International Journal of Scientific Research & Engineering Trends
Volume 6, Issue 3, May-June-2020, ISSN (Online): 2395-566X

Product Recommendation System

Vasudha Mishra, Aditi Rathod
Shri Shankaracharya Engineering College Bhilai, India
vasudha2194@gmail.com,aditi250199@gmail.com

Abstract **– Today the problem isn't about having paucity of information but rather it's about how we are efficiently using that information in order to get the best results in the near future. In this research paper, I have enlisted all my findings regarding the recommendation system, how can we prepare its model using machine learning, its importance in today's technological and commercial sector. I will elucidate about the applications of non- personalized recommenders, content-based recommenders and collaborative recommenders. The proposed model will take input from the customer and as a result will recommend all the relatable products in order to maximize the marketing strategy. After the extraction of the required features which are based on the type of product, the model will quantify the behaviour of the data and apply the concept of cosine similarity, it will transform the data and after arranging it in the decreasing order, it will show the resultant output of the related products. All these are the findings from various different sources which in turn throw insight of my understanding while researching through this topic.**

Keywords **– Machine learning, recommendation system, cosine similarity, content-based recommenders, collaborative filtering.**

I. INTRODUCTION

With the advancing world, where everyone wants maximum productivity within limited time and that too prefer working smart rather than hard. The same concept arises in the online business world as well. In the online business world, the main profit margin depends on figuring out the customer transaction behaviour, be it Amazon, where we are looking out for a particular product but end up buying 5-6 more, listening songs of specific genre in Spotify straight for an hour, or while watching movies or web series in Netflix with loads of similar movies already been lined below waiting for its turn.

"A lot of times people don't know what they want until you show it to them": Steve Jobs

With its highly competitive nature, recommender system becomes an essential component to all the small, medium and big enterprises in order to figure out the accurate behaviour in timely manner to make the most of the profit with the required time range. Recommender Systems have been extensively used in various sectors, leveraging the advancements of embedded sophisticated algorithms and the profusion of supported knowledge bases. Hence, Recommendation System (RSs) have brought a plethora of benefits spanning from e-business, to health informatics, to social networks, to entertainment, to many other applications. RSs in the context of software development provide proposed services to fulfil developers' needs considering their skills and the project's specifications. Every single enterprise is using

recommender system to improve their retention as continuously catering to user preferences makes them more likely to become loyal subscribers, increase their revenue manifold by using accurate "You might like" product recommendations and accelerate work by saving analyst's time up to 80% when served suggestions for materials necessary for customer for their further research. The idea is to narrow down the pool of selection options for your customers to a few meaningful choices, they are more likely to make a purchase now, as well as come back for more down the road.

Be it small and local start up or mega service like Amazon, Flipkart, Google-music, LinkedIn, Facebook etc. Having a recommendation system in their online application portal is the primary need of that enterprise to increase their sale and meet their customer demand.

Obviously, there is always something to make current machine a better self. Similarly, this recommendation system has some limitation as well, like it requires huge chunk of database to work properly, the larger the database, more accurate will be the recommendations; also with changing preference of customer, we need to take care of all the trending items that can suit customer's requirement and earlier browsing history to give them better recommendations.

But the effective boom of this particular model makes me so interested to research about this as to how it works? What are its limitations? How performance of different algorithm affects the result? And how to overcome its possible limitation, if possible, in any way? Here I have studied about overall working of product recommendation system using machine learning, what all concepts it use to quantify the ratings and past history of items and what all algorithms are used to calculate differently the cosine angles and apply training and testing over the dataset to give the best possible recommendations to their existing customers.

II. BODY

Recommender systems are information filtering systems that deal with the problem of information overload by filtering vital information fragment out of large amount of dynamically generated information according to user's preferences, interest, or observed behaviour about item. It has the ability to predict whether a particular user would prefer an item or not based on the user's profile.

Recommender systems are beneficial to both service providers and users. They reduce transaction costs of finding and selecting items in an online shopping environment. Recommendation systems have also proved to improve decision making process and quality. In e-commerce setting, recommender systems enhance revenues, for the fact that they are effective means of selling more products. In scientific libraries, recommender systems support users by allowing them to move beyond catalogue searches. Therefore, the need to use efficient and accurate recommendation techniques within a system that will provide relevant and dependable recommendations for users cannot be over- emphasized.

There are majorly six types of recommender systems which work primarily in the Media and Entertainment industry:

- Collaborative Recommender system,
- Content-based recommender system,
- Demographic based recommender system,
- Utility based recommender system,
- Knowledge based recommender system and
- Hybrid recommender system.

1. Collaborative filtering

Collaborative filtering requires collecting and analysing data about customers' behaviours and preferences to identify patterns and provide accurate recommendations based on the similarity to other users. To illustrate: if John likes items A, B, C, and D, while Mike likes items A, B, and C, he will most probably also like item D. Collaborative filtering system tries to find other like-minded users and then recommends the movies that are most liked by them. Although there are many collaborative filtering techniques, they can be divided into two major categories [15]: Memory Based approaches

2. Model Based approaches

Memory based Approach Memory-based techniques continuously analyse all user or item data to calculate recommendations and can be classified in the following main groups: CF techniques, Content-Based (CB) techniques and hybrid techniques. CF techniques recommend items that were used by similar users in the past; they base their recommendations on social, community-driven information (e.g., user behaviour like ratings or implicit histories). CB techniques recommend items similar to the ones the learners preferred in the past; they base their recommendations on individual information and ignore the offerings from other users. Hybrid techniques combine both techniques to provide more accurate recommendations.

3. Existing Similarity Measures

The most important first step in memory-based CF is similarity evaluation. The CF system in this step evaluates the similarity between the target user and other users for common rating items. The similarity is used as a weight for predicting the preference score. Various similarity metrics have been proposed in previous studies. These are as follows:

Tanimoto coefficient: It is similarity between two sets. It is a ratio of intersections. Assume that set X is {B, C, D} and set Y is {C, D, E}. The Tanimoto coefficient T of two set A and B is 0.5. This metric doesn't consider the user rating but the case of a very sparse data set is efficient.

Cosine similarity: The Cosine similarity is known as the Vector similarity or Cosine coefficient. This metric assumes that common rating items of two users are two points in a vector space model, and then calculates cosϴ between the two points.

The main agenda to calculate the cosine similarity is to quantify the similarity present between the content.

Person's Correlation: The Pearson Correlation measures the strength of the linear relationship between two variables. It is usually signified by r, and has values in the range [-1.0,1.0]. Where -1.0 is a perfect negative correlation, 0.0 is no correlation, and 1.0 is a perfect positive correlation.

4. Formation of Nearest Neighbour:

The second step after the similarity evaluation is generation of nearest neighbourhoods. To improve performance, many methods have been proposed by CF researchers. The methods for selecting nearest neighbourhoods include classification using K- means, a threshold for the number of common rating items and a graph algorithm. In general, it selects similar users greater than a given threshold or high rank users.

5. Prediction of Preference Score:

The last step in memory-based CF is to predict the preference score of the target user for non-rating items. It predicts the preference score of non-rating items for the target user, based on the rating of nearest neighbourhoods.

Various methods have been proposed, and Weighted Mean is used as most general algorithm.

6. Merits and Demerits of Memory Based Approach

User-based techniques correlate users by mining their (similar) ratings and then recommend new items that were preferred by similar users. Item- based techniques correlate the items by mining (similar) ratings and then recommend new, similar items. The main advantages of both techniques are that they use information that is provided bottom-up by user ratings, that they are domain independent and require no content analysis and that the quality of the recommendation increases over time. CF techniques are limited by a number of disadvantages. First of all, the so-called „cold start" problem is due to the fact that CF techniques depend on sufficient user performance from the past. Even when such systems have been running for a while, this problem emerges when new users or items are added. New users first have to give a sufficient number of ratings for items in order to get accurate recommendations based on user-based CF (new user problem). New items have to be rated by a sufficient number of users if they are to be recommended. Another disadvantage for CF techniques is the sparsity of the past user actions in a network. Since these techniques deal with community-driven information, they support well-liked tastes more strongly than unpopular tastes. The learners with an unusual taste may get less qualitative recommendations, and learners with common taste are unlikely to get unpopular items of high quality

recommended. Another common problem is scalability. RSs which deal with large amounts of data, like amazon.com, have to be able to provide recommendations in real time, with the number of both the users and items exceeding millions.

Content-based filtering: Content-based filtering focuses on the attributes or descriptive characteristics of items. In this approach, keywords are used to describe an item, and user profile is built to show what kind of items the user likes. The recommended items are similar to items the user has previously displayed or liked. This approach is often used for recommendations of articles or other text documents.

7. Demographic Based Recommender System:

This system aims to categorize the users based on attributes and make recommendations based on demographic classes. Many industries have taken this kind of approach as it's not that complex and easy to implement. In Demographic-based recommender system the algorithms first need a proper market research in the specified region accompanied with a short survey to gather data for categorization. Demographic techniques form "people-to-people" correlations like collaborative ones, but use different data. The benefit of a demographic approach is that it does not require a history of user ratings like that in collaborative and content-based recommender systems.

8. Utility Based Recommender System:

Utility based recommender system makes suggestions based on computation of the utility of each object for the user. Of course, the central problem for this type of system is how to create a utility for individual users. In utility-based system, every industry will have a different technique for arriving at a user specific utility function and applying it to the objects under consideration. The main advantage of using a utility-based recommender system is that it can factor non-product attributes, such as vendor reliability and product availability, into the utility computation. This makes it possible to check real time inventory of the object and display it to the user.

9. Knowledge Based Recommender System:

This type of recommender system attempts to suggest objects based on inferences about a user's needs and preferences. Knowledge based recommendation works on functional knowledge: they have knowledge about how a particular item meets a particular user need, and can therefore reason about the relationship between a need and a possible recommendation.

10. Hybrid recommendation systems

Both collaborative and content-based filtering have some drawbacks, so a hybrid recommendation system was created to solve this issue. Hybrid recommendation systems make use of both the representation of content and the similarities between users. Netflix is an example of a hybrid recommendation engine, combining the data on habits of similar users and similar characteristics to content previously liked by a user to provide awesome movie recommendations.

Here we have work on collaborative based recommendation wherein using the concept of cosine similarity, we will quantify the similarity of the objects. Then we will calculate the distance between them by using either: Euclid distance concept or Angular distance concept

Following are the steps that we will be following while designing recommendation system:

Read the required CSV file

Select features from the available dataset in order to work with them and apply recommendation

Create a column in the data frame which combine all the selected feature. Using the transform function, we will parse individual row in order to combine it vertically.

Create a count matrix from the new column. The method used herein will count the number of words to perform working within it.

Compute cosine similarity in order to quantify the similarity present within the content.

Get the index of the product from its title and arrange it in ascending order

International Journal of Scientific Research & Engineering Trends
Volume 6, Issue 3, May-June-2020, ISSN (Online): 2395-566X

Print the titles of 1st fifteen movies thus completing the concept of recommendation.

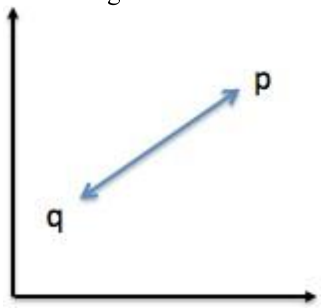
The basic terminologies and method used in our project:

Cosine similarity: It is a metric used to measure how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, higher the cosine similarity.

Euclidean distance: It used for continuous data and can be applied to both 2 and 3-dimensional system. It is just a distance measure between a pair of samples p and q in an n-dimensional feature space:

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$$

For example, picture it as a "straight, connecting" line in a 2D feature space:



The Euclidean is often the "default" distance used in e.g., K-nearest neighbours (classification) or K-means (clustering) to find the "k closest points" of a particular sample point. Another prominent example is hierarchical clustering, agglomerative clustering (complete and single linkage) where you want to find the distance between clusters.

Angular distance: The angular distance (angular separation, apparent distance, or apparent separation) between two point objects, as viewed from a location different from either of these objects, is the angle of length between the two directions originating from the observer and pointing toward these two objects. The term angular distance (or separation) is technically synonymous with angle itself, but is meantto suggest the (often vast, unknown, or irrelevant) linear distance between these objects (for instance, stars as observed from Earth).

In our project we are using the concept of angular distance using the concept of cosine similarity, as here we are interested in calculating the occurrence of the word.

In order to calculate the distance between them, we need to create vectors, for which we will use method od count vectorizer which will count words occurring specified number of times using scykit learn library of python by including class of count vectorizer and then based on its counting we will locate that on the graph with respect to origin in order to calculate the distance between them so as to find similarity among them. After which we apply the concept of cosine similarity by passing the count matrix into it which we have calculated earlier. Since the range gives the value from 0 to 1, it will give a brief as to how close the products are form one another.

After getting numerical values, we can easily analyse the data and train the model to give proper recommendation to the required system in order to increase the transaction work properly.

III. FINDINGS

Some domains are rich in non-textual content, which is currently still difficult to utilize effectively, such as multimedia content. So far, multimedia recommendation mostly relies onimplicit and explicit user feedback, but we expect that the rapid developments in ML algorithms will enable more structured and effective ways of extracting and using semantic and stylistic features in the future, as done in Messina in this issue4.

Content-based recommendation has, because of its reliance on complex textual content, traditionally been inspired by the developments in the fields of computational linguistics and natural language processing (NLP). For example, (word) embeddings are beginning to see widespread application in recommendation after its successful application in many NLP tasks.

Most of the recommender systems suffer from the cold start problem because users usually do not provide adequate ratings to hotels to enable collaborating filtering-based recommendation, which can lead to an issue called as cold start problem. Opinion-based sentiment analysis resolves this issue by considering four types of contexts: (i) guest type, which can be "business," "couple," "solo," "group," and "family"; (ii) hotel name; (iii) location; and (iv) rating about different hotels. Opinion-based sentiment analysis calculates the polarity of each sentence of review to find its score effectively solving the problem of cold start and improving the accuracy5.

Similarly, one cannot imagine manually sorting through thousands of comments, customer support conversations, or customer reviews, as there is too much data available to process. Machine learning- based sentiment analysis or classification is used to classify and provide recommendations to the users. Here the system learns from the data input given to it and then uses this learning to classify new recommendations. In supervised machine learning techniques, two types of data sets are required: training dataset and test data set. An automatic classifier learns the classification factors of the document from the training set, and the accuracy in classification can be evaluated using the test set. The key step in the supervised machine learning

5 https://www.hindawi.com/journals/sp/2019/5941 096/ this website gives a brief about all the technique is feature selection. The classifier and feature selection determines the classification performance. Some researchers have introduced another approach known as cluster-based approach. Collaborative filtering based on clustering reduces the computation time and focuses only on time efficiency improvement as the clustering phase is performed offline5.

The proposed system uses the heterogeneous nature of data (textual and numerical). We have performed following four steps to process a natural language text review as follows:

- Lexical analysis
- Syntax analysis
- Semantic analysis
- Feature extraction

Once we get proper required data, we will be able to apply various methods into it to calculate the relative similarity among t1hen and add it to our recommendation model.

IV. CONCLUSION

Recommendation systems have been anticipated are based on collaborative filtering, content-based filtering and hybrid recommendation methods and so far, most of them have been able to resolve the problems while providing improved recommendations. However, due to Several information explosion, it is required to work on this research area to explore and provide new methods that can provide recommendation in a wide range of applications while considering the quality and privacy aspects. Thus, the current recommendation research that has been made in the field of recommendation system using machine learning system needs enhancement for present and future requirements of better recommendation qualities.

Nowadays recommender system is widely used in every single enterprise and this paper has reviewed more than 15 articles in order to conclude that various enhancement has been made in the past few years and various algorithms of machine learning has been tested majority of which prefer Bayesian or decision tree approach in order to overcome all the existing limitation within the system. But still there are limitation such as the size of dataset and the efficiency in the interaction system in order to make the recommendation as effective as possible.

During the systematic review, the investigation has been made in the domains in which proposals of RSs with a ML algorithm were validated. The three most used domains are movies, documents, and product review, mainly because of the ease in accessing data. For the movie domain, Movie Lens and IMDb are two online datasets of movie ratings. For the documents and product review domains, the author notes that text is one of the most common forms of processing data, and several ML algorithms were created for text processing.

To conclude, although, Artificial intelligence and Machine Learning has made tremendous growth in every single field using recommendation system, still there are areas where we need to study and find out still effective approach for a model to achieve the best outcome.

V. ACKNOWLEDGMENT

I have great pleasure in the submission of this project report entitled Product Recommendation System Using Machine Learning for Name of the Company in partial fulfillment the degree of B.E. Computer Science & Engineering. While Submitting this Project report, I take this opportunity to thank those directly or indirectly related to project work.

I would like to thank my guide Prof. Sreejit Panicker of the guide in Company who has provided the opportunity and organizing project for me. Without his active co-operation and guidance, it would have become very difficult to complete task in time.

I would like to express sincere thanks and gratitude to Director of the College (Dr. P.B. Deshmukh), Principal, Head of Department, (Computer Science & Engineering).

While Submission of the project, I also like to thanks to Project Coordinator (Prof. Yogiraj Bhale), and the staff of Shri Shankaracharya Engineering College for their continuous help and guidance throughout the course of project.

Acknowledgement is due to our parents, family members, friends and all those persons who have helped us directly or indirectly in the successful completion of the project work.

REFERENCES

[1]. International Journal of Computer Applications Technology and Research Volume 6–Issue 2, 97-100, 2017, ISSN: -2319–8656
[2]. Sigmoidal.io/recommender-system-recommendation-engine/
[3]. https://readwrite.com/2009/01/28/5_problems_of_recommender_systems/
[4]. https://link.springer.com/article/10.1007/s11257-019-09231-w
[5]. https://www.hindawi.com/journals/sp/2019/5941 096/Figure.1. flowchart of Medical Expert System Using Artificial Intelligence.

International Journal of Scientific Research & Engineering Trends
Volume 6, Issue 3, May-June-2020, ISSN (Online): 2395-566X

AUTHOR'S PROFILE

Vasudha Mishra
SSEC alumni B.E., CSE
4th year
vasudha2194@gmailcom

Aditi Rathod
SSEC alumni B.E., C.S.E
4th year
aditi250199@gmail.com