Sequential Purchase Recommendation System for E-Commerce Sites

Shivani Saini, Sunil Saumya, Jyoti Prakash Singh

Sequential Purchase Recommendation System for E-commerce Sites

Shivani Saini, Sunil Saumya, and Jyoti Prakash Singh

National Institute of Technology Patna, Bihar, India
Email id: (shivani.cspg15,sunils.cse15, jps)@nitp.ac.in

Abstract. To find out which product should be recommended to the customer and when to recommend is done by the recommender system. Different approaches by using customer profile and product description are used to build recommender system. Although these information are not enough to recommend, sometimes buying of some products occurs in a stepwise manner, where buying of one product follows the buying of other products. The purpose of this research is to find the sequences followed by customers while purchasing products to improve the efficiency of recommender system. Sequence pattern mining is used to find out the order of purchasing products. The duration we find tells the time gap between the purchased product and recommendation of next sequential products.

Keywords: Data mining, Sequential Pattern, Recommendation System, E-commerce

1 Introduction

Recommendation of products to attract their customers have become norm of every e-commerce website. A good recommendation system surely increases business of these sites as users may find their choice without too much searching. The analysis of popular e-commerce website, such as *flipkart.com*, *amazon.in* and *snapdeal.com* etc. reveals that recommendations to the users are made based on their browsing history or user's previous purchase pattern. Most of these recommendation system is applied to new products or services. But there are several merchandise which is regularly used by the users and brought in by the user at regular interval. An example of such products and recommendation is recently started by Amazon, which they have termed *subscribe and save* option. With *subscribe and save* option, Amazon offers some extra discounts on some selected products. On careful and detailed analysis, it was found that most of these items belong to grocery and packaged food items such as *Coffee, Tea* etc. Further analysis shows that not all grocery items are put under this option. For example, Amazon has given *subscribe and save* option for the product "Bru instant coffee 100g" but not for "Nescafe instant coffee 100g". This was the main motivation behind this work as why and how Amazon has decided some products to put them under *subscribe and save* option. Another interesting thing is that in Amazon *subscribe and save* option only same items are offered for discount. In this article, we are trying to find the sequence of all items which are brought regularity. We are not only finding the same product purchased every month, but, also the different products purchased one after another in

a sequence. This type of mining generally used for sequential data, such as Books (divided into parts or the story in a sequence), TV serials, Movies (divided into parts or the story in a sequence). But, we believe this type of sequences can also exist between one or more products. User buy some products in a sequence, for example, most of the user buy mobile phone and mobile cover in a sequence. So, we are trying to find out such kind of sequences, in online shopping as shown in Table 1.

Table 1. Products purchased by the users

User	**Month**	**Items Purchased**
User1	January	Soap, Coffee, Mobile Phone
User1	February	Book, Coffee
User1	March	Mobile cover, Coffee
User2	January	Coffee,Tea
User2	February	Coffee, Book
User2	March	Coffee
User3	January	Mobile Phone
User3	March	Mobile cover

From the Table 1, it is clear that the purchasing nature of the different user may not be similar. User1 and user3 have the similar purchasing patterns. They first purchase mobile phone, then purchased the mobile cover. Similarly, user1 and user3 has the similar purchasing patterns, as they have repeatedly purchased coffee every month. In this article, we are trying to find out the common purchase sequences among all the users. The sequences may consist the same items or different ones. Our main objective of this article is to find out the sequences in the online product purchasing system, i.e., the sequences frequent among all users and Intra-duration in the sequence.
The rest of the article is organized as follows: Section 2 is the literature review. In section 3 we discuss the methodology by which we are finding the frequent purchase pattern sequences. The results are explained in section 4, in section 5 we discuss our findings and section 6 is concluding our work.

2 Literature Review

In this section, a brief introduction about the recommendation system is presented. Recommender Systems are software tools and techniques that give the suggestion to users to see or buy the items based on their browsing history, previous purchase history or by using their pattern of purchase history [10] [3]. A recommendation system is widely used in almost every field such as movie recommendation, music, book, news, television shows, community question answer website, product recommendation, and many others. Since, taste of persons is not similar so, the recommendation is also not similar for all users.

A recommendation system is basically divided into three types: a) Content based filtering [6], b) Collaborative filtering [2] and c) Hybrid approaches [4].

a) Content-based filtering: This works with data that are provided by the users either explicitly (ratings) or implicitly (clicking on a link). Based on these data a user profile is generated to perform the recommendation to the similar user. The more participation of a user leads more accurate recommendation. Recommendation using the content is performed using the similarity score between the user profile and item profile, and finally, the top score item is recommended to the user. Since, the recommendation is performed based on user previous purchase history so, the most difficult problem of this approach is recommendation for new users, as there is no purchase history availability of new users.

b) Collaborative filtering: It is a technique of making an automatic prediction system about the user with the help of other similar user's choice or information. Assumption used in collaborative filtering is to select and aggregate other user's opinion to provide a better recommendation of the active user's preferences. Probably, they assume that, if users agree about the quality or relevance of any items, then they may agree about other items. For example, if a group of user like the same product as *user x*, then *user x* is likely to like the product they like which he hasn't yet seen.

c) Hybrid filtering: The concept of content based filtering and collaborative filtering is combined, to predict the next item more accurately. A work introduced by Liu et al. [8] used hybrid recommendation method that combines the segmentation-based sequential rule method with the segmentation-based KNN-CF method. The proposed method is based on user's RFM values. Where RFM (R= Recency, F= Frequency, and M=Monetary) is indicating the user activeness on the e-commerce website. The RFM value will be used to group the user in various clusters. Choi et al. [5] proposed a work which is the hybrid of implicit rating and explicit rating. They integrate collaborative filtering approach with a sequence pattern algorithm for improving the recommendation quality.
Mcauley et al. [12] built the recommendation system on the basis of product image and its matching accessories. Another work proposed by Mcauley et al. [11] built a network of substitutable and complementary products.
None of the above talked recommendation system focused on the sequences occur in the user's previous purchase history in the online purchase system. The problem of sequential pattern mining (SPM) was first introduced by Agrawal and Srikant [1]. In [1], the SPM was defined as follows: From a given database of sequences, where each sequence consists of a list of different transactions ordered by transaction time and a set of items, sequential pattern mining basically mines all such kinds of sequential patterns with a user specified minimum support value. Minimum support of a pattern is defined as the number of data sequences that contains such patterns. The discovery of such sequence required for various types of algorithms [1]. Many approaches are used to find out what would be the next product purchased by the user. Haiyun Lu [9] proposed an idea for recommendation of items which is based on sequential pattern mining. They used the users previous purchase history data to analyze the user purchase behavior at a particular location. The patterns are used to recommend the next category purchase item to a user in a particular location. Huang et al. [6] proposed a system based on

sequential pattern which predicts the customer's time-invariant purchase behavior for food items in a supermarket. Khandaga et al. [7] proposed a mechanism which focused food recommendation system. As, today it is the biggest question "WHAT TO EAT". People always getting confused with their food choice. If a system recommends a right food items, then the user may like the system.

3 Methodology

It may be possible that a user purchase more than one item together but not always. There is a high possibility that if item1 is purchased today, then after a few days item 2 would be purchased. Which item would be purchased together have well explained by Agrawal and Srikant [1]. They introduced Apriori algorithm in which, the whole dataset is scan number of times and with the help of user input minimum support and confidence value, the frequent purchase item set was extracted. For example, if item A and B are frequent pattern, then the association rule might be either $A \rightarrow B$ *or* $B \rightarrow A$ or it may be possible that item A and B purchased together. But, Aprioi is not able to find out the exact order in which the product might be purchased by the user. To resolve this issue, Sequential Pattern Discovery using Equivalence classes (SPADE) algorithm was introduced by Zaki et al.[13].
In this article, we are working with amazon dataset. With the help of SPADE algorithm we are trying to find out Frequent Sequential Purchase Pattern. The flow chart of our proposed work is shown in Figure 1. In Figure 1, U1, U2, U3 are users and A, B,C,D are products. Since, the structure of the dataset is not formatted as we required, so we have done some pre-processing steps to convert the dataset in our required format. In the next step, we apply sequence mining algorithm [13] to find out the sequences available in the dataset. Next we find out the time gap between the purchase of first product and next sequential product.

3.1 Dataset

To perform our analysis, we download the amazon dataset, which is available online[1]. It contains 82,677,139 (approx. 82 million) ratings of 9,874,213 products given by 21,176,523 users. Ratings are given by the user, since the year 1997 to 2014. Our proposal consists some assumption that is listed below:

Assumptions

- The transaction data is not available due to security and privacy concern. So, we are assuming that the user has given the review after purchasing the item.
- We are not concerned about the rating given by the user.

The Amazon dataset format is shown in Table 2. Here Product ID is *asin* (Amazon Standard Identification Number) number of the product which is used by Amazon to uniquely identify the products.

[1] http://jmcauley.ucsd.edu/data/amazon/

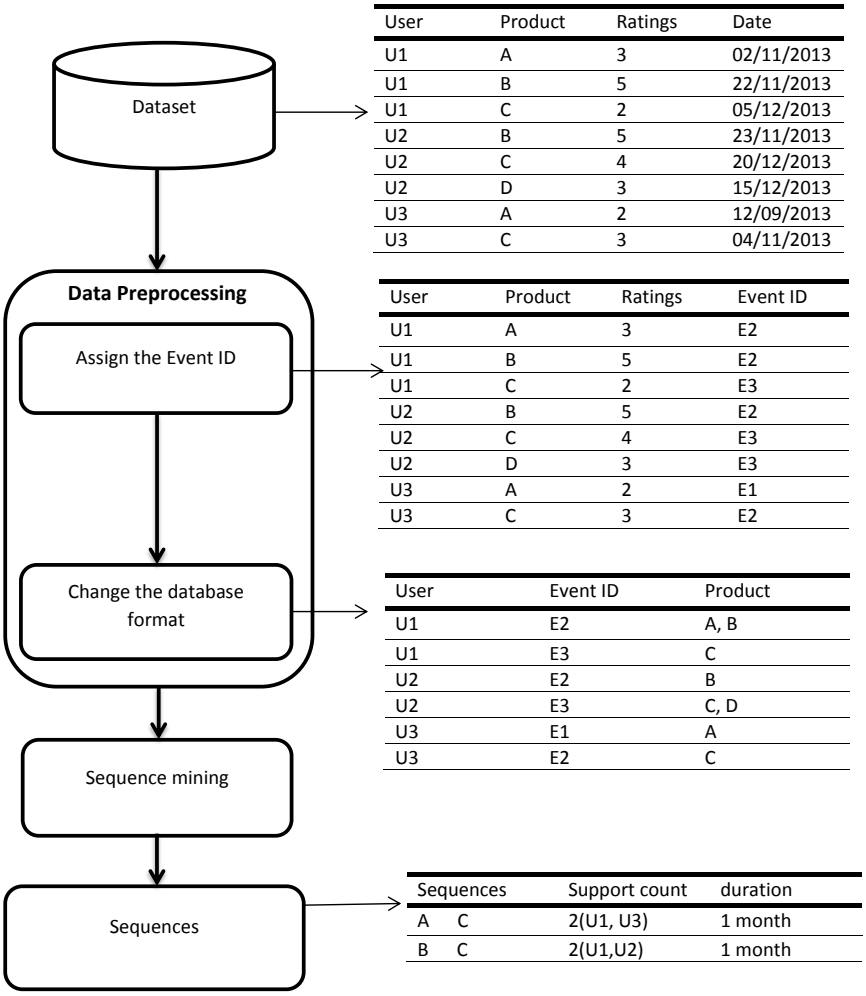

Fig. 1. Flowchart of proposed work

Table 2. Snapshot of dataset

USER ID	PRODUCT ID	Ratings	DATE
A1CCQTW8Q1XJ6E	B0002QB9NE	3	15-12-2013
A14R9XMZVJ6INB	B00EM5POSW	5	21-12-2013
A14RFF9JUIM34U	B00004Z1SX	2	02-12-2013
A14RFF9JUIM34U	B0002L5R78	1	14-10-2013
A14S2P9NK1V9VW	B000GQVVU6	3	26-01-2013

3.2 Data Preprocessing

In this section, we discuss about the data preprocessing steps. An example of the data preprocessing steps is shown in Table 3, in which the Table 3(a) is the same dataset format that we downloaded from amazon website and the Table 3(b) is coming after the preprocessing step.

Table 3. Change the database format

USER	DATE	ITEMS
U1	02/01/2016	A
U1	05/01/2016	B,C
U1	01/02/2016	D
U1	20/02/2016	A
U2	03/01/2016	A
U2	05/01/2016	B

Table 3(a): Before preprocessing step

SID	**EID**	**ITEMS**
U1	E1	A,B,C
U1	E2	D,A
U2	E1	A,B

Table 3(b): After preprocessing step

Where, SID is a sequence ID. We are considering one user as a one sequence as we are finding sequences trending among all users, EID is an event ID. We are binding whole month transaction with the same event ID and ITEMS are the product purchased by the user in a month. In the above example A, B, C, D are the products. In Table 3:

- E1= Items purchase in January 2016
- E2= Items purchase in February 2016

Set the event: Set the event such as week, month, year, etc. If we choose month as event, then we will assign the same event ID for that month and we get monthly sequences, e.g.,

$$A \rightarrow B$$

The user has purchase A then after some months B will be purchased by the user.

3.3 Sequence Mining Algorithm

Any sequence mining algorithm can be used to find out the sequences. Here we are using SPADE algorithm. Sequence mining is generally used for sequential or episodic data. Two types of sequence on a product:

1. **Same products repeating:** Users repeatedly buy the same items monthly (or weekly, yearly etc)basis. This type of sequences falls in this type.

   $$A \rightarrow A$$

   Example: Sequence found in a serial or episodic data, i.e.,books, TV serials, Movie series

2. **One after another**: If a user buys different items in a sequence, then this type of sequences will come under this category.

$$A \rightarrow B$$

Example: Mobile phone → Mobile case

3.4 Intra-duration

There is one more important aspect of recommender system is when to recommend the recommended product. The efficient recommender system should recommended user when they need it. So time plays an important role in recommender system. Here we find out the time elapsed between the purchase of first product and the next sequential products. For example, if we have sequence A → B then we find after how many months the user is purchasing B once he purchased A. For this, we are finding mean and mode of the duration followed by all users. Here, mean gives the average time gap between products, whereas, mode gives the duration followed by most of the users.

4 Result

The algorithm for preprocessing data and finding sequences are implemented in Python. The algorithm was executed on a 64 core server having 64 GB of RAM. To evaluate the result we split our dataset into train dataset and test dataset as shown in Table 4. On train dataset we built our recommender system however, test dataset was used to check its performance.

Table 4. Train test split

Dataset	No. of record	No. of user	No. of product
Train Dataset	6,22,528	3,627	3,36,489
Test Dataset	2,66,240	1,555	1,74,561

Table 5 represents some of the frequent item sets returned by our system . The first and second row of the Table 5 contains one item set while row 5 contains 2 item sets brought together. Row three and four of Table 5 contains the sequence of two items brought in order A → B where A represents the first item and B represents the second item. The supports counts (Number of users bought the items) of the frequent items are also shown in column three. We were only interested in the sequences of the item that are purchased by the user. In our dataset we got 268 such sequences.

Table 5. Frequent items

Sr.no	Frequent items(asin)	Support count
1	B00934WBRO	199
2	B0026ZYZ7Q	145
3	B00934WBRO → B00B9AAI9S	86
4	B0026ZYZ7Q → B00B9AAI9S	81
5	B001AIJZQ6, B0021YV8LS	57

Table 6 represents the frequent sequence along with the duration between purchasing of first product and the next sequential products. The fourth column of Table 6 shows the average duration represented as ***d1***. The next column of the same table shows duration followed by most of the user represented as ***d2***. Both ***d1*** and ***d2*** represents duration in months (as described in Section 3.4).

Table 6. Sequences

Sr.no	Frequent items(asin)	Support count	*d1*	*d2*
1	B00934WBRO → B00B9AAI9S	86	3	3
2	B0026ZYZ7Q → B00B9AAI9S	81	2	2
3	B007FK3CVM → B00934WBRO	73	2	3
4	B00934WBRO → B00C88DV6M	71	4	4
5	B0013OQGO6 → B00B9AAI9S 65	57	4	5

4.1 Validation

To check the performance of the system, we used the following metrics. *Accuracy:* The accuracy of the recommendation is defined as the ratio of users who are purchasing products in a specific sequence to the users who purchase the product together or in different sequence. Say N1 number of users purchase products P1 and P2 either together or in any sequence. N2 is the number of users who are purchasing products P1 and P2 in the sequence P1 → P2. Then accuracy can be defined as

$$Accuracy = \frac{\sum \frac{N2}{N1}}{\grave{n}} \tag{1}$$

where, $\grave{n}$ be the number of the sequences followed by some users (at least one user). The accuracy measures on the scale of 0 to 1, where 1 refers 100% and 0 refers 0% accuracy. We calculated N1, N2 and N2/N1 for our test dataset and the details can be seen in Table 7. We got accuracy of 0.9 for our test dataset.

Table 7. Test results

P1	P2	N1	N2	N2/N1
B00934WBRO	B00B9AAI9S	34	34	1
B00934WBRO	B00C88DV6M	30	30	1
B00934WBRO	B009FKNGGQ	26	26	1
B00934WBRO	B0021YV8LS	25	25	1
B00934WBRO	B001AIJZQ6	24	24	1
B00934WBRO	B001AIJZQ6	24	24	1
B0026ZYZ7Q	B00B9AAI9S	35	34	0.971428571
B007FK3CVM	B00934WBRO	28	27	0.964285714
B007FK3CVM	B00934WBRO	28	27	0.964285714
B007FK3CVM	B00934WBRO	28	27	0.964285714
B0026ZYZ7Q	B00C88DV6M	26	25	0.961538462

5 Discussion

Our proposed system extracted around 268 sequences that are found to be frequent for the dataset used. The system also calculated the mean and mode duration after which these sequences are followed. Our result includes most of the items listed in Amazon's *subscribe and save* option which supports our results. Since, Amazon's *subscribe and save* option includes single item which is repeated after specified month. The current proposal enhanced the recommendation system by recommending different items which are brought one after another after a gap of some months.

6 Conclusion

Sequential pattern mining has played an important role for accurate recommendation system. As, if we are able to find out the purchase sequence of users with respect to the time then we recommend, the more accurate product to the users that helps to minimize the user search time as well as improve the companies sell. In this article, we find out such purchase sequences of the user from amazon data set using SPADE algorithm and time duration within the sequences. So, we can recommend the next sequential product to user after some months. Here we evaluated those sequences which had a time gap of more than one month. We can decrease these time gaps to 1 day or a week. With this modification we would have more sequences which occur in short duration of time. There are some sequences which are common among all the users, so we have found only those sequences which are popular among all the users. However the future work can find sequences for specific user, or similar user by applying the same method. Future work can also include sequences which are followed by the user in different years.

References

1. R. Agrawal, R. Srikant *et al.*, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, 1994, pp. 487–499.

2. J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1998, pp. 43–52.
3. R. Burke, "Hybrid web recommender systems," in *The adaptive web*. Springer, 2007, pp. 377–408.
4. Y. H. Cho, J. K. Kim, and S. H. Kim, "A personalized recommender system based on web usage mining and decision tree induction," *Expert systems with Applications*, vol. 23, no. 3, pp. 329–342, 2002.
5. K. Choi, D. Yoo, G. Kim, and Y. Suh, "A hybrid online-product recommendation system: Combining implicit rating-based collaborative filtering and sequential pattern analysis," *Electronic Commerce Research and Applications*, vol. 11, no. 4, pp. 309–317, 2012.
6. C.-L. Huang and W.-L. Huang, "Handling sequential pattern decay: Developing a two-stage collaborative recommender system," *Electronic Commerce Research and Applications*, vol. 8, no. 3, pp. 117–129, 2009.
7. S. Khandagale, S. Mallade, K. Kharat, and V. Bansode, "Food recommendation system using sequential pattern mining," *Imperial Journal of Interdisciplinary Research*, vol. 2, no. 6, 2016.
8. D.-R. Liu, C.-H. Lai, and W.-J. Lee, "A hybrid of sequential rules and collaborative filtering for product recommendation," *Information Sciences*, vol. 179, no. 20, pp. 3505–3519, 2009.
9. H. Lu, "Recommendations based on purchase patterns," *International Journal of Machine Learning and Computing*, vol. 4, no. 6, p. 501, 2014.
10. T. Mahmood and F. Ricci, "Improving recommender systems with adaptive conversational strategies," in *Proceedings of the 20th ACM conference on Hypertext and hypermedia*. ACM, 2009, pp. 73–82.
11. J. McAuley, R. Pandey, and J. Leskovec, "Inferring networks of substitutable and complementary products," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 785–794.
12. J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel, "Image-based recommendations on styles and substitutes," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015, pp. 43–52.
13. M. J. Zaki, "Spade: An efficient algorithm for mining frequent sequences," *Machine learning*, vol. 42, no. 1-2, pp. 31–60, 2001.