Journal of Studies in Science and Engineering

EngiScience Publisher

Research Article

Shopping and Basket Analysis by Using an Improved Apriori Algorithm in WEKA

Shahab H. Kaka Ali [1,*], Ibrahim Berkan Aydilek [1]

[1] Department of Computer Engineering, Faculty of Engineering, Harran University, 63050 Şanlıurfa, Turkey

*Corresponding Author: Shahab H. Kaka Ali, E-mail: shahabengit@gmail.com

Article Info

Article History

Received Nov 20, 2021

Revised Nov 30, 2021

Accepted Dec 01, 2021

Keywords

Recommendation system

Apriori Algorithm

E-commerce

Data mining

Association rules

Improvement of Apriori Algorithm

Abstract

In the past years, e-commerce and online shopping have grown fast. It became more helpful by letting people buy the desired product online. Also, to help their users to find the product of their desire easily and make the process simpler, the online shopping websites use some kinds of algorithm to provide recommendation systems. Often, these systems use techniques like basket analyzing and association rules which is finding the relation between the products together or between users too, so the apriori algorithm is one of the famous ones among the recommendation systems. Although it has some limitations while implementing, which makes the algorithm less confident or even useless, Let us assume we have 100K records in the sold item list in a system in which about 10K refers to the customers buying only one or two items in their purchase. Therefore, this ten per cent will not affect finding the relation between the items, at the same time these records will make the system less efficient and take more time to analyze, in this paper, we try to show how we can improve the apriori algorithm efficiency and accuracy by some preprocessing on the dataset before applying apriori algorithm by eliminating the unnecessary records, this process helps to make the algorithm better because of reducing the number of transactions, hence finding strong relationships between items easier for the rest of the records.

1. Introduction

Data mining consists of the process of finding knowledge from a dataset by using some algorithms like preprocessing on data, pattern recognition, classification, association rule mining, and clustering etc., which could be using these algorithms to get information and find discovering some patterns which could be important for decision making. Usually, the Knowledge Discovery in Database is used as a synonym for data mining word, and sometimes it is described as the heart of KDD. A non-intuitive operation for identifying information which could be correct patterns, new, helpful, and understandable [1], So recommendation systems (RS) are a kind of the very prevalent of using data mining in electronic business which became base tools to help improve businesses such as their sales. RS in software engineering is the

implementation of software that generates knowledge about items assessment which could be helpful to software engineering processes for every step [2]. A search engine would be a critical filtering system, is making our search easier for an item [3]. Some recommender systems are earning renown to their qualifications within drowning the side effect of discovering an area to seeking users [4], request for advice about a specific subject are helped and increased natural social task by recommendation system [5], Within this paper, we try to show some way to create a small online item recommendation system retailer. The system will specifically be made to find the requirements of retailers by a few data sets which has small processing power, and it checks accuracy, efficiency, and scalability with real-life information from a little online retailer. The process of discovering these objects which have a strong relationship is a wasting time process. So, We need a fast and clever algorithm that could find knowledge automatically in a short time to effectively search for finding meaningful information in a massive space of the dataset that could be possible. These tools used for assisting users by supplying advantageous offers are Recommendation systems, thus decreasing their find out time [6]. So we explain the attempt to design my project (Recommendation system in e-commerce). We should think about the association rules between the objects(items) in the sold table (our dataset) according to the same basket or same customer instance it means first needs to think about basket analysis, so I chose Apriori-Algorithm for my purpose, which is most common and powerful in this filed. Basket Analysis is a crucial method known and utilized by substantial retailers to reveal relationships between products, like bread, butter etc. It works by searching for a mix of products that happen together now and then in exchanges. Deshpande with Paranjape-Voditel executed association rules about the stock market sector. They were developing a portfolio recommendation system and market basket analysis [7]. To give it another perspective, they enable retailers for generating relations between objects who each customers was purchasing. When information technology was growing, a massive amount of information was gathered and warehoused by companies. Then the important thing for companies is changing the information to valuable knowledge in the dynamic market as decision making. This value-added information discovered from Market Basket Analysis is essential for decision-making. If it will be known that the customers purchase an item like (A) and they will purchase another item like (B), So it will be possible for the system to know which items will be purchaser together, or the purchasers of target prospects for the second product.

1.1. Recommendation System

The recommendation system is an implementation that supplies and offers a product in creating a useful decision by the servant. Currently, the main recommendations include content-based recommendation, collaborative filtering, recommendation depending on association rules, depending on the activity of recommendations, wallet recommendations depending on information [8]. usage of recommendations regularly commonly makes foresee a product, like movies, fruits, and music offers. The method executes in a couple of paths, namely with gathering customer information straight or indirectly. Electronic commerce sites could be allowed by recommendation systems to personalization and customization by automated and fast, and letting the sites increase sales, they allow the sites to generate more sales by tailoring for customer's requirements and turning them for buyers, increasing selling more items with packaging most linked goods together, and customer loyalty increase [9] [10]. Customer loyalty becomes carried out by recommending products for customers that they need time for knowing their requirements and for collecting extra concerning them [11]. Direct gathering data has required the customer to rate a product. unlike indirect gathering data is by observing the items, later an e-commerce web's customer can see it. According to the discussion about various recommender systems and various usages, it could be understanding with the issues and limitations which notable by all users which choosing buying items in online markets then the recommendation's reliability is also not convinced by the 58.4% of somebody [12]. After the monitoring data is collected, later it is run to execute a specific algorithm like the Apriori algorithm. next to that, the outcomes will turn back to ten customers like a product recommendation with that user's parameters.

2. Related Works

Repeated item sets mining is one of the significant phases within association rule mining to find repeated item sets within the database records. The important data mining task is trying for discovering interesting patterns within datasets, like association rules, classifiers, clustering, and correlation, etc. There are a lot of approaches suggested to discovering repeated item sets. However, these algorithms could be indexed for a couple of classes, such as generating of candidate or development of modality. It could be used Apriori algorithm as a way for the declaration of the candidate producing. It produces length (k+1)

candidate item sets depending on length (k) frequent item sets. The item set's frequency is observed by summation of their repeating transactions.

3. Recommender System in E Commerce

The aim from the recommendation system is to recommend items depend on the user's priorities. Famous implementing of RS could be observed from the book's scopes [13], music [14], [15], and movies [16]. There are many early articles about collaborative filtering discussed recommender systems before a quarter century ago [17]. Recommender system's name is widely familiar because of doing comprise content-based filtering and collaborative filtering in addition hybrid algorithms.

4. Apriori Algorithm

It is the algorithm that is used to find out the association rules between objects. That means how two objects are associated and related to each other like two are three products or user behaviors. This means the apriori algorithm is an association rule learning that analyzes that "People who bought item X also bought item Y. The purpose of the apriori algorithm is to generate the association rule between objects. So it tells us how two or three items are correlated to each other. Apriori algorithm consists of a couple of stages to association rule mining, stage one consists of scanning all information in the data warehouse or database to discover all repeating item sets, and the next stage finding association rules in the discovered sets. The first stage is more difficult than the second step to manage because that first stage is decided on the execution of the mining association rule, so often all articles concentrate on the set one issue like eliminating the data before applying the second step [18]. Agrawal and Srikant had invented an approach by the name Apriori algorithm which could be mentioned as the basal algorithm for association rules [19]. The most important concepts of the Apriori algorithm are support and confidence.

Support = number of these customers who bought item1 / Total number of customers.

Confidence (item1-> item2) = customers who brought item1 and item2 / customers who brought item1.the apriori algorithm generates some strong rules between the items by filtering threshold with minimum support and confidence. For practicing this algorithm about the recommendation system I use Weka and I discuss all steps one by one.

5. Limitations of Apriori Algorithm

This algorithm suffers some problems in the execution time such as in a database that contains a large amount of records of these customers who bought only one or two items in their purchase, So this amount

of records will not effect on finding the relation between the items, at the same time the system will waste a long time to scan the records, because of these the system will be less efficient and less accurate, it is clear that these algorithms depend on (support and confidence), let's discuss them by an example assume that 10% of market customers bought only (A) item, so in the definition of the Apriori algorithm are:

support (item A) = number of all records which contains (A)/ number of the whole records

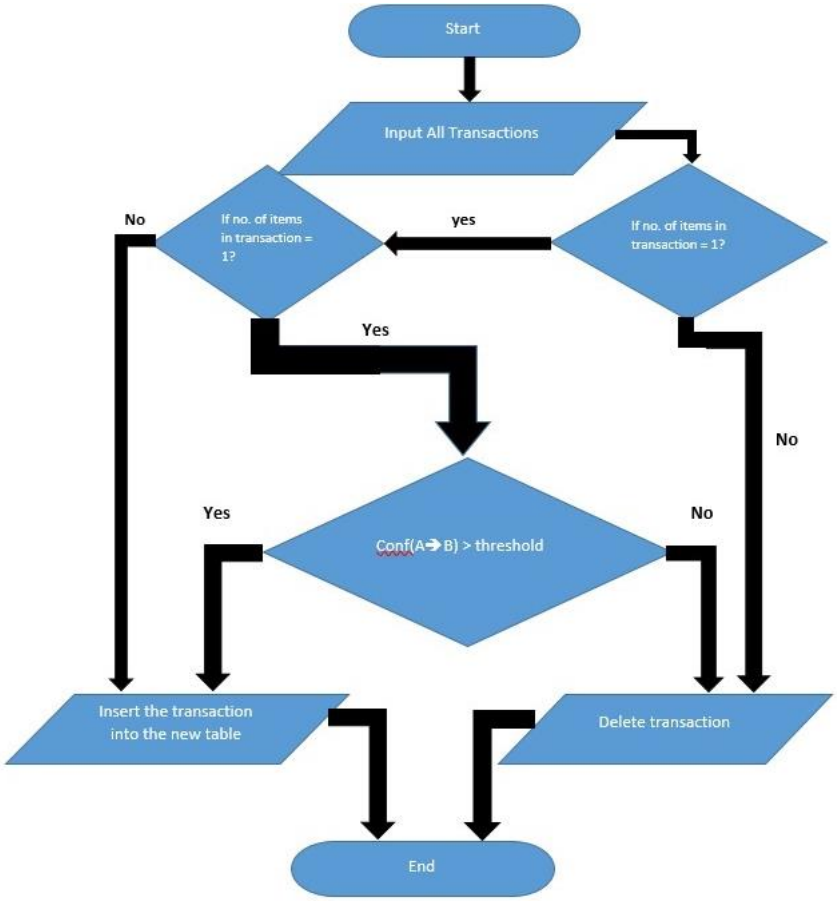
confidence (A➔B)= support(A U B)/support(A), and we can write this formula like below:

confidence (A➔B)= support(A U B)* number of the whole records /(number of all records which contains (A))

As observed in the formulas, a large number of (A) has a bad effect on the confidence of all item set that contains (A) item, and on the other hand, it leads the system to waste a lot of time implementing a large amount of data however 10% of them are unnecessary. In some cases, else if the database contains some records which are the costumers bought two items like (A, B), (C, D), …..., and (Y, Z) if these records appear only one time, also it has a same bad effect on the system too. In this point of view, this algorithm has some limitations, and it needs to improve.

6. Improvement of Apriori Algorithm

We can make some preprocessing on the datasets before Appling the apriori algorithm as an improvement of these limitations as we mentioned before like improvement its efficiency which is consist of some steps such as deleting all transaction which contains less than at least two sold items directly. And Second deleting all these records which contain these item sets in which consist of two sold items together if has lower frequents together means deleting all records in which refers to these customers bought only two items and these two items has lower repetition than the selected threshold in the dataset. Also, we can repeat the same process again and again till the datasets can be shortened without losing the important transaction mean these transactions contains a group of items that have a strong relationship together. As shown in (flowchart 1), all the datasets are inputted to the system, then the system checks if any transaction containing only one item sold directly will be deleted by the system. After that, it checks these transactions with two items then compare the confidence (A➔B) with the minimum confidence, if anyone has less confidence from the minimum that selected also, will be delete, finally the remaining transactions input to a new table or remaining in the table by deleting low confidence item sets.



Flowchart 1. Preprocessing on the dataset

Let's Discuss It By An Assumption Example That Showed By Table 1, 2, 3:

Table 1. Table of Transactions With Useless Taransaction

Transaction(T)	Sold Items(I)
T1	I1,I3,I4,I7
T2	I3,I4,I7
T3	I1,I6
T4	I2,I3,I5,I4
T5	I1
T6	I3,I6,I7
T7	I1,I3,I7
T8	I2
T9	I1,I3,I7
T10	I3
T11	I1,I2
T12	I3,I7

Eliminating the Transactions in (Table 1.) by deleting transactions T5, T8, T10 means deleting all the records with only one sold item like (Table 2.).

Table 2. Table of Transactions After Deleting Some Useless Taransaction

Transaction(T)	**Sold Items(I)**
T1	I1,I3,I4,I7
T2	I3,I4,I7
T3	I1,I6
T4	I2,I3,I5,I4
T6	I3,I6,I7
T7	I1,I3,I7
T9	I1,I3,I7
T11	I1,I2
T12	I3,I7

After deleting (T5, T8, T10) then these transactions which contain only two Items with low frequent together will be deleted like T3 and T11 which contains (I1, I2) and (I1, I6) because these sets only one time accrue in then table unlike (I3, I7) which repeated three times, so the remaining sets are showed in Table 3.

Table 3. Table of Transactions After Deleting Some Useless Transaction

Transaction(T)	**Sold Items(I)**
T1	I1,I3,I4,I7
T2	I3,I4,I7
T4	I2,I3,I5,I4
T6	I3,I6,I7
T7	I1,I3,I7
T9	I1,I3,I7
T12	I3,I7

After executing the apriori algorithm before and after our technique with a groceries dataset which is consist of about 7k records with 10 items such as: (whole milk, other vegetables, rolls/buns, soda, yogurt, bottled water, root vegetables, tropical fruit, shopping bags, sausage), the results like the table below:

A. Test (1)

This is the result before preprocessing:

Min. support is 0.1 having 707 instances

Min. metric <confidence> is 0.3

cycle performed is 18 Cycles

large item sets which produced is:

Size of set of large item sets L(1) is 10

Size of set of large item sets L(2) is 1

Best rules found:

other vegetables=t 1903 ==> whole milk=t 736 conf:(0.5) lift:(1.09) lev:(0.01) [59] conv:(1.05)

B. Test (2)

This is the result after preprocessing:

Min. support is 0.1 having 398 instances

lower metric <confidence> is 0.3

performed cycles are 18 cycles

large item sets which produced is:

Size of set of large item sets L(1) is 10

Size of set of large item sets L(2) is 8

Best rules found:

1. root vegetables ==> whole milk=t 481 <conf:(0.52)> lift:(1.07) lev:(0.01) [33] conv:(1.07)
2. root vegetables=t 929 ==> other vegetables=t 466 <conf:(0.5)> lift:(1.3) lev:(0.03) [107] conv:(1.23)
3. yogurt=t 1138 ==> whole milk=t 551 <conf:(0.48)> lift:(1) lev:(0) [2] conv:(1)
4. other vegetables=t 1536 ==> whole milk=t 736 <conf:(0.48)> lift:(0.99) lev:(-0) [-4] conv:(0.99)
5. tropical fruit=t 882 ==> whole milk=t 416 <conf:(0.47)> lift:(0.98) lev:(-0) [-9] conv:(0.98)
6. rolls/buns=t 1360 ==> whole milk=t 557 <conf:(0.41)> lift:(0.85) lev:(-0.02) [-98] conv:(0.88)
7. whole milk=t 1918 ==> other vegetables=t 736 <conf:(0.38)> lift:(0.99) lev:(-0) [-4] conv:(1)
8. yogurt=t 1138 ==> other vegetables=t 427 <conf:(0.38)> lift:(0.97) lev:(-0) [-12] conv:(0.98)

9. rolls/buns=t 1360 ==> other vegetables=t 419 <conf:(0.31)> lift:(0.8) lev:(-0.03) [-105] conv:(0.89)
10. other vegetables=t 1536 ==> root vegetables=t 466 <conf:(0.3)> lift:(1.3) lev:(0.03) [107] conv:(1.1)

Table 4. Table of implementing apriori algorithm before and after the new technique

Apriori	Min. Support	No. of Item sets	Item sets
Old Method	0.3	1	other vegetables ➔ whole milk conf:(.39)
New Method	03	8	other vegetables ➔whole milk conf. (48),with 7 rules else.

Eliminating unnecessary records from the dataset has a good effect on the speed of algorithm two because as shown in (Chart 1.) the number of records was eliminated for about half of the records.

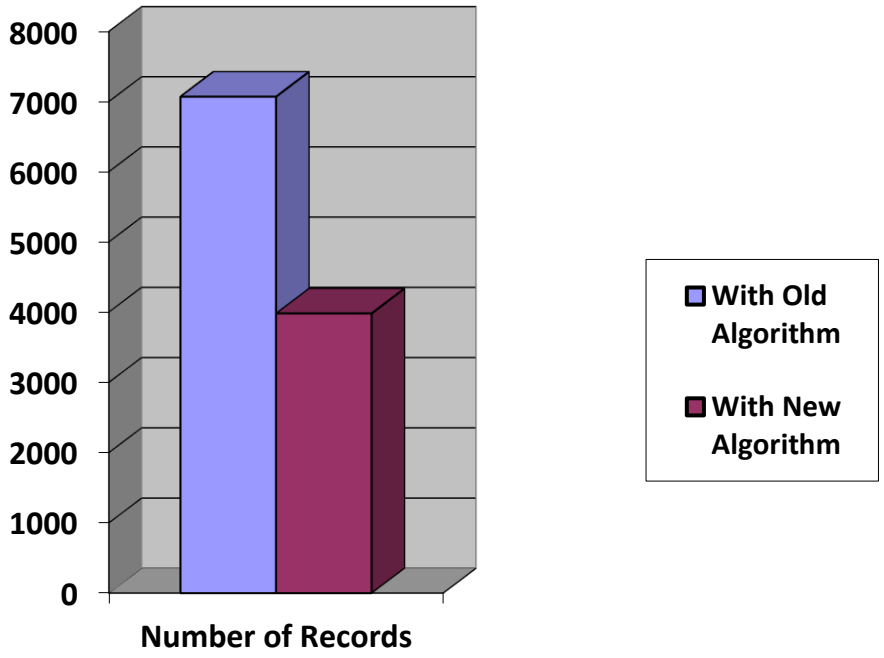


Chart 2. Number of The Records In Before And After Preprocessing

7. Conclusion

Using recommendation systems in online shopping is the best idea for improving the business and for customers too because it has a good effect on the customer which satisfy and finding all their requirement easily and for sealers like an online shop too which could sell more product at the same time

compared to before. so we can use the apriori algorithm as the best approach for finding the relationship between products that can sell together or relation between customers too, this algorithm consists of two stages like fining frequent itemsets and finding a relation between them, however using apriori algorithm for recommendation system has some limitations especially in these datasets contains a large number of itemsets and records, so these will be reasons to minimize the efficiency of the system by reducing the certainty of item sets in the generated rules and in other hands the system's time consuming for scanning a large number of the records but these can improve by using some techniques such as it discussed in the article by preprocessing on the dataset which is eliminating the number of records before applying the algorithm and it has good effect for the recommendation system.

Declaration of Competing Interest: The authors declare that they have no conflict of interest.

References

[1] Fayyad, U. M., Piatestky-Shapiro, G., Smyth, P. "From Data Mining to Knowledge Discovery: An Overview", AAAI Press / The MIT Press, pp. 1-34, 1996

[2] M. P. Robillard and R. J. Walker., 2014. An Introduction to Recommendation Systems in Software Engineering. In *Recommendation Systems in Software Engineering* (pp. 01 -11). Springer, Berlin, Heidelberg.

[3] Dhawan, S. and Singh, K., 2015. High rating recent preferences based recommendation system. *Procedia Computer Science*, *70*, pp.259-264.

[4] Resnick, P. and Varian, H.R., 1997. Recommender systems. *Communications of the ACM*, *40*(3), pp.56-58.

[5] Resnick, P. and Varian, H.R., 1997. Recommender systems. *Communications of the ACM*, *40*(3), pp.56-58.

[6] Dhawan, S. and Singh, K., 2015. High rating recent preferences based recommendation system. *Procedia Computer Science*, *70*, pp.259-264.

[7] Paranjape-Voditel, P. and Deshpande, U., 2013. A stock market portfolio recommender system based on association rule mining. *Applied Soft Computing*, *13*(2), pp.1055-1063.

[8] Zhang, Z. and Qian, S., 2012. The research of e-commerce recommendation system based on collaborative filtering technology. In *Advances in Computer Science and Information Engineering* (pp. 507-512). Springer, Berlin, Heidelberg.

[9] Wei, C.P., Shaw, M.J. and Easley, R.F., 2016. A survey of recommendation systems in electronic commerce. In *E-Service: new directions in theory and practice* (pp. 180-211). Routledge.

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

[11] Sarwar, B., Karypis, G., Konstan, J. and Riedl, J., 2000, October. Analysis of recommendation algorithms for e-commerce. In *Proceedings of the 2nd ACM Conference on Electronic Commerce* (pp. 158-167).

[12] Venkatesan, R. and Sabari, A., 2020. ISSUES IN VARIOUS RECOMMENDER SYSTEM IN E-COMMERCE–A SURVEY. *Journal of Critical Reviews*, *7*(7), pp.604-608.

[13] Linden, G., Smith, B. and York, J., 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, *7*(1), pp.76-80.

[14] McCarthy, J.F. and Anagnost, T.D., 1998, November. MusicFX: an arbiter of group preferences for computer supported collaborative workouts. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work* (pp. 363-372).

[15] Chao, D.L., Balthrop, J. and Forrest, S., 2005, November. Adaptive radio: achieving consensus using negative preferences. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work* (pp. 120-123).

[16] Ling, K., Beenen, G., Ludford, P., Wang, X., Chang, K., Li, X., Cosley, D., Frankowski, D., Terveen, L., Rashid, A.M. and Resnick, P., 2005. Using social psychology to motivate contributions to online communities. *Journal of Computer-Mediated Communication*, *10*(4), pp.00-00.

[17] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J., 1994, October. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work* (pp. 175-186).

[18] Guo, Y., Wang, M. and Li, X., 2017. Application of an improved Apriori algorithm in a mobile e-commerce recommendation system. *Industrial Management & Data Systems*.

[19] Agrawal, R. and Srikant, R., 1994, September. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).