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].

Collaborative filtering is the process of filtering items based on the opinions of other users. Here we are going to discuss the process of collaborative filtering and how it is used in recommending the products.

Keywords-collaborative filtering, k-mean clustering, correlation, Single value decomposition, cosine 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.

Collaborative 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.

Generally, collaborative systems report better performance than content-based, but its success depends on the availability of an effective number of user ratings [3-7]. Such systems have the problem that it suffers from the item cold-start problems which occur when recommendations must be made on the basis of few recorded ratings [3]. These problems occur because the similarity is not enough. In these situations the use of a content-based approach works as an alternative. However, this approach has its own limitations. For example, the keywords used to represent the content of the items might not be very representative. Also, content-based approaches have the limitation on making accurate recommendations to users with less ratings.

A common way to solve the problems of the above techniques is to integrate both content-based information and collaborative information into a hybrid recommendation system [9].

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.

II. Literature survey

We have implemented a simple product recommendation model using collaborative filtering and clustering. In which the data of all the details of the products and of the users who bought that product was collected.

Arthur F. [12] used User K-NN, Item K-NN, and SVD algorithms to handle the supervised learning dataset. It uses two or more recommendation algorithms as a view point Rating prediction scenario for the user who did not rate the purchased products.

Orit Raphaeli[13] used Statistical analysis Sequential and Association rule mining algorithms and advantage of work is to help the online retailers to engage the customer in mobiles at any time and anywhere and limitation is Only 3 kinds of web engagement measures are considered out of 8 measures.

Wei Wang [14] which is based on User trust factors and accurate prediction. Using Cosine Similarity measures. Here, limitation of algorithms is the mapping of individual pie charts which leads to complexity and relation between the members which is not mentioned. With the help of interactivity, we can find specific nodes in the group which support digging of data and make a way to get intermediate feedback.

III. Implementation

A. We used python to implement collaborative filtering to recommend the products to the users.

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.

$$Cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \ \|\vec{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2} \sqrt{\sum_1^n b_i^2}}$$

Higher the similarity smaller is the angle

C. We have used inbuilt libraries to find the cosine similarity and also to read the excel file.

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.

In brief, what happens in our text based clustering is that, based on the similar words of product description the clusters are made and in return a particular cluster is recommended to the user which matches the most to their input attribute.

So the step by step of the implementation approach of clustering is as follows :
At first we had converted the description rows to an array form using vectorizer. This is done because to perform any kind of operations to a string it is necessary to convert it in a binary form. Here this mapping of a text-data to a vector array is known as feature extraction.

After that we had used stopwords function from natural language toolkit (nltk). What this does is that words such as (the,is,not etc) which are there in product description, get removed or are not taken into consideration. This is done because this kind of word has no contribution to the schematic meaning of a text.

Further we need to know how a particular word in a row of product description is important to us. Thus for doing this we had used a method having two steps i.e. term frequency(TF) and inverse document frequency(IDF) .

Term Frequency

$$tf(t,d) = log(1 + freq(t,d))$$

t - term or word
d - document
tf - Frequency

Here the term ‘tf’ measures the frequency of a term t in a given document d and as a result gives higher importance to more frequent occurring terms.

Inverse Document Frequency

$$idf(t,D) = log(\frac{N}{count(d \in D : t \in d)})$$

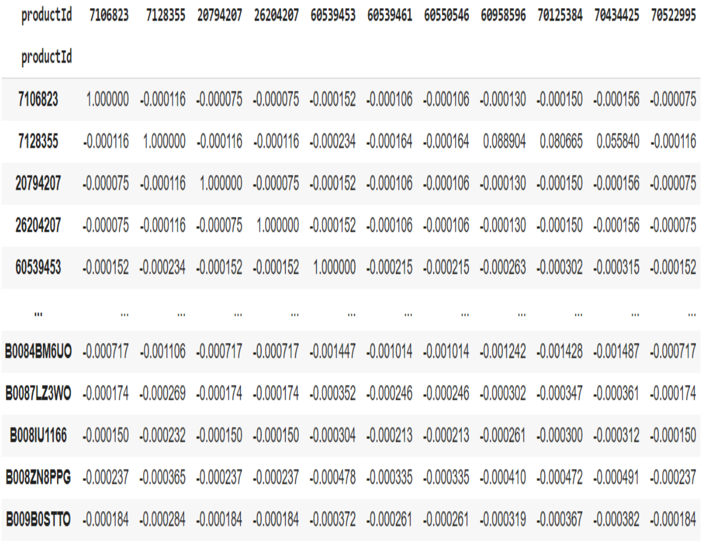
D - **Set** of document

The term ‘idf’ measures the importance of a term *t* in a set of documents D. Thus the result obtained lowers the importance of frequently used terms with low importance and increases the weight of rare terms which gives more meaning to a text.
Thus the combination of both the terms gives us the result measures of each term.

E. As a final step we used k means on the obtained vector and then plotted the group of clusters.
Therefore when a user enters a key word the function of show recommendation gives a particular relevant cluster as a recommendation.

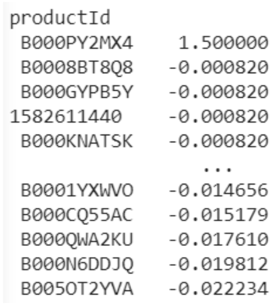
IV. Results

A. Cosine similarity matrix between the products



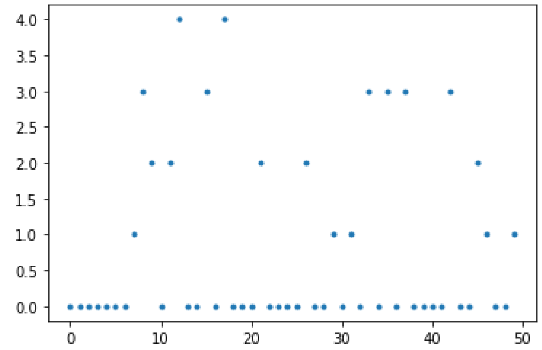
productId	7106823	7128355	20794207	26204207	60539453	60539461	60550546	60958596	70125384	70434425	70522995
productId											
7106823	1.000000	-0.000116	-0.000075	-0.000075	-0.000152	-0.000106	-0.000106	-0.000130	-0.000150	-0.000156	-0.000075
7128355	-0.000116	1.000000	-0.000116	-0.000116	-0.000234	-0.000164	-0.000164	0.088904	0.080665	0.055840	-0.000116
20794207	-0.000075	-0.000116	1.000000	-0.000075	-0.000152	-0.000106	-0.000106	-0.000130	-0.000150	-0.000156	-0.000075
26204207	-0.000075	-0.000116	-0.000075	1.000000	-0.000152	-0.000106	-0.000106	-0.000130	-0.000150	-0.000156	-0.000075
60539453	-0.000152	-0.000234	-0.000152	-0.000152	1.000000	-0.000215	-0.000215	-0.000263	-0.000302	-0.000315	-0.000152
...	...	...	...	...	...	...	...	...	...	...	...
B0084BM6UO	-0.000717	-0.001106	-0.000717	-0.000717	-0.001447	-0.001014	-0.001014	-0.001242	-0.001428	-0.001487	-0.000717
B0087LZ3WO	-0.000174	-0.000269	-0.000174	-0.000174	-0.000352	-0.000246	-0.000246	-0.000302	-0.000347	-0.000361	-0.000174
B008IU1166	-0.000150	-0.000232	-0.000150	-0.000150	-0.000304	-0.000213	-0.000213	-0.000261	-0.000300	-0.000312	-0.000150
B008ZN8PPG	-0.000237	-0.000365	-0.000237	-0.000237	-0.000478	-0.000335	-0.000335	-0.000410	-0.000472	-0.000491	-0.000237
B009B0STTO	-0.000184	-0.000284	-0.000184	-0.000184	-0.000372	-0.000261	-0.000261	-0.000319	-0.000367	-0.000382	-0.000184

B. Products recommended based on the similarity

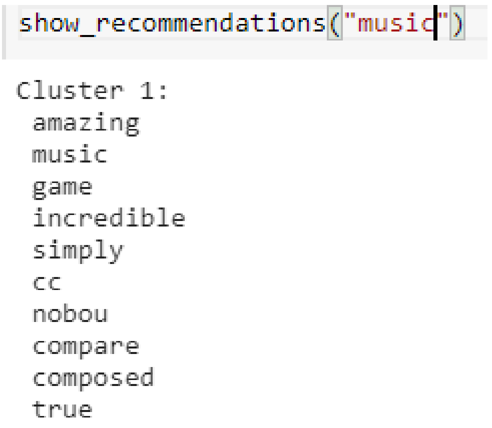


```
productId
 B000PY2MX4    1.500000
 B0008BT8Q8   -0.000820
 B000GYPB5Y   -0.000820
1582611440    -0.000820
 B000KNATSK   -0.000820
                  ...
 B0001YXWVO   -0.014656
 B000CQ55AC   -0.015179
 B000QWA2KU   -0.017610
 B000N6DDJQ   -0.019812
 B005OT2YVA   -0.022234
```

C. no. of clusters vs Features



D. Show recommendation function result



```
show_recommendations("music")

Cluster 1:
 amazing
 music
 game
 incredible
 simply
 cc
 nobou
 compare
 composed
 true
```

V. Conclusion

Recommendation system gives users new opportunities of retrieving personalised information on the internet. This paper discussed mainly collaborative filtering. Collaborative filtering has the ability to provide recommendations that are relevant to the users even without the content present in the user's profile. Here cosine similarity is used to find the similarity between the products as it gives us better results then while using euclidean distance between the points. Smaller the cosine angle higher is the correlation between the products and based on that correlation, products are recommended.

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.

References

[1] P. Resnick, H.R. Varian, Recommender systems, Communications of the ACM 40 (3) (1997) 56–58.

[2] S. Kangas, Collaborative filtering and recommendation systems, in: VTT Information Technology, 2002.

[3] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, IEEE Transactions on Knowledge and Data Engineering 17 (6) (2005) 734–749.

[4] [4] B. Marlin, Collaborative Filtering: A Machine Learning Perspective, Master's thesis, University of Toronto, 2004.

[5] H.J. Ahn, A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem, Information Sciences 178 (2008) 37–51.

[6] Q. Li, B. Kim, Clustering approach for hybrid recommender system, in: IEEE/WIC Proceedings of the International Conference on Web Intelligence, 2003, pp. 33–38.

[7] A. Gunawardana, C. Meek, A unified approach to building hybrid recommender systems, in: RecSys'09: Proceedings of the third ACM Conference on Recommender Systems, 2009, pp. 117–124.

[8] P. Melville, R. Mooney, R. Nagarajan, Content-boosted collaborative filtering, in: ACM SIGIR 2001 Workshop on Recommender Systems, 2001.

[9] R. Burke, Hybrid recommender systems: survey and experiments, User Modeling and User-Adapted Interaction 12 (4) (2002) 331–370.

[10] R.D. Burke, Hybrid Web recommender systems, Lecture Notes in Computer Science 4321 (2007) 377–408.

[11] "Issues In Various Recommender System In E-Commerce – A Survey," *Journal of critical reviews*, vol. 7, no. 07, 2020

[12] AbdulHussien, A. A. (2017). Comparison of machine learning algorithms to classify web pages.International Journal of Advanced Computer Science and Applications (ijacsa),8(11).

[13] 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.