2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-1451-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/Confluence51648.2021.9377058

Analysis of Online Grocery Recommendation Systems

Lamiyah Khattar
Amity School of Engineering and Technology
Amity University
Noida,India
lamiyah.1921@gmail.com

Dr. Geetika Munjal
Amity School of Engineering and Technology
Amity University
Noida,India
munjal.geetika@gmail.com

Abstract— **Recommender system has been recognized as the most effective method for information overload problem. But a generic approach with small changes is used to get results in most types of recommender systems. Grocery recommendations are unique to this generic approach because of the possibility of reordering items. This reordering criterion can very well be taken as a measure of preference compared to general rating system that is followed. This paper focuses on using the idea of reorders to make and compare different systems of online grocery recommendations. Two types of systems have been tested for grocery recommendation taking reorders into account and a comparison between them has been made at the end. The methods and metrics used for making these two recommender systems including truncated SVD, cosine similarity and k-nearest neighbours. The paper has also discussed the results of reordering criteria for grocery sales.**

Keywords—recommender systems, collaborative filtering, SVD, k-NN

I. INTRODUCTION

With the evolving demand for e-commerce, people possess access to a wide variety of products in any field/genre/requirement they can dream of. This introduces to people to more options than one can fathom or dream of looking through and generally just ends up with them choosing the first thing that catches their eye, which may or may not be a fit for their requirements. When these options fit in perfectly, customers tend to leave a good review but when it doesn't, the re-views are bad [1]. This leads to a 50-50 chance for a good or bad review. Recommender systems come in handy at this point, to reduce the possibility of a bad review and to increase customer satisfaction while only showing them products which they are more likely to prefer.

Recommender systems have changed from just a novelty to a demand, as it has been proven; better recommender systems have a direct impact on increasing sales and traffic at a particular website[. Almost all of the largest commercial Web sites use recommender systems to help their customers in finding to products to purchase. Generally, recommender systems automate personalization on the Web, enabling individual personalization for each customer. Recommender system enhances E-commerce in 3 ways.

1.1 Feasting on people's likes

Many times people open websites just to look through products. By recording every page, they open i.e. pages that intrigue their interest. Recommender system gain knowledge of a person's preferences and starts recommending / showing them items, which they might actually end up buying[3].

1.2 Cluster cross linking sales

Recommender systems are mainly based on the perspective of reading patterns. If a recommender system can read patterns in products that are bought together, next time someone adds in cart one of the items from the bundle, it can easily recommend items that are most likely to be bought with the carted item hence, cross linking the increase in sales.

1.3 Customer allegiance

Recommendations made by the previously frequently visited site would be more on point. Even if customer tries to swap to a new website, it might take weeks or months to reach the level of pin point accurate recommendation the previously frequently visited website possessed allowing them to have a preference to continue using the website that was used previously. Hence presents us long term customer allegiance to a website [4]. Thus the goal of a Recommender System is to generate meaningful recommendations to a collection of users for items or products that might interest them. Suggestions for books on Amazon, or movies on Netflix, are real world examples of the operation of industry-strength recommender systems. The design of such recommendation engines depends on the domain and the particular characteristics of the data available [5]. This paper focuses on analyzing different types of systems that can be used in grocery recommendation.

II. LITERATURE REVIEW

A lot of studies are done on recommendation system for example, Kutuzova [6] focused on analyzing the customer's behavior based on online market interaction. The authors have used the concept of item-to-item collaborative filtering and similar buying pattern of clients. It was noticed that same products may possess different characteristic based on the seller. Within a single area, the products sold might seem similar but then again be completely different. This is what made the paper a challenge, and for correct accumulation and merging, it was analyzed through trial and tests that structure and contents acted as the main important force in building the system. The study suggested an adapted recommendation system and compared it where the recommended system had used heterogeneous dataset. In another study Mindi Yuan [7] Walmart grocery recommendations have been personalized for users. The model used is a multi-level recommender system which focused on classifying and grouping bunch of item or selling them as individuals, the basket of a customer based on product weights and quality is analyzed. A multi-level co-bought model is used to make recommendation of items for each of the respective purpose. At time of selection, the different types of preferences are incorporated and a decision is found. The final result

presented that in offline cases a 11% hit rate was reached, but in case of online tests, a 25% hit rate was noticed. Gilbert Badaro[8] introduces a hybrid approach to allow people to solve the problem of finding the ratings of unrated items, by utilizing a user-item ranking matrix using the help of a weighted combination of user-based and item-based collaborative filtering. The technique provided improvements in addressing the 2 major challenges of recommender systems that are faced in today's world: accuracy of recommender systems and sparsity of data by simultaneously incorporating users' correlations and items ones. The final evaluation of the system showed the superiority of the proposed model. Furthermore, the proposed algorithm dealt with cold start users by relying on item similarity and with cold start items by relying on user similarity. Yadagiri[9] uses an anonymously entered dataset of purchases made related to online grocery shopping, and a recommender system has been presented for the prediction of future purchases. The method use is that if a utility matrix and collaborative filtering method have been applied on it so as to be able to pair users based on similarity and dissimilarity of their purchases. Those recommendations have been made and a further approach has also been taken in the paper, that is, natural language processing has been used to determine the constituents of nutrients in the products and to improve recommendations based on that fact too. The results then, provided user recommendations of the healthiest options by basing it on their previous purchase records. George Lekakos[10] In previous models, the recommendation techniques are applied in the same way along all customers, thus failing to meet the customer's need by not keeping in mind how many people a single customer is buying for, when was the last time someone bought something etc. The paper focuses on applying recommender system techniques to a physical world scenario and checking the applicability based on the results of different recommendation techniques and for their possible use in physical environments. The result is a huge success as he e-commerce recommending techniques work quite better than traditional retail ones. Ling Yanxiang[11] proposed an approach for the cold-start user problems faced in the world of recommendation. Although many efforts have been made on the "Cold-Start" problem, it is still an open problem and had become a very emergent issue in social network analysis. In this paper, the approach proposed, applies the character capture and clustering methods to address the cold-user problem (producing recommendations to new users who have no preference on any item). The use of vector cosine method was made to cluster different users into different groups. For each group, they produced the Top-N recommendation by averaging ratings of every item and choosing the top N items on the list. SongJie Gong[12] proposed a new recommendation technique to overcome the drawbacks of the majorly used collaborative filtering techniques. The author pro-posed utilizing the results of singular value decomposition to fill the vacant ratings and then use the item based method to produce the prediction of unrated items. The experimental results proved that the algorithm combined SVD method and item-based method to be reliable.

Recommender Systems are introduced as an intelligent technique to deal with the problem of information and product overload. Their purpose is to provide efficient personalized solutions in economic business domains. This paper focuses on the different systems of recommendation by employing a recommendation technique for each of them and basing it on grocery recommendation to give is the better system as well as provides any results extracted from it. Thus maintaining the integrity of the specifications

III. METHODOLOGY

Before adopting any model, pre-processing is required which includes i) Filtering out orders based on prior ordering: Orders marked prior are the orders that were placed online. The dataset also con-sisted orders made to nearest grocery stores which were not placed online, hence the need to filter them out for the process. ii) Removing aisle ids : The positioning at which aisle they were placed at, does not matter as the focud of the research is limitted to online transactions that take place. The key 'aisle_id' acted as a foreign key for different datasets that were combined in these calculations and hence editing of all data frames was necessary. iii)Removing users without enough data: In a dataset as huge as the one used, lack of cases for training a particular user would just cause inconsistencies. So, users who had placed less than a select number of orders were filtered out and removed from the data frames. iv) Reorders Calculation: Since the dataset does not provide us with a rating system, the ratings are calculated based on the data available using the number of reorders as a basis. What makes grocery recommenders different from other recommenders is that rating for recommenders such as movies, places visited etc. cannot be calculated on the basis of reorder/re-watching as someone might just feel like watching a movie once or twice, and it doesn't mean that someone doesn't like a movie if they have watched it only once. In the case of online grocery recommendation though, grocery purchase is a weekly/repetitive task. If someone has only ordered something once it can be presumed they didn't like it, and if someone it ordering a product again and again, it is safe to assume that the product is preferred by the user. Even in the likelihood cases of a rating being present, rating is not the best basis of calculation due to the possibilty of the product not being highly rated but being an afforable choice. Recommender systems aim to increase sales by recommending products likely to be bought. c)Calculating Ratings*:* The number of reorders tell us how many times a person ordered something, but it does not mean that if the person ordered something 2 times and something else 20 times, that he/she did not like what they ordered twice. Since the usage of sum of reorders calculation is not possible directly, an average rating of 2.5 was set for anything ordered by the user once and geometric progressive sum was set as rating with an increment rate of 0.5 and the number of iterations as number of reorders. These values were taken to maintain ratings on a scale of 0-5. The median of the numbers of reorders was taken as an equivalent to cap rating of 5.

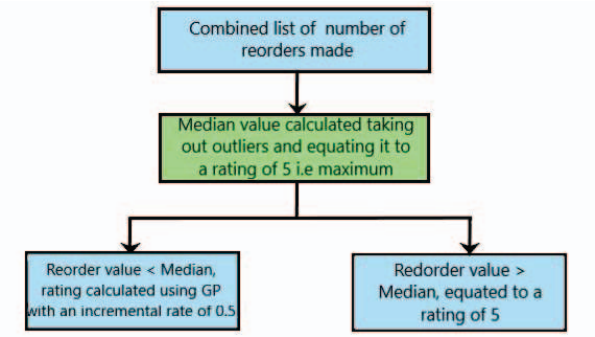

Fig 1: Flowchart representing the calculation of ratings based on reorders

1.4 Recommendation Methods

1) Item Based Recommendation: Generally, in item based recommendation, a profile specific to the user is created consisting details of the items that the user has consumed, Item based recommendation focuses on recommending items which are similar to the items mentioned in the user's profile. The advancement in machine learning has played a vital role in comparing item similarity. [13] For the implementation process, a correlation matrix is created to check the relationship between the observed phenomenon. A positive value indicated the present of a positive/direct relation between the components while a negative value indicated the presence of inverse relation and hence, a lack of similarity between the products. The values in a correlation matrix vary between -1 to 1. Similarity is later checked and based on the higher value of correlation, the product is chosen to be recommended [14].

Singular Value Decomposition

Singular value Decomposition can be seen a method used to identify and order the dimensions where the most variation is observed in accordance to the data points. Once the area of mots variation has been identified, it if possible for one to figure out the best approximate data points which are few in dimension but can be used instead of original data points. Hence, SVD can be seen as a method for data reduction. [15] (a rectangular matrix is taken to be A, such that A is considered to be n x p matrix). The Singular value decomposition theorem states:

$$A_{nxp} = U_{nxn} S_{nxp} V^T{}_{pxp} \quad (1)$$

In (1), Columns present in U are considered to be the left singular vectors; S (possesses the same dimensions as A) consists of singular values and is calculated in diagonal (*mode amplitudes*); whereas V^T consists of rows that are values of the right singular vectors (*expression level vectors*). The SVD is useful in representing an expansion of the original data in the form of a coordinate system where the covariance matrix is non – ascending diagonal consisting of values which are real in nature.

2) User Based Recommendation: In user-based Collaborative Filtering Recommender Systems (CFRS), the items possessing a similar rating, given by different users are grouped with respect to their similarity index using a particular clustering technique. Items present in a unique cluster of users are then recommended to the new users who possess characteristics similar to the users in the group[16]. To make recommendations specific to user, the first step involved is calculating the mean of ratings by a particular user for every user. By calculating mean per user, it is possible to segregate users into a high rater or a low rater. The ratings have been calculated by basing likeability on the amount of times someone reordered something, it is not plausible to make a direct comparison between two users and some users might order 100 times, while others might only order 30 times. After calculating the value of mean ratings per user, the mean value is subtracted from actual value to receive a set of ratings which are used further on.[17] By subtracting, a general value is obtained which can be used to represent a user's favourability in general. Following this, pivot tables were created with respect to the 24693 products, and were used, one representing the rating given by user to each product, and the other representing the general favourability rating that a user possesses for each product which were then used as variable input for further steps. The missing values were filled accordingly in these pivot tables .The cosine similarity was calculated and represented based on users. Top 30 users who had similar preferences as the user who the products are to be recommended are extracted. further cosine similarity was taken for the table created representing relation with the products. Identification codes of the products which are common between the users and were given a similar rating are extracted.

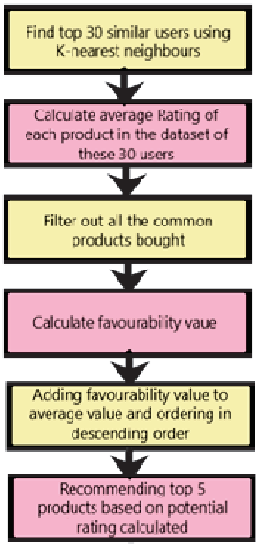

Fig 2: Flow charts representing the method on the basis of which products in user-based recommendation were recommended

The average scores the products is calculating by finding average favourability of top 30 similar users and adding it to the average rating the user gives to and product. A function is created to filter out common products bought by similar users and are considered to the products under considerations. This function is also responsible for calculate the actual favourability the user(whom which product has to recommended) would possess for the products . The correlativity of these products to the products bought by user is calculated on the previously made

function and by adding the favourability value received the average rating given by user to any product and the top 5 rated products are then recommended to the user[18]

1.5 Co-relativity Calculation Tools

. The process of clustering of data aims at collection of concert similar physical objects or maybe documents which are generally taken on the basis of some sort of likeness gauge which is additionally a challenging undertaking due to the fact most acknowledged clustering algorithms can't be generalized. So some basic calculation tools used in the clustering part of the following paper have been explained :[19]

1) k-nearest neighbours : K nearest neighbours makes use of dataset directly to make predictions. Predection are made for a new attribute (x), here which is the preferences of the user, by going through all the training dataset and searching for *k* number of instances that appear most similar attributes and summarizing the output received. In case of regression this appears to be the mean output variable, while for classification it appears to be the mode, that is ; the most common class value.

For one to be able to determine, which of the K attributes in the training dataset appear to be most similar to the new input, a distance measure is used. The lesser the distance, more similar the product is. In cases of real valued input variables, the most commonly used distance measure and the one that has been utilised is the Euclidean distance.

$$Distance(x, xi) = \sqrt{sum((xi - xii)^2)} \quad (2)$$

2) Cosine Similarity: Cosine similarity is a measure calculated between two vectors. It is said to be a measurement of orientation and is the comparison between documents on a nor-malized space. It does not follow the act of taking into account the magnitude of each word count but calculates the cosine of the angle existing between them.

$$Similarity = cos\theta = \frac{A.B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3)$$

When plotted on a multi-dimensional space, where each dimension is used to represent to a word in the document, the cosine similarity is then used to capture the orientation (the angle) of the documents.

IV. Result

A dataset provided by instacart in an open source format was used. The dataset contained orders from more than 200,000 users and each user had made between 4-100 orders. This dataset was made public for people to be able to use it to product future orders of users based on previous orderings. The experiment was performed on a filtered out sub –dataset. The filtering strategy used was so as to filter out people with respect to amount of products ordered and amount of orders made so as to be able to possess a dataset with concrete columns. The columns of the table represented different views of the dataset. Small sub-datasets were filtered out taking prospective usability into consideration. These datasets were used later to obtain a rating for the products which has later been used for predictions. The ratings are used to create a table which tells us the correlation between different products. That' correlation can be seen in the form of the graph given above. By the means of using truncated SVD [19], it was made possible to create an online grocery recommender which recommended products based on the items chosen by using the correlativity graph shown below. The correlativity comparison was checked and only those products with a large value of correlativity were recommended.

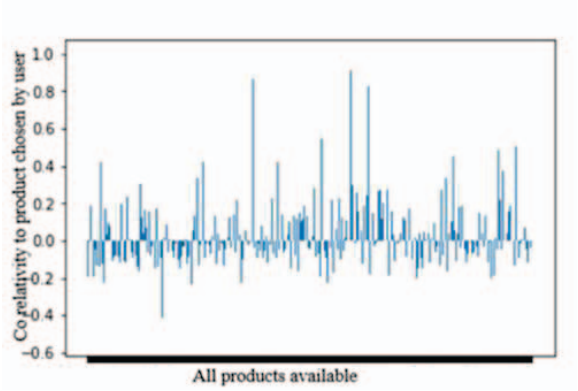

Fig 3: Graph representing the correlativity of products in comparison to the product input

In case of item based recommender 10 products were checked namely: -1) #2 Coffee Filters,2) #2 Cone White Coffee Filters,3) #2 Mechanical Pencils, 4)#4 Natural Brown Coffee Filters,5) & Go! Hazelnut Spread + Pretzel Sticks,6) +Energy Black Cherry Vegetable & Fruit Juice,7) 0 Calorie Acai Rasp-berry Water Beverage,8) '0 Calorie Fuji Apple Pear Water Beverage',9) 0 Calorie Strawberry Dragon fruit Water Beverage, 10) 0% Fat Black Cherry Greek Yogurt, the correlativity level was kept above 95%. It was noted that Products 2,7,8 and 10 received 0 recommendations while product 6 received as many as 49 recommendations. It is to be noted that this represents cases where unpopular products will not be recommended anything on being chosen, while number of recommendations for popular products will keep increasing.

As one can see, item based recommending is not the best kind of recommendation process for grocery recommendations. The user based recommendation system, takes into account all the products ordered by all users. It uses previous purchases and ratings (calculated using the amount of times they had reordered), and uses cosine similarity to calculate the similarity between different users. By the method of k-nearest neighbours, these users are filtered out. The products commonly ordered by all these users are filtered out and a relative rating is calculated for these products based on the similarity in ratings the possessed with the main user. Then the potential rating of user (to whom the items are to recommended) would be calculated. The products possessing the top 5 potential ratings are then recommended

TABLE I Products recommended to first 10 users based on user based collaborative filtering

User	Recommendations	User	Recommendations
1	Spring Water Natural Spring Water Clementines Strawberries Natural Arteisan Water	7	Banana Egg Whites Small Curd Lowfat 2% Milkfat Cottage Cheese Organic Baby Spinach Smartwater
2	Original Plain Yoghurt Extra Fancy Unsalted Mixed Nuts Organic Eggs Large Egg Whites Small Curd Lowfat 2% Milkfat Cottage Cheese	10	Bag of Organic Bananas Baby Cucumbers Grade AA Large White Eggs Organic Baby Spinach Trail Mix Fruit & Nut Chewy Granola Bars
3	Organic Whole Milk Vitamin D Organic Whole Milk Organic Raspberries Bag of Organic Bananas Smartwater	11	Banana Small Curd Lowfat 2% Milkfat Cottage Cheese Organic Strawberries Organic Baby Spinach Egg Whites
4	Banana Organic Strawberries Cocktail Style Cubes Spring Water Ice Fresh Organic Blueberries Smartwater	12	Spot's Stew Grain Free Wholesome Chicken Recipe Cat Food Banana Chicken & Lobster Formula Canned Cat Food Italian Sparkling Mineral Water Cocktail Style Cubes Spring Water Ice
5	Cocktail Style Cubes Spring Water Ice Smartwater Lime Sparkling Water Natural Artesian Water Banana	13	Banana Bag of Organic Bananas Organic Baby Spinach Cocktail Style Cubes Spring Water Ice Smartwater

TABLE II TOTAL ORDERS OF PRODUCTS RECOMMENDED REPETITIVELY IN TABLE I

Name of Product	Total orders of repetitive recommendations
Natural Artesian Water	98
Bag of Organic Bananas	1823
Smartwater	93
Organic Strawberries	1722
Banana	2028
Organic Baby Spinach	1552

From the Tables I and II, most of the products that have been recommended repetitively have been ordered a very high number of times. From this it is safe to conclude that, at the end of the day, the most recommended items would be the most ordered ones because of the excess amount of data present related to them.

V. CONCLUSION AND FUTURE WORK

An online grocery recommender was created taking the orders given by users into consideration. The recommender was capable of recommending products based on item based collaborative filtering. The research is capable of recommending items based on the similarity of items when checked. Future goals would be directing recommendations based on how often the order the products by taking the time interval between orders in the account. The segregation of recommending products based on the brands frequently purchased can also be done. The predictions made can be further defined by taking these factors into account and labelling products as healthy or unhealthy or neutral based on their departments since manually filtering each product into those could be considered a machine learning research in itself.

REFERENCES

[1] Reichheld, F. F., & Schefter, P. (2000). E-loyalty: your secret weapon on the web. *Harvard business review*, *78*(4), 105-113.

[2] Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. In *Recommender systems handbook* (pp. 257-297). Springer, Boston, MA.

[3] Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. Egyptian informatics journal, 16(3), 261-273.

[4] Schafer, J. B., Konstan, J., & Riedl, J. (1999, November). Recommender systems in e-commerce. In Proceedings of the 1st ACM conference on Electronic commerce (pp. 158-166).

[5] Dataset : "The Instacart Online Grocery Shopping Dataset 2017", Ac-cessed from https://www.instacart.com/datasets/grocery-shopping-2017

[6] Tatiana, K., & Mikhail, M. (2018). Market basket analysis of heterogeneous data sources for recommendation system improvement. Procedia Computer Science, 136, 246-254.

[7] Yuan, M., Pavlidis, Y., Jain, M., & Caster, K. (2016, November). Walmart online grocery personalization: Behavioral insights and basket recom-mendations. In International Conference on Conceptual Modeling (pp. 49-64). Springer, Cham

[8] Badaro, G., Hajj, H., El-Hajj, W., & Nachman, L. (2013, July). A hybrid approach with collaborative filtering for recommender systems. In 2013 9th international wireless communications and mobile computing con-ference (iwcmc) (pp. 349-354). IEEE

[9] Bodike, Y., Heu, D., Kadari, B., Kiser, B., & Pirouz, M. (2020, March). A Novel Recommender System for Healthy Grocery Shopping. In Future of Information and Communication Conference (pp. 133-146). Springer, Cham

[10] Charami, M., & Lekakos, G. (2009). Applying Recommendation Tech-niques In Conventional Grocery Retailing. In MCIS (p. 135)

[11] Yanxiang, Ling, et al. "User-based clustering with top-n recommendation on cold-start problem." 2013 Third International Conference on Intelli-gent System Design and Engineering Applications. IEEE, 2013

[12] Gong, SongJie & Ye, HongWu & Dai, YaE. (2009). Combining Singular Value Decomposition and Item-based Recommender in Collaborative Filtering. 769 - 772. 10.1109/WKDD.2009.132

[13] Process of data science, data analysis being a part - uploaded by Andrea G. B. Tettamanzi

[14] Xue, F., He, X., Wang, X., Xu, J., Liu, K., & Hong, R. (2019). Deep item-based collaborative filtering for top-n recommendation. ACM Transactions on Information Systems (TOIS), 37(3), 1-25.

[15] Baker, K. (2005). Singular value decomposition tutorial. The Ohio State University, 24.

[16] Breese, J. S., Heckerman, D., & Kadie, C. (2013). Empirical analysis of predictive algorithms for collaborative filtering. *arXiv preprint arXiv:1301.7363*..

[17] Middleton, S. E., Shadbolt, N. R., & De Roure, D. C. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 54-88.

[18] John S. Breese, David Heckerman, and Carl Kadie 1998. Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98), pp 43-52.

[19] Gulati, S., & Munjal, G. (2015, March). Algorithms for clustering XML documents: A review. In 2015 International Conference on Advances in Computer Engineering and Applications (pp. 654-658). IEEE.

[20] Zhou, X., He, J., Huang, G., & Zhang, Y. (2015). SVD-based incremental approaches for recommender systems. *Journal of Computer and System Sciences*, *81*(4), 717-733.