International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Profitable Association Mining rules based on Casual Survey Approach Using Apriori Algorithm

Anjali Sharma[1], Neeraj Kumar Choudhary[2]

[1, 2]Computer Science and Engineering, Amity University, Noida, India

Abstract: ***In this technology specific era, everyone in industry look for smart methods to get profit in business. The biggest contribution is given by the introduction of Data Mining branch of Computer Science Engineering. The huge size of data is monitored, compressed and mined to get interesting facts. Different approaches are introduced day by day so as to formulate hidden associations between data. Following to which, new association rules are being generated. The base idea to find most occurring events was provided in Apriori Algorithm. In this paper we will analyse classical Apriori algorithm. We will introduce the concept of Casual Transaction database that includes direct survey from users i.e, how they trend to go for selections. We will propose an itea to integrate Casual transaction database with Real transaction database (what customer really does). We will analyse frequent itemset and association rules. We will then try to fit this approach with Apriori Algorithm. Accordingly we will keep this paper for our vision towards implementation of this approach and for our future work area.***

Keywords: Association rules, Data Mining, Apriori Algorithm, frequent itemset, Casual Survey Approach.

1. Introduction

We are dealing with enormous amount of data in different ways. This could be the case of reading complete books for exams, analysis of case study of the patients in order to diagnose disease, calculating daily productivity of a store, monthly task sheet of an employee and so on. However it's very tedious and time consuming to go around voluminous data every time when you want get some results out of it. To facilitate this, data mining approach came into light. It is very much similar like the traditional approaches of coal mining where, number of steps has to be gone through in order to get the finest diamond. Hence if we apply the same strategy in our technology, we do while writing in exams by answering with best points out of a complete book(as it is almost impossible to trace a complete book in exam) , we can make the use of pivot table, pie charts, graphs, etc. To diagnose a disease out of the case study of a patient, to get a better sight of a business the more frequent and less frequent buying products can be listed and special discounts can be made to increase productivity, in order to enhance an employee's productivity we can mine the areas where an employee is lacking. For all the above approaches there is a need to filter the data so as to get useful information that is called data mining. Once we have fine data, we can find relations between the data by observing the nature and behaviors' of the occurrence of events in data [3,4,8,9].The associations are represented in the form of rules. These rules provide the hidden interesting dependencies between different events. Such association between events in a database is called association rules [3,4,8,9].

To frame the concept of association rules, the basic classical Apriori Algorithm was introduced by R. Aggarwal and R. Srikant in 1993[9]. This algorithm provided a way to find frequent itemsets with the help of user defined support count and confidence. In this paper, we would like propose an idea of finding efficient association rules by introducing two types of transactional databases. 1) Casual transaction database. 2) Real transaction database. We will integrate these two transaction tables and then we will make the use of Apriori Algorithm to compute association rules. We will analyze the real transaction database table in Apriori Algorithm. Then we will set forth an ideato compare the results of both the approaches in order to analyze the better approach out of them. Hence the sequence of our review paper will include the description of Classical Apriori Algorithm in next section. Then we will discuss new proposal. We will review it along with Apriori Algorithm. Followed to this, in our next section we will present our idea of future insight to this paper. The last section concludes the paper.

2. Classical Apriori Algorithm with Association Rules

Consider a set of items I, These items are the entries of transaction database D. Each transaction is considered as T with Transaction id TID in such a way that $T \sqsubseteq I$

(1) Definition 1: Let I={I1,I2,I3,… , In} is an itemset, then D={<TID, T>| TID$\sqsubseteq$I} [2]
(2) Definition 2: Let $X \sqsubseteq I$, $Y \sqsubseteq I$, and $X \cap Y = \emptyset$, then $X \rightarrow Y$ is an association rule [2].
(3) If in transaction database c is the percentage of transaction where A and B itemset are available such that A contains B, that means the rule $X \rightarrow Y$ has confidence c in D. Similarly if s is the percentage of transaction in D that contains both $A \cup B$, then $X \rightarrow Y$ has support s in D.
(4) Minsupp and minconf is the term that represents minimum support and minimum confidence percentage set by the user based on profit or gain of the business.

From the Classical Apriori Algorithm [8] , we find:
(1) Frequent itemsets if support (X) $\geqslant$ minsupp then X is a frequent itemset.
(2) Well off association rules from frequent itemsets. If itemset $B \subset A$ and $B \neq \emptyset$ and support (A)/support (B)$\geqslant$ minconf(A-B).

Apriori Algorithm is used to mine association rules with frequent itemsets. It is a level wise search [6] where k

itemsets searches for k+1 itemsets. It makes the use of Apriori property i.e. The subsets of frequent itemsets should be non- empty and these subsets are also frequent. It is a two step process [6]:-

(1) **The Join Step:** Count of each and every itemsets is kept in a transaction database table; we call it candidate k itemsets. Lk is created by combining Lk-1 with itself in such a way that itemsets are arranged in lexicographic order [Han]. And Lk-1 itemsets are joinable if and only if Lk-1 of first item of an itemset > Lk-1 of second item of itemset. This property keeps check on replicacy of itemsets.
(2) **The prune step:** Candidate itemset is always a superset of Lk itemsets. Hence Candidate itemsets can be large in numbers. In order to minimize we make the use of Apriori property described before.

Apriori Algorithm can be used in business analysis in order to improve productivity. Consider a scenario of a food store. Where different types of customers visit frequently. Out of which some are foodie in nature. Such types of customers buy the food items that they would love to eat. They do not care for money. There are customers who are foodie but they cannot spend freely on food items. Most of the customers belong to this category. And the third type of customers trend to have the food item which would come in bulk and less in price. If we find a transaction statement of all types of customers and build a transaction database table. We can estimate few food items that all three categories of customers usually buy. To get a firm idea about the buying behavior of the customer we can efficiently make the use of Apriori Algorithm. Once we will apply this algorithm, we will get few association rules that depict about the food items that are in demand. We can get to know about the associations of different food items that are fairly in demand. We can make few generalizations over such items. We can apply some business formula in order to enhance productivity of such food items. And there are few food items that rae purchased rarely. We can enhance productivity of such food items by changing or editing some flavors in it. Apart we can put few food items in combinations. Customers can be attracted by choosing different types of strategies. Thus Apriori Algorithm helps in mining association rules between different food item sets. We can get to know about the frequently buying item sets. We can find confidence and support. Food Itemsets with high confidence are more frequently buying patterns and in demand. Item Sets near to the high confidence can be brought in high confidence by applying few business strategies on it. It helps to improve the quality of the business. For finding association rules by using Apriori Algorithm, we have WEKA tool that is easily available online. We can even analyze the outliers in the food items that are making great no factors to the customers. It is a user friendly tool.

3. Casual Approach with Apriori Algorithm

Casual Survey Approach can be defined as the way to collect response of some short or close ended questions from public. The source would be the text messages wherein we can send text messages to number of individuals. We can do this survey with the help of whatsapp application (via internet) and we can send emails. The source of the survey is pretty inexpensive. In surveys like this the research may paralyze with biased results. However in order make the research personalize and impartial, Casual Survey will be done with one on one basis. To send the survey to individuals on the phone numbers of clients located in different regions. With the help of Whatsapp, Line, etc we can fire survey to different audiences. We can make the survey more efficient by personalizing the results and ensuring this fact to the end users. In this approach, we will consider two transaction tables:- casual transaction table T1 and formal real database transaction table T2.

Casual transaction table takes the item sets from the Casual Survey. In this transaction table, transactions will be taken from different customers on the basis of few questions via text messages, on whatsapp, casually. By doing this, we can have better idea about the customers preference. We will integrate the casual transaction table with the real database transaction table (transactions will be based on real transaction values collected from the billing of a store). We will then analyze the result by applying Apriori Algorithm on our integrated table. Later on, the results that we will get from individual casual transaction table and real database transaction table by using Apriori Algorithm will be compared with the integrated table. And with the outcomes we will mine various association rules. Based on the interesting facts if we get some outcome out of customer and producer both, then there are the chances of best productivity. Therefore in this review paper we would like to propose our above stated idea so as to mine best association rules that would seem profitable to both the parties. In this way we will have three different findings. First we will consider transaction table T1 that we will get from Casual Survey. We will apply Apriori Algorithm in order to get association rules with best confidence. In the second part we will work on Real database transaction table T2. We will apply Apriori Algorithm over this table so as to get strong association rules with high confidence. Then in the final part, we will integrate the transactions of both the tables and then we will apply Apriori Algorithm on the integrated database table so as to get strong association rules. We will compare the results of the entire three tables. We will analyze Casual Survey Approach thereafter. For the analysis of this approach, we will use Weka data mining tool . It is one of the most user friendly tools. With the help of Weka the business system can identify the areas where they are losing money and where exactly they need to work upon in order to get profit. The availability of the tool is pretty much easy on Google. Below is the schematic diagram for Casual Survey Approach Fig. 1.

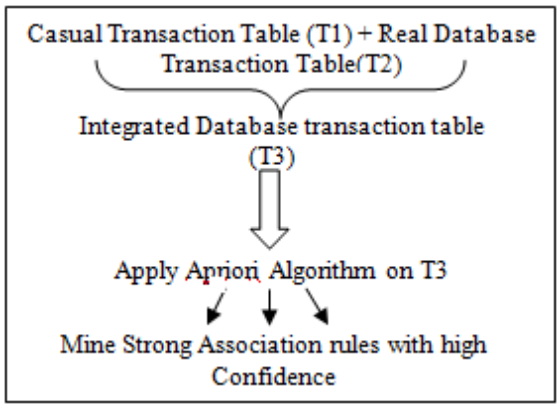

Figure 1: Casual Survey Approach

4. Future Work Aspects

In our next paper, we will provide implementation to our proposal. Based on the idea of our research, we will work on various data collection from casual survey and real transaction database. We will mine association rules and we will try association rules strategy on a store. We will then analyze at the end of a month if this really a benefit to both the customers and the producers. The analysis will be done by using WEKA tool. We will analyze in terms of time complexity that is how our search varies in all the three transaction tables with respect to time. We will try to get an efficient approach in terms of space complexity.

5. Conclusions

In this review paper we have discussed basic concepts of data mining and association rules. We then discussed Apriori Algorithm in order to get frequent itemsets and association mining rules. Based on the Apriori Algorithm we put forward our concept of Casual Survey Approach in an insight to mine profitable association mining rules. We have discussed pros and cons of Casual Survey Approach. We have then provided a brief overview of Weka tool that we will use to mine association rules for all three transaction tables. We have used Fig. 1. Casual Survey Approach above, so as to explain the approach in more clear form with the help of the diagram.

References

[1] A.B.M. Rezbaul Islam and Tae-Sun Chung, "An improved Frequent Pattern Tree Based Association Rule Mining Technique"., published in IEEE 2011.
[2] Aggarwal R and Srikant R. Fast algorithms for mining association rules.In Proc. 1994 Int. Conf. Very Large Data Bases, Santigo, Chile.487-499, Step.1994.
[3] Arun K pujai "Data Mining techniques". University Press IIndia) Pvt. Ltd.2001.
[4] Ashok Savasere, E. Omieeinski and Shamkant Navathe"An efficient Algorithm for mining Association Rules in large databases". Proceedings of the 21st VLDB conference Zurich, Switzerland,, 1995.
[5] Huiying Wang and Xiangwei Liu., "The research of Improved Association Rules Mining Apriori Algorithm"., In proceedings of the 2011 Eight international Conference on Fuzzy Systems and Knowledge Discovery(FSKD) Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
[6] Jiawei Han and Micheline Kamber, "University of Illinois at Urbane Champaign in second editions 2006 Morgan Kaufmann Publisher.
[7] Liqiang Geng, Howard J. Hamilton. "Interestingness measures for data mining: A survey." ACM Computing Surveys, 2006,pp, 1-32.
[8] R. Agarwal, T. Imienlinski, and A. Swami.. "Mining association rules between sets of items in large databases". In proceedings of the 1993 ACM SIGMOID International Conference on Management of Data, pages 207-216, Washington, DC, May 26-281993.
[9] R. Aggarwal and R. srikant "For Mining Association rules "Proceedings of VLDB conference, pp487- 449, Santigo Chile, 1994.
[10] Ruowu Zhong and Huiping Wang, "Research of Commonly Used Association Rules Mining Algorithm in Data Mining"., presented in 2011 International Conference on Internet Computing and Information Services.
[11] Xiufeng Piao, Zhanlong Wang and Gang Liu, " Research on Mining Positive and Negative Association Rules Based on Dual Confidence"., presented in 2011 fifth International Conference on Internet Computing for Science And Engineering.
[12] Zhun Zhou, Bingru Yang, Yunfeng Zhao, Wei Hou.,"Research on Algorithms for Associations Rules Mining based on FP-tree". Published in IEEE 2008

Author Profile

Anjali Sharma is pursuing Master of Technology in Computer Science and Engineering from Amity University, Noida. She has done Bachelors of Technology in Computer Engineering from Rajasthan Technical University, Kota. Her area of research is in Data Mining and Computer Networks.