CSE4077- Recommender Systems

J Component – Review 1 Project Report

Comparative study on recommendation systems of leading Ecommerce websites- Amazon , BigBasket & Flipkart

By

19MIA1037	B N Shrikirthi
19MIA1066	Madasu Deepika
19MIA1069	G.Harinisri
19MIA1080	Hanchate Samyuktha

M.Tech CSE with Specialization in Business Analytics

Submitted to

Dr.A.Bhuvaneswari,
Assistant Professor Senior,
SCOPE, VIT, Chennai

School of Computer Science and Engineering

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computing Science and Engineering

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

FALL SEM 22-23

Worklet details

Programme	M.Tech CSE with Specialization in Business Analytics
Course Name / Code	Recommender System / CSE4077
Slot	E1+TE1
Faculty Name	Dr.A.Bhuvaneswari
Component	J – Component
J Component Title	Comparative study on recommendation systems of leading Ecommerce websites- Amazon , BigBasket & Flipkart
Team Members Name | Reg. No	19MIA1037 B N Shrikirthi 19MIA1066 Madasu Deepika 19MIA1069 G.Harinisri 19MIA1080 Hanchate Samyuktha

ABSTRACT

In this era of massive data growth and innovations, people often face distress in decision making when it comes to choosing products for themselves. Everyday new products are launched and new products are introduced in the e-commerce we often use. So the dependency on recommendation systems has grown in a unimaginable way. Recommendation systems have become widely popular due to surge of information and this led us to do our comparative study between three massive e commerce web sites. E commerce web sites are gaining a lot of customers and revenue based on this recommendation system. According to McKinsey report amazon's recommendation algorithms drives 35% of its sales.

Amazon, Flipkart and big basket are one of the leading e-commerce websites and customers take the recommendations given by these web sites into serious consideration. An ample amount of sales for these three web sites is given by their respective recommendation systems. So we are comparing three e commerce websites Amazon, Flipkart and big basket.

In this project we are aiming to compare the recommendation systems of these three websites and we will be applying Machine Learning algorithms. After applying the ML algorithms we will be comparing which algorithm gives best recommendations for the given E-commerce website. We will be applying some such as KNN, cosine similarity algorithms and then see which algorithm gives good accuracy for the respective Ecommerce platform.

INTRODUCTION

E-commerce is the business of buying and selling varied products through the medium of internet. With the adoption to busy lifestyles, e-commerce platforms like Amazon and Flipkart have seen a surge in the number of purchasers over the last few years. From cleaning basic products to tech gadgets, from basic lifestyle choices to gifting choices, E-commerce platforms have it all. Based on the customer requirement and conscious lifestyle choices, many e-commerce sites created a niche market for themselves.

Amazon and Flipkart are two renowned and leading e-commerce sites in India that sell everything ranging from electronics, to apparel's and kitchen appliances thus catering to a large group of customers. Big Basket is majorly a grocery e-commerce site that delivers fresh vegetables, meat, fruits, snacks and groceries at the comfort of the customer.

For the purpose of this project, the three e-commerce sites have been taken into consideration. The data was cleaned and preprocessed before performing operations. Algorithms like XGBoost, Logistic Regression, Cosine similarity and KNN were applied on the three datasets and the better accuracy model was used for recommending the products from the three ecommerce sites. Data visualization and exploratory data analysis have been further performed for better understanding of the data.

LITERATURE SURVEY

Sl no	Title	Author / Journal name / Year	Technique	Result
1	Unifying Collaborative and content based filtering	Justin Basilico, Thomas Hofmann ICML ’04: Proceedings of twenty- first international conference on Machine Learning	Proposed an online algorithm that generalizes perceptron learning.	Upon designing a suitable kernel between user item pairs the results show significant improvements over standard approaches.
2	Using Content-Based Filtering for Recommendation	Robin van Metern, Maarten van Someren	Analysis the collection of web pages and extracts terms and calculates their tf-idf weights which are then stored in the database. The profiler and membership component both react to user requests.	The results are Negatively influenced by the fact that the same concept can usually be described with several terms and many terms have more than one meaning. This makes the user profile less accurately, especially because the documents in the collection are relatively short and normally only a few documents about the same topic are selected by the user.
3	THE WAR BETWEEN FLIPKART AND AMAZON INDIA: A STUDY ON	Dr Samrat Bharadwaj	The study also ponders upon investigating the major factors that ultimately impact customer	It is observed that both Flipkart and Amazon India are into deep

	CUSTOMER PERCEPTION	Journal on Marketing,2019	satisfaction towards Flipkart and Amazon. The questionnaire focuses upon the various domains which customers generally emphasises upon while shopping online like order tracking and delivery, website usage, product availability, payment procedures etc. The paper concludes by stating that in the war between Flipkart and Amazon; Flipkart wins by providing an efficient delivery system, user-friendly website and exact tracking facility.	neck competition and are in a brutal war where one tries to wipe the other out. Both these firms are seen to apply various strategies fromtime to time in order to make the other feel their presence. Though people are more attracted towards Flipkart, yet it was observed that many of them choose not to retain with Flipkart, rather switch to Amazon India for its better quality and range of products. Flipkart should learn from its mistake and should make a balance between convenience, quality and quantity in order to retain its leadership position in the long run.
4	SECURE GROCERY RECOMMENDATION SYSTEM USING BLOCKCHAIN.	RAI, AKARSH; SHAIKH, SHAKIB; VISHWAKARMA, RAHUL Journal on Software Engineering,202	The main goal of developing a hybrid recommendation system model, helps the user to get the best product available without making any extra efforts online. This recommendation will keep track of user search history and carts which were previously created. The user credentials and personal information will be secured using blockchain and the data analysis will be carried out using various algorithms.	This recommendation system will provide the latest and best price from different online stores such as Amazon, Bigbasket, Grofers. It will make a recommendation based on recent searches and the most frequent item which a user may require.

5	A Systematic Study on the Recommender Systems in the E-Commerce	Pegah Malekpour Alamdari, Nima Jafari Navimipour, Mehdi Hosseinzade, Ali Asghar Safaei, Aso Darwesh Journal on recommender system,2020	Reviewing the pros and cons of traditional techniques.Expressing some of the main challenges of relevant solutions.Pointing out some aspects of RSs to improve their accuracy and functionality for future studies.	The results confirmed that most of the studies work to improve the accuracy of recommendations, but security, response time, novelty, diversity, serendipity are not considered in many papers. In this study, we found that collaborating filtering techniques were used more than all other methods.

DATASETS AND TOOLS

1) **BigBasket Products-** BigBasket Entire Product List (~28K datapoints)

Analyzing BB Products and their performance across.

Dataset name : BigBasket Product.csv

This dataset contains 10 attributes with simple meaning and which are described as follows:

1. index - That is the serial number
2. product - Title of the product
3. category - Category into which product has been classified
4. sub_category - Subcategory into which product has been kept
5. brand - Brand of the product
6. sale_price - Price at which product is being sold on the site
7. market_price - Market price of the product
8. type - Type into which product falls
9. rating - Rating the product has got from its consumers
10. description - Description of the dataset present in detail

2**) Flipkart - Flipkart Products**

Dataset name: Products.csv

This is a pre-crawled dataset, taken as subset of a bigger dataset (more than 5.8 million products) that was created by extracting data from Flipkart.com, a leading Indian eCommerce store.

This dataset has following fields:

1. product_url
2. product_name
3. productcategorytree
4. pid
5. retail_price
6. discounted_price
7. image
8. isFKAdvantage_product
9. description
10. product_rating
11. overall_rating
12. brand
13. product_specifications

3) **AMAZON**: Amazon product data from https://jmcauley.ucsd.edu/data/amazon/

Dataset name : product.csv

This dataset contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014.This dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).

2) reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
3) asin - ID of the product, e.g. 0000013714
4) reviewerName - name of the reviewer
5) helpful - helpfulness rating of the review, e.g. 2/3
6) reviewText - text of the review
7) overall - rating of the product
8) summary - summary of the review
9) unixReviewTime - time of the review (unix time)

TOOLS:

- Google colaboratory for exploratory data analysis, preprocessing / cleaning, model planning ,model building and evaluation.

- Tableau / power BI for data visualization.

ALGORITHMS

- What is recommendation system?

Recommender systems are a type of machine learning algorithm that provides consumers with "relevant" recommendations. When we search for something anywhere, be it in an app or in our search engine, this recommender system is used to provide us with relevant results. They use a class of algorithms to find out the relevant recommendation for the user.

They are required because as the choice around us is overwhelming the customers are in need to choose amongst the infinity choices , to avoid such cases recommendation system would be very helpful.

- What is content based recommendation system?

Content-based filtering uses item features to recommend other items similar to what the user likes, based on their previous actions or explicit feedback.

Because the recommendations are tailored to a person, the model does not require any information about other users. This makes scaling of a big number of people more simple .The model can recognize a user's individual preferences and make recommendations for niche things that only a few other users are interested in. New items may be suggested before being rated by a large number of users, as opposed to collective filtering.

1) TF-IDF:

TF-IDF stands for term frequency-inverse document frequency.

Term frequency works by looking at the frequency of a particular term you are concerned with relative to the document. There are multiple measures, or ways, of defining frequency: Number of times the word appears in a document (raw count).

Term frequency adjusted for the length of the document (raw count of occurences divided by number of words in the document). Logarithmically scaled frequency (e.g. log(1 + raw count)). Boolean frequency (e.g. 1 if the term occurs, or 0 if the term does not occur, in the document).

Inverse document frequency looks at how common (or uncommon) a word is amongst the corpus. IDF is calculated as follows where t is the term (word) we are looking to measure the commonness of and N is the number of documents (d) in the corpus (D).

2) COSINE SIMILARITY:

Using the Cosine Similarity, Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The output value ranges from 0–1. 0 means no similarity, whereas 1 means that both the items are 100% similar.

It is often used to measure document similarity in text analysis.Our recommendation model takes both the pre-processed features and the user’s preferences as input. Then, the user’s input and add it as a row to the metadata. The model then vectorizes the word soup using CountVectorizer function from the scikit-learn python library. CountVectorizer takes documents (different strings) and returns a tokenized matrix. Each word soup is encoded into frequencies of words in that word soup.

Our recommendation model utilizes all properties and the metadata to calculate and find the most similar item to the user input. We use the cosine function to compute the similarity score between products, where each product will have a similarity score with every other product in each of our dataset.

3) KNN:

Content based approach utilizes a series of discrete characteristics of an item in order to recommend additional items with similar properties.KNN is a machine learning algorithm to find clusters of similar users based on common product ratings, and make predictions using the average rating of top-k nearest neighbors.

4) XGBOOST:

The Xgboost algorithm is used to implement multi-classifiers to predict the data, and the multi-class calculation results are calculated to form the recommended results. The experimental results show that the sparseness of the matrix is solved to a certain extent, and the recommendation accuracy is improved.

5) LOGISTIC REGRESSION:

Logistic regression is a linear model that does not handle complex non-linear data features.The logistic regression method as a classifier to predict the user's preferences to recommend products

Logistic regression is a machine learning algorithm used to predict the probability that an observation belongs to one of two possible classes.Logistic regression is a statistical and machine learning technique for classifying records of a dataset based on the values of the input fields.

REFERENCES

1) Basilico, Justin & Hofmann, Thomas. (2004). Unifying Collaborative and Content-Based Filtering. Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004. 10.1145/1015330.1015394.

2) Meteren, Robin. (2000). Using Content-Based Filtering for Recommendation.

3) Bharadwaj, Dr. (2019). THE WAR BETWEEN FLIPKART AND AMAZON INDIA: A STUDY ON CUSTOMER PERCEPTION.

4) P. M. Alamdari, N. J. Navimipour, M. Hosseinzadeh, A. A. Safaei and A. Darwesh, "A Systematic Study on the Recommender Systems in the E-Commerce," in IEEE Access, vol. 8, pp. 115694-115716, 2020, doi: 10.1109/ACCESS.2020.3002803.

5) Alamdari, Pegah & Navimipour, Nima & Hosseinzadeh, Mehdi & Safaei, Ali & Darwesh, Aso. (2020). A Systematic Study on the Recommender Systems in the E-Commerce. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.3002803.

6) D.N.V.Krishna Reddy, D. R. (2015). A Study On Customer's Perception And Satisfaction Towards Electronic Banking In Khammam District. IOSR Journal of Business and Management (IOSR-JBM), 17 (12 (II)), 20-27.

7) Dahiya, R. (2012). Impact of Demographic Factors of Consumers on Online Shopping Behaviour: A Study of Consumers in India. International Journal of Engineering and MAnagement Sciences, 3 (1), 43-52.