN8PTG	0.998872	0.998722	0.998872	0.998872	0.998571	0.998797	0.998797	0.998722	
B009BOSTTO	0.999474	0.999323	0.999474	0.999474	0.999173	0.999398	0.999398	0.999323	

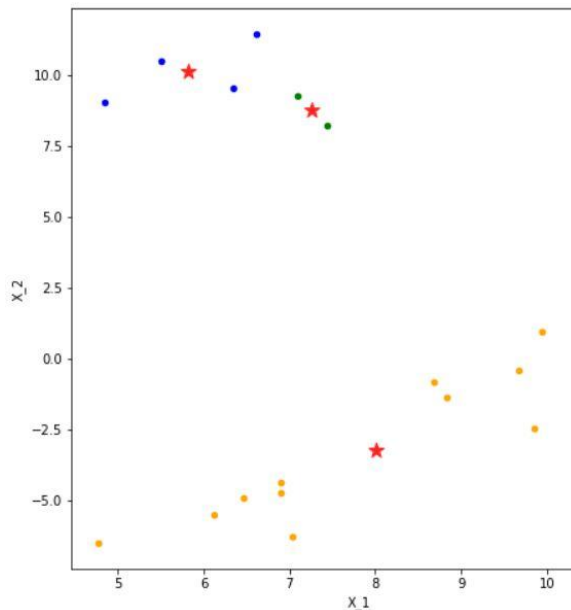
productId	
B000PY2MX4	1.500000
B00009LW53	1.493232
B0000A7XQR	1.493232
B0000C3XXU	1.493232
B0000C3XXS	1.493232
	...
B000CQ55AC	1.454655
B0001YXWVO	1.452173
B000QWA2KU	1.437961
B000N6DDJQ	1.423184
B0050T2YVA	1.399496

D. USER-BUILT K-MEANS CLUSTERING:

E. HYBRID SIMILARITY:

```
C:\Users\s.shah\AppData\Local\Programs\Python\Python36-32\python.exe
Cosine Similarity : 0.9922778767136677
Jaccard Similarity :0.2
SIM : 0.5961389383568338

Process finished with exit code 0
```



V. CONCLUSION

Recommendation system gives users new opportunities of retrieving personalised information on the internet. This paper discussed mainly k-means algorithm, jaccard similarity and the difference in results generated after using jaccard similarity and cosine similarity. We saw that jaccard similarity just took the ratings which were common between two users and did not consider the values of the ratings which may lead to bad recommendations whereas cosine similarity took all the values of the ratings into consideration and recommended the product to the user based on the similarity value obtained after calculation, Cosine similarity therefore gives better results than jaccard similarity. But using the average of both the results we get the best results.

For clustering k-means is one of the efficient and popular machine learning algorithms. Here the datasets are classified into a k number of clusters. Also for text based clustering it involves Natural Language Processing (NLP). This method of finding groups in unstructured texts can be applied in many domains such as research segmentation and news related organizations. K means algorithm Accuracy is mainly dependent on how many clusters we make. The number of clusters can be determined using the elbow method.

VI. REFERENCES

- [1] Filippini, D., Alimelli, A., Di Natale, C., Paolesse, R., D'Amico, A., & Lundström, I. (2006). Chemical sensing with familiar devices. *Angewandte Chemie International Edition*, 45(23), 3800-3803.
- [2] Advanced Computational Intelligence: An International Journal (ACII), Vol.3, No.1, January 2016 DOI:10.5121/acii.2016.3103 25 A LITERATURE SURVEY ON RECOMMENDATION SYSTEM BASED ON SENTIMENTAL ANALYSIS Achin Jain1 , Vanita
- [3] Francesco Ricci, Lior Rokach, Bracha Shapira and Paul B. Kantor - Recommender Systems Handbook; First Edition; Springer-Verlag New York, Inc. New York, NY, USA, 2010.
- [4] Anshul Yadav, Sakshi Dhingra "A REVIEW ON K-MEANS CLUSTERING TECHNIQUE" International Journal of Latest Research in Science and Technology, Volume-5, Issue4: PageNo.13-16, July-August 2016
- [5] Michael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander, "OPTICS: Ordering Points to Identify the Clustering Structure", Proceeding of ACM SIGMOD International Conference Management of Data Mining, May 31-June 3, ACM Press, Philadelphia, Pennsylvania, United States, pp: 49-60.
- [6] Soumi Ghosh, Sanjay Kumar Dubey, Comparative Analysis of K-Means and Fuzzy C-Means Algorithms, International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013.
- [7] Shi Na, Liu Xumin, Guan Yong, Research on K-means Clustering Algorithm: An Improved K-means Clustering Algorithm, Intelligent Information Technology and Security Informatics, 2010 IEEE Third International Symposium on 2-4 April, 2010 (pp. 63-67).

VII. GITHUB REPOSITORY LINK

<https://github.com/Nisargpatel16/CSE523-Machine-Learning-AlphaElite>