BIG DATA AND MARKET BASKET ANALYSIS: A SURVEY

Mohan Rao Mamdikar, V.E.C. Lakhanpur, SargujaUniversity,Ambikapur (C.G.), India
(mohanrao.mamdikar@gmail.com)
Amit Kumar Dewangan, CVRU, Bilaspur (C.G.), India (amit.nitrr@gmail.com)

ABSTRACT

Knowledge discovery in databases (KDD) is an area whose aim is to extract good knowledge from a huge amount of data. Many applications have applied knowledge discovery especially for the company or any applications that use large database in their company activities. This is because the knowledge discovery is very important to extract the useful knowledge or structured data to improve marketing strategy. Big data is repository which is collection of very huge data sets with several types of data so that it is difficult to process normally. Analyzing the behavior of customer and decision- making on such data became a huge and challenging problem for organization mainly for retailers in a current competitive Market to survive in the market. Technological inventions such as data mining tool andHadoop, Map Reduce (Presented by Google) became an important tool for analyzing the huge amount of data and making decision correctly. Using these tool future trends will be evaluated by analyzing the huge data. Market basket analysis is a powerful tool to do so. In this paper, our main aim would be trying to analyze the huge amount of data and find what would be future behavior of customer and make the correct decision to survive in competitive market. Analysis will be done with the help of association rule using Market basket analysis to evaluate the future market trends or patterns.

KEYWORDS:KDD,BigData,Hadoop,HDFS, Map/Reduce, Frequent itemset, data mining, association rule, Brute-Force method, market basket analysis.

INTRODUCTION

Big data is beyond the Data ware house, with a study 90% of the data on the planet has been made in recent years. Big data is collection of huge diverse data such as sensors from atmosphere data, posts from social media (Facebook, twitter and Whatsappetc), images, and videos, audios and GPS signals from mobile phones. These types of structured, semi-structured and unstructured data may in petabytes (1,024 terabyte) or even more Exabyte's (1,024 petabytes).

Data mining has attracted a great deal of attention in the database industry and in society as whole recent years. The information and good knowledge extracted by using Data mining can be used in various applications such as market analysis, fraud detection, customer retention, and to market control etc.In this paper, we have discussed an association rule to find out customer behavior using market basket analysis. Latter on we will conclude that how market basket analysis will be used in Big data. There are already many algorithms proposed but our main concern with association rule and Brute-Force method.

In Section 2, we discuss about big data, challenges and applications. In section 3 we define Hadoop. In section 4, we discuss about Map reduce framework. In section 5, we define the data mining and some pre-processing techniques. In section 6, we define association rule. In section 7, we define market basket analysis and its applications. Finally, in section 8, conclusion points.

.

BIG DATA

Big data is an abstract concept.It does not have exact definition. Bid data is really "big data" or "massive data" which cannot be processed with traditional techniques.

The current international population exceeds 7.6 billion and 3.96 billion of these people are using internet. Even more 5 billion individuals are using mobile devices. According to Mckinsey (2013) As a result these millions of people are generating massive amount of data. Sensors continuously generates enormous of structured, unstructured and semi-structured data. Such type of data can managed in petabytes or in Exabyte's or furthermore. This type of data is referred as "Big Data".

Nowadays huge amount of data is collected from internet (Facebook, Google, Twitter etc). Big data related to the services of internet companies growing rapidly and generating massive amount of structured, unstructured and semi-structured data daily. Google itself processes data of hundreds of petabytes, Facebook generates enormous amount of log data of over 1PB/month. There are many such big data related companies which are generating in and processing such a "Big Data".

Big data is produced from different devices and applications over the world-

i. Black Box data- Black box is the essential part of airplane, aircraft, helicopter and jet plane which captures every moment's data such as voice of flight crew, recordings of microphones, earphones and performance information of aircraft.
ii. Social Media Data- Social media such Facebook, twitter, Whatsapp holds massive amount of data in the form of Audio, video, images and GIF. The views posted by the millions of people over the world.
iii. Power Grid data- Power grid holds the information how much power is consumed by a particular node.
iv. Search Engine data- Search engine such as Google retrieves large amount of data from different database.

There are many fields which are generating and processes huge data such as healthcare, transport data, other electronic devices. Therefore we could say that big data is collected from different sources.

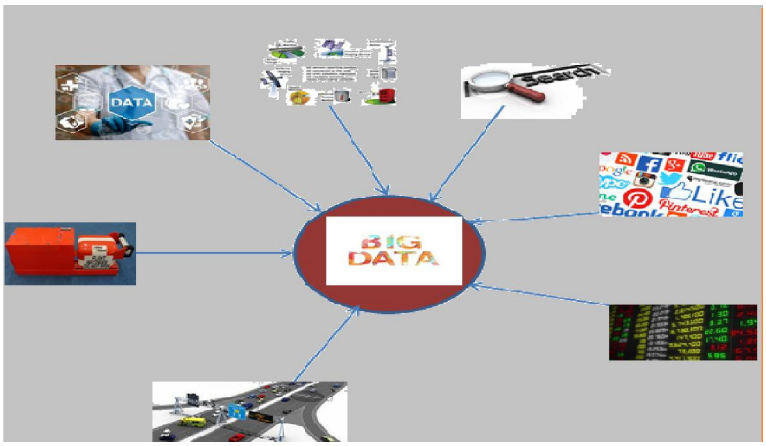

Figure 1: Big data sources

There are main characteristics of big data – As a matter of fact, big data has been characterized into 5Vs model, where 5Vs namely volume, velocity, variety, Veracity and Value.

a) Volume –Presently big number of systems are generating huge amount of data constantly, which growing petabytes or Exabyte's or thousands of Exabyte's.
b) Velocity- Data is increasing at rapid speed which should be collected and analyse for maximum commercial utilization timely.
c) Variety- Variety indicates the different types of structure, unstructured and semi structure data.
d) Veracity- The quality of data which is captured can vary constantly affects the accurate analysis.
e) Value- Value is derived from big data, to convert big data into value to access big data. If big data is not converted into value then it is useless.

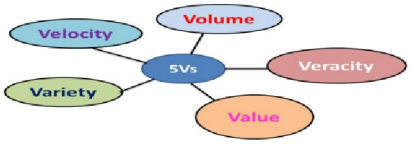

Figure 2. 5Vs of Big Data

Application of big datais in the field of marketing, Industry, Business, healthcare, fraud detection, market prediction etc.

HADOOP

There are various technologies in the market from different companies such as Google, IBM, AMAGON, Microsoft are to handlewith big Data.

Dough cutting, Mike cafarella and their team started a open source project called Hadoop in 2005 by Google.

Hadoop provides an efficient framework for running jobs on multiple clusters. It is not only storage system but is a platform for data storage as well as data processing. There are certain characteristics of Hadoop-

Scalable- More nodes can be added, can increase size of memory, processors.

Fault Tolerant-Even if nodes failed or go down, data can be processed by another node.

Reliable- Hadoop can handle if any type of software/Hardware failure.

Distributed- Data is replicated among the other nodes, if any type of failure by the particular node it can be processed by another node.

Hadoop works as data storage and as mining. It can mine any types of data such as structure, unstructured and semi structured data.

Hadoop architecture, works in master-slave manner. There is one master node which manages, maintains and monitors the slave nodes. On the other hand slave nodes are actual worker nodes.

Master nodes always store the data about data but slaves nodes stores data only. Data is stored in hadoop in cluster manner distributedly.

The main Component of Hadoop is HDFS. As we know that was Hadoop motivated by Google's Map/Reduce. Dough Cutting named the Hadoop system after his youngster full toy elephants. This Hadoop has a component called hadoop distributed file system (HDFS) providesstorage. HDFS is highly reliable, scalable, fault tolerant and distributed.

MAP REDUCE

Map/reduce is a processing layer of Hadoop.It is designed for processing large volumes of data in parallel.

Map/reduce programming which is written on functional programming paradigm.Haddop is powerful and very efficient due to Map/reduce by which parallel processing is performed. The map/reduce is originallyreferred to the proprietary Google technology.

Map/reduce mainly composed of following operations-

- Map- Each work node applies the map function to the local data, and writes the output to a temporary storage. A master node ensures only replica of redundant output.
- Shuffle- Worker nodes distribute data based on the output keys.
- Reduce-Worker nodes now process each group of output data/per key in parallel.

ASSOCIATION RULE

Frequent patterns or frequent itemsets are patterns that appear in a data set frequently. For example, a set of items, such as milk and bread that appear frequently together in a transaction data, set is a frequent itemset. A subsequence, such as buying first a PC, then a mobile, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern. Finding such frequent patterns plays an important role in data mining association, and correlations. This helps data classification, clustering and other data mining tasks. One major field of data mining is the problem of deriving association rule among the data.In this problem set of items and large collection of transactions area given. Which is the subset of this items.The task is to find relationship among the various items within the transaction.

Association rule is very popular procedure for finding interesting relations between various variable in large database. As introduce by RakeshAgrawal et al in the year 1994. We are discussing popular example of Association rule that is market basket analysis in the next section. An association rule is an expression form $A \Rightarrow B$, where A and B are frequent itemsets in the transactional database say $T=\{t_1,t_2,t_3 \ldots \ldots t_n\}$and $A \cap B = \emptyset$ The rule $A \Rightarrow B$ is translated as "if itemset$A$ occurs in a transaction T , the itemset$B$ will also occur in the same transaction T ". Through such kind of knowledge, retailers or market analyst can place two itemsets that is A and B closely so that encourage of sale of two itemsets together based on association/correlation rules.

MARKET BASKET ANALYSIS

Typical example of association rule is none other than the Market Basket Analysis. This process analyzes the customer purchasing habits by discovering associations between different itemsets that customer place his/her "shopping basket. Finding such association among different items may help retailers or market analysts to improve marketing strategy to increase the sale of items and to provide discount offer by analyzing frequently purchased items by customers. If we look at of different supermarkets' layouts like Big Bazaar, Food Bazaar, and many other markets, based on market basket analysis only.

In supermarkets, an analyst or domain expert may learn many things about buying behaviors of customer "Which itemsets are likely to be purchased by customers on a particular trip from store?" By using market basket analysis, we can answer this question by performing association rule on the retail data of customer purchasing transaction at supermarket or store. Based upon this, analysis is used to plan marketing and discount strategies. In this, items that are frequently purchased by customer are placed close together in order to encourage the customer to further sale of such items. If we think of the universe as the set of items available at store, then each item in the store has a Boolean variable (representing the presence or absence of that item). by the Boolean vectors of values assigned to these variables. The Boolean vectors can be analyzed for buying habits that reflect the items which are frequently purchased together. These patterns or habits can be represented using association rule. This is illustrated using an example, the knowledge or information that customer who purchase laptop also buy an anti-virus software at the same time, which is represented in Association Rule as follows:

$$Laptop \Rightarrow Anti\text{-}virus\ software$$
$$[Support = 30\%;\ confidence = 70\%] \quad (3)$$

Remark 1. The conditional probability p is sometimes referred to as the 'accuracy' or 'confidence' of the association rule.

The support and confidence are two measures of rule interestingness. They reflect the usefulness and certainty of discovered rules respectively. The support of 30% of association rule above means that 30% of all the transactions under analysis show that laptop and antivirus software are purchased together. A confidence of 70% means that 70% of the customers who purchased a laptop also bought the antivirus software. Association rules are assumed to be interesting if they satisfy both user defined minimum support threshold and minimum confidence threshold. If set of frequent k-itemsets commonly denoted as L_k, we can have from equation 2

$$Confidence(A \Rightarrow B) = p(B/A)$$

$$\frac{support(AUB)}{support(A)}$$

$$\frac{support_{count}(AUB)}{support_{count}(A)}$$

Now we will take an example and perform association rule using market basket analysis to derive association between data sets. We will take transactional record as input for analysis and output of analysis will be now information directly derived from stored transactional data using association rule. We have collected data from market and transaction data is shown in table 1.

Table 1: Transaction databases

Transactional ID	***Items purchased by customer***
1	*sugar,milk,tea,bread*
2	*sugar, bread*
3	*milk, bread*
4	*bread,milk,tea*
5	*milk, bread*
6	*sugar, milk*
7	*sugar,tea,bread*

Let us take the items currently seen by the customer as*X*(independent variable) and other items will be associated with those currently seen items say *Y* (dependent variable). If we have only two items say *A* and *B*, then we have only two possible rulessuch as $A \Rightarrow B$ and $B \Rightarrow A$. If we have only three items say *A, B*, and *C*, then we have 12 possible rules.Similarly, if we have four items then we have total 50 possible association rules and for 5 items there are 180 possible rules. Therefore we can derive this relation in general for n number of items, there are $R = 3^n - 2^{n+1} + 1$ possible rules .This is computational prohibitive because it is exponential to theItems (*n*). It is obvious that if we have thousand number of item then difficult to compute rules as shown in table 2

Table 2: Relation between R and n

n	1	2	3	4	6	10	100	500
R	0	2	12	50	602	57002	5.15378E+47	3.636E+238

Now we will convert our transaction database given in table V into binary data, and then apply Brute-force Method of association rule. Binary data as given below:

Table 3: Combination of Association rules

Transactional ID	*Items Purchased by customer*	*A*	*O*	*G*	*P*
1	*sugar,milk,tea,bread*	1	1	1	1
2	*sugar, bread*	1	0	0	1
3	*milk, bread*	0	1	0	1
4	*bread,milk,tea*	0	1	1	1
5	*milk, bread*	0	1	0	1
6	*sugar, milk*	1	1	0	0
7	*sugar,tea,bread*	1	0	1	1
SUM		4	5	3	6

For more readability and understandability, we call the items by first letter (*S* for *Sugar*, *M* for *Milk*, *T* for *Tea* and *B* for *Bread*).Let us rename this binary transaction record table into *B*in Record named as *S*_, M_, *T*_, and *B*_ . We also set an array of 1, if items in the row match with same item in the column, otherwise 0, to help computation of support and confidence respectively and we named it as BiOne.

The support can be computed by using formula given below.

$$support = \frac{n(X \Rightarrow Y)}{N}$$

WhereN is number of transactions and support is the ratio of support count and the number of transactions. Similarly confidence also computed simply by taking ratio between support counts of union of the independent variable and support count of dependent variable. It formulated as follows.

$$confidence = \frac{n(X \Rightarrow Y)}{n(X)}$$

Now compute association rules of market basket analysis based on binary transaction record and set two important threshold criteria those are minimum support and minimum confidence. For example we set 40% minimum support and 75% minimum confidence. Minimum support =40%, Minimum confidence = 75%. We can obtain 3 association rules based on user defined thresholds, minimum support= 40% and minimum confidence =75% as given in table 4.

Table 4. Computation of support and confidence

NM.	*X*	⇒ *Y*	*n(XUY)*	*N*	*%*	*n(X)*	*%CMnf.*	*I in Rule?*
1	*S*	*M*	2	7	29%	4	50%	×
2	*S*	*T*	2	7	29%	4	50%	×
3	*S*	*B*	3	7	43%	4	75%	✓
4	*S*	*MT*	1	7	14%	4	25%	×
5	*S*	*TB*	2	7	29%	4	50%	×
6	*S*	*MB*	1	7	14%	4	25%	×
7	*S*	*MTB*	1	7	14%	5	25%	×
8	*M*	*S*	2	7	29%	5	40%	×
9	*M*	*T*	2	7	29%	5	40%	×
10	*M*	*B*	4	7	57%	5	80%	✓
11	*M*	*ST*	1	7	14%	5	20%	×
12	*M*	*TB*	2	7	29%	5	40%	×
13	*M*	*SB*	1	7	14%	5	20%	×
14	*M*	*STB*	1	7	14%	3	20%	×
15	*T*	*S*	2	7	29%	3	67%	×
16	*T*	*M*	2	7	29%	3	67%	×
17	*T*	*B*	3	7	43%	3	100%	✓
18	*T*	*SM*	1	7	24%	3	33%	×
19	*T*	*MB*	2	7	29%	3	67%	×
20	*T*	*SB*	2	7	29%	3	67%	×
21	*T*	*SMB*	1	7	14%	3	33%	×
22	*B*	*S*	3	7	57%	6	50%	×
23	*B*	*M*	4	7	57%	6	67%	×
24	*B*	*T*	3	7	43%	6	50%	×
25	*B*	*SM*	1	7	14%	6	17%	×
26	*B*	*MT*	2	7	29%	6	33%	×
27	*B*	*ST*	2	7	29%	6	33%	×
28	*B*	*SMT*	1	7	14%	9	11%	×
29	*SM*	*T*	1	7	14%	9	11%	×
30	*SM*	*B*	1	7	14%	9	11%	×
31	*SM*	*TB*	1	7	14%	9	11%	×
32	*ST*	*M*	1	7	14%	7	14%	×
33	*ST*	*B*	2	7	29%	7	29%	×
34	*ST*	*MB*	1	7	14%	7	14%	×
35	*SB*	*M*	1	7	14%	10	10%	×
36	*SB*	*T*	2	7	29%	10	20%	×
37	*SB*	*MT*	1	7	14%	10	6%	×
38	*MT*	*S*	1	7	145	8	13%	×
39	*MT*	*B*	2	7	29%	8	25%	×
40	*MT*	*SB*	1	7	43%	8	13%	×
41	*MB*	*S*	1	7	14%	11	9%	×
42	*MB*	*T*	2	7	14%	11	18%	×
43	*MB*	*ST*	1	7	14%	11	9%	×
44	*TB*	*M*	2	7	29%	9	22%	×
45	*TB*	*M*	2	7	29%	9	22%	×

46	*TB*	*SM*	1	7	14%	9	11%	×
47	*SM*	*TB*	1	7	14%	12	8%	×
48	*SM*	*BT*	1	7	14%	15	7%	×
49	*ST*	*BM*	1	7	14%	13	8%	×
50	*MT*	*BS*	1	7	14%	14	7%	×

From the above table we can provide threshold value to the support and confidence to get association rules. Here 40% support means the purchasing items *X* and *K* together and 75% of minimum confidence means the occurrence of *X* and *K*of the transaction if customer buying *X* then also buy item *K*.
From the table V, it is obvious that only transaction ids 3, 10 and 17 satisfy the minimum support threshold and minimum confidence threshold, due to higher value of support and confidence than user define threshold value (40%, 75%). By observation, these items or products are placed together and these items are purchased by customers certainly.
To achieve this analysis, we have Brute-Force method, defined as follows:

```
Input   :  Transaction  record
output  :  Set of  association  Rules
begin:mst\\minimum support threshold ;
mct\\minimum confidence threshold;
for (all data items)do
begin {
find all association rules}
end ;
 for (all association rules) do
   begin {
   find support and confidence
   if (support >=mst&& confidence >=mct) then
begin {save association rule}
   end if ;
  end for ;
end ;
```

CONCLUSION AND FUTURE WORK

From the above calculation and observation, it is found that data mining tools will be used for discovering new patterns associated with changing behaviors of various transactions. To find future trends or patterns we have used Brute-Force method. Using this method we have successfully performed associations among various products purchased by customers and they can be placed together to improve marketing strategies, by computing support and confidence. Same operation can be done for "Big Data" by which this kind of analysis extensively used in cross selling, product placement, and fraud detection, market prediction.
As we have seen that using Brute-Force Method is not much efficient with huge data items such as "Big data". Therefore in a broad market analysis, it is little more complicated to do analysis using this algorithm. So to get more efficient analysis of market we may use apriori algorithm.

REFERENCES

Philip chen C.L., Zhang C. Y. (2014). Data-intensive applications, Challenes, techniques and technologies: A survey on big Data. Information science,275 : 314-347.

Min chen, Shiwen Mao, Yunhao Liu "Big Data : A survey"MobilenetwAppl(2014) 19:171-209.

Nawsher khan, IbrarYaqoob, Ibrahim Abaker, TargioHashem, ZakiraInayat, WaleedKamaleldin Mahmoud Ali, Muhammad Alam, Muhammad Shiraz, Abdulla Gani "Big Data : Survey, Technology, Opportunities, and Challenges" , Hindawi publishing Corporation, The scientific world Journal, Volume 2014.

Worldometers, "Real time statistics," 2018, http://www.worldometers.info/world-population/

PSG Aruna Sri, Anusha M "Big Data-Survey" Indonesian journal of Electrical Engineering and Informatics (IJEEI), Volume 4, No.1 March 2016, pp-74-80.

Campos, M.M.; Stengard, P.J.; Milenova, B.L (2005.)"Machine Learning and Applications",Proceedings.Fourth International Conference.

Agrawal R. and Srikant (1994). Fast Algorithms for Mining Association Rules, 20th Int'l Conf. on Very Large Data Bases, Santiago, Chile.

Mamdikar M. R. (2013). Pandey S. and Kumar V. (2013), Evaluation of market trends using association rule by basket analysis, International Journal of Information & computational Technology, 3(1).

https://data-flair.training/blogs/hadoop-tutorial/

Agrawal, R., and Ramakrishnan S., JPillai J. (2011).User centric approach to item set utility mining in Market Basket Analysis, International Journal on Computer Science and Engineering (IJCSE).

Han J. and Kamber M. (2006). Data Mining: Concepts and Techniques, 2nd ed. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers.

Brin S., Motwani R. and SilvertrinBeyound C.(1997). Market Baskets: Generalizing association rules for correlation, SIGMOD Record, (ACM Special Interest Group on Management of Data), 26(2): 265.

Nan-chanHsich, Kuo-Chang cha(2009)enhancing consumer behavior analysis by data mining techniques.

Agrawal R., Imilienski T. and Swami A (1993)., Mining Associations Rules between Sets of Items in large databases. Proc. of the ALM SIGNOD. Intl conf. on management of Data, 207-216 .