International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 17 (2018) pp. 13272-13281
© Research India Publications. http://www.ripublication.com

Integrating Application with Algorithms of Association Rule used in Descriptive Data Modelling, through which Data Mining can be Implemented for Future Prediction

Sumita Mukherjee[1]
Associate Director
University Advancement & Internationalization Riara University P.O. Box 49940-00100 Nairobi, Kenya
Research Scholar, Manav Rachna University, Faridabad, Haryana, INDIA

Guide-Dr. Prinima Gupta
Associate Professor
Manav Rachna University, Faridabad
Co-Guide- Prof. Felix Musau
Riara University Nairobi Kenya

Abstract

Data mining is a most appropriate discipline which clubs up statistics, database technology, knowledge discovery, pattern recognition, machine learning, business, natural disaster and other areas. It interferes and integrates in such a manner which is the most suitable for identifying a valid, logical and understandable pattern to influence the expansion and productivity of an operation, profitability, increase in sales and retail measurements, prediction of natural disaster, less faulty production, desired human resources.

What is Data? A fact and figure What is knowledge? Data collected in such a manner that we know about the data. What is information? Flowing of proper data through communication. What is Statistics? Based on data, knowledge and information is a decision maker and systematic evidence based on intellectual science.

This Paper will give an idea of the importance and significance of data mining in a manner that the knowledge acquired by learning descriptive data mining can be easily applied on predictive action. It also targets to fragment the fascinating correlations,the maximum used arrangements and alliances among the products in product range in the proceeding database.Once the understanding and concept on descriptive data modelling is clear the application on predictive model becomes easy, productive and more appropriate to take decision positively to footprint the accomplishment of the overall system.

Objective: - This paper mainly discusses various types of Association Rule application so that various algorithms used for association rule can be easily understood. With the proper pictorial example, we can integrate the application and the need for algorithm to find the pattern and use the algorithm for future prediction. The main objective is to assess the association pattern and its impact on various business-like retail, customer handling, cross marketing, future prediction. Though, different objective measures define different association patterns with different properties and applications.

The first measure is the level of break apart between product in a product range which is bounded up with sales strategy, and the second is the objective measure that intends to discover unexpected rules and handling the same in the database.

INTRODUCTION

Data mining can be bifurcated into three groups a) Descriptive b) Predictive c) Perspective

Descriptive Data Mining: In simple language we collect a big data and then break into smaller parts and then take a view on the smaller part in an understandable format. It is easy to identify the most suitable pattern on a small data and also use the past format for future analysis. Thus, descriptive data mining gives an inner idea of what is been happening in the company for last few years. It avails previous information in an easily digestible format for the benefit of a wide business audience. This data mining does not concentrate much on cause and effect relationship. Most unusable methods for descriptive analysis are observations, surveys and case studies. One example of day to day life is going to a super market and buy only the things required by you or jotted on your shopping list. Other examples are Inventory system, Sales report, HR System etc. of any business.

The methods used for Descriptive Data mining are a) Association Rule b) Clustering

Association: This defines a data mining function that gives an idea of the likelihood of the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as association rules. Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories. The data mining process which includes Association rule which in turn, observe rules that may govern associations and causal objects between sets of items. So, in a given transaction with multiple items, it tries to find the rules that govern how or why such items are often bought together. For example, flowers, candle, gift, chocolates are many times bought together because a lot of people like to present flower bouquet with candle chocolate, gift together for birthday celebration.

Support and Confidence are the two primary coins of association rules. This concept identifies the relationships and rules generated by analyzing data for frequently used if/then patterns. Association rules are usually needed to satisfy a user-specified minimum support and a user -specified minimum confidence at the same time.

The main applications of Association rule mining:

Basket data analysis - is to analyze the association of purchased items in a single basket or single purchase as per the examples given below.

Meet Carolyne she has her own business of floral shop where she sells different kinds of flowers, flower vase, manure, seeds, fountains, candle, bouquets, gift items etc. Of course, she has many fixed customers, still she has noticed some customer feels a bit frustrated after shopping in her shop if they want to buy flowers as well as few more items like pots, candle, manure etc. She has realized that the way she has spread her items at the shop the customers either pick only one item and look for other items not only that the customers pick the varieties like rose and lily or complaint of varieties like tuberose, glandulous, chrysanthemum etc. She also noticed that the bouquets, candles and gift items are not moving well as they are not placed at proper place. She realizes that there is a need to reorganize the store by doing a research and survey for better sale by picking more than one item by customer at one go.

Market basket analysis is the process of dealing with combinations of items that are often purchased together in one transaction. So, if a customer buys one item, according to market basket analysis, they are certain to an extent likely to buy another item.

Items are the things that the customers are purchasing. For example, each item is a product that Carolyne has in her store. When there is a group of items, this is called a range of items.

Transactions are the groups of items that are purchased together. So, if a customer buys a candle and a bouquet all items occur together in one transaction.

The probability that a customer will purchase an item is known as support. In other words, if Caroline is able to identify popular products that many or most of her customers purchase, then she would have support of that item or item set. For example, if most of her consumers buy flowers and candles then both the flowers and candles will have great support.

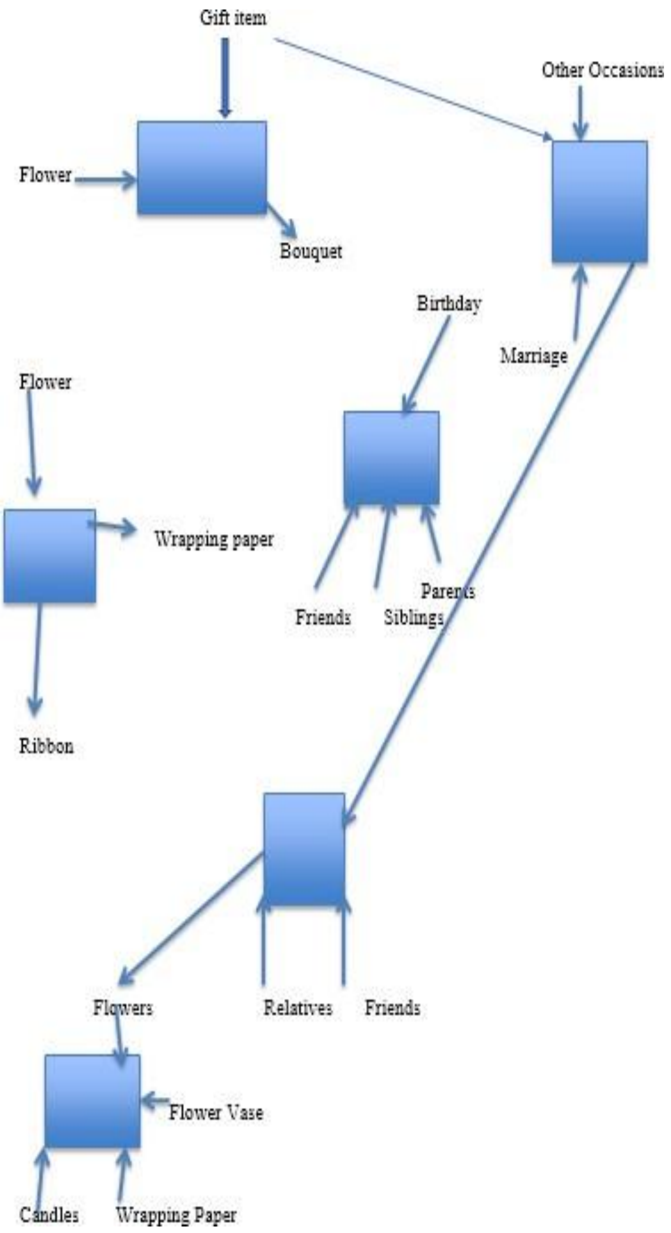


Figure 1. Represents Market Basket Analysis

Cross marketing - deals with promotion of one product with other as the products not only complement each other but also carry out other businesses that supplement your own business but not contenders. For example, Airtel, Videophone showrooms campaign for varieties of data plan for obvious reasons. Cross-promotion is a form of marketing promotion where customers of one product or service are targeted with promotion of a related product. A typical example is cross-media marketing of a brand University for example University promotion will include the website, different courses, student’s achievement etc.

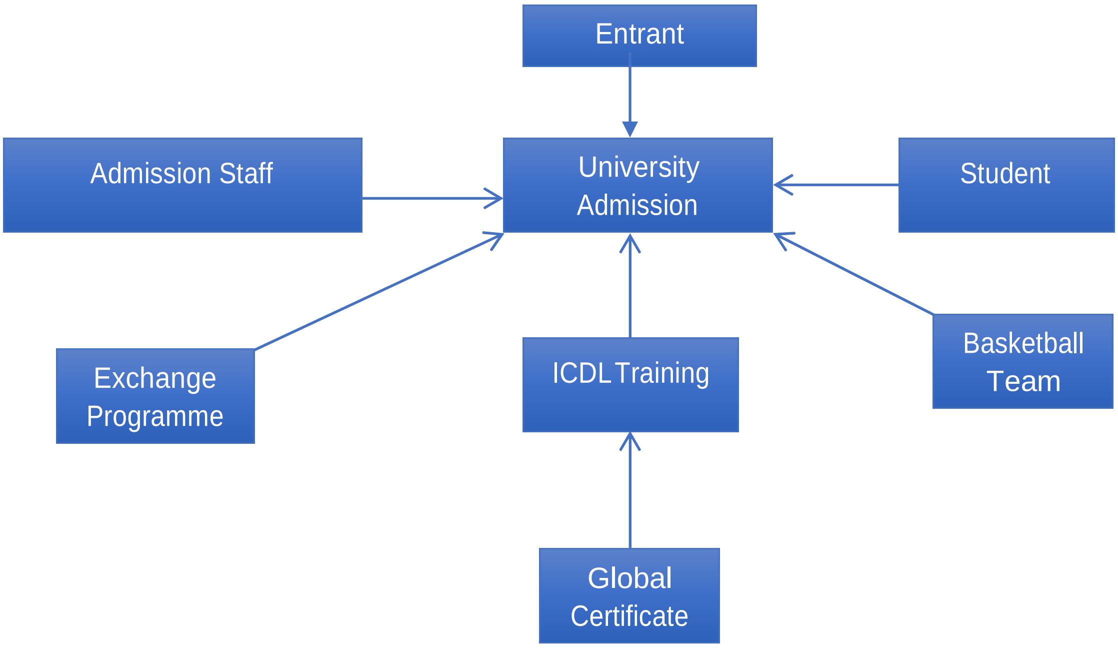


Figure 1.1 Represents Cross Marketing

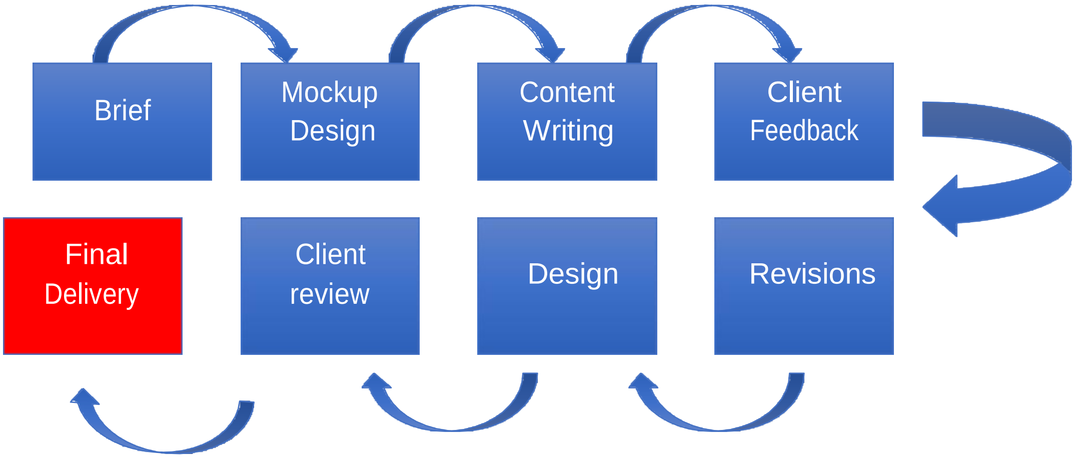


Figure 1.2 Represents Catalog Design

Catalog Design - the selection of items in a business' catalog are often designed to complement each other so that buying one item will lead to buying of another. For example, a car showroom will present a catalogue with different company cars, car accessories, engine oil etc. Thus these items are often complements or much related and not only useful for the seller but also to customer.

Loss leader Analysis :This process includes bringing more number of customer i.e it concentrates more on foot falls of customer rather than making profit. For example, a book seller will like to get more books sold having less profit rather than making more profit by selling few books.

POLITICAL • Governance • Accountability • State subordination	**FICTIONAL** • Plot • Character • Conflict • Secrecy
RELIGIOUS • Vision • Mission • Salvation • Priest	**ADVENTURE** • Authenticity • Purpose • Inspiration • Focus

Figure 1.3 Represesnts Loss Leader Analysis

Some Important Terms:-

Product Range: A group of one or more products. For example, from the above table {Flowers, Candles}

An item set is known P- product range if it contains P items. Support count {} Number of times product range occurs i.e. the number of occurrences {Gift items, Vase} = 3 {Flowers, Candles} =2

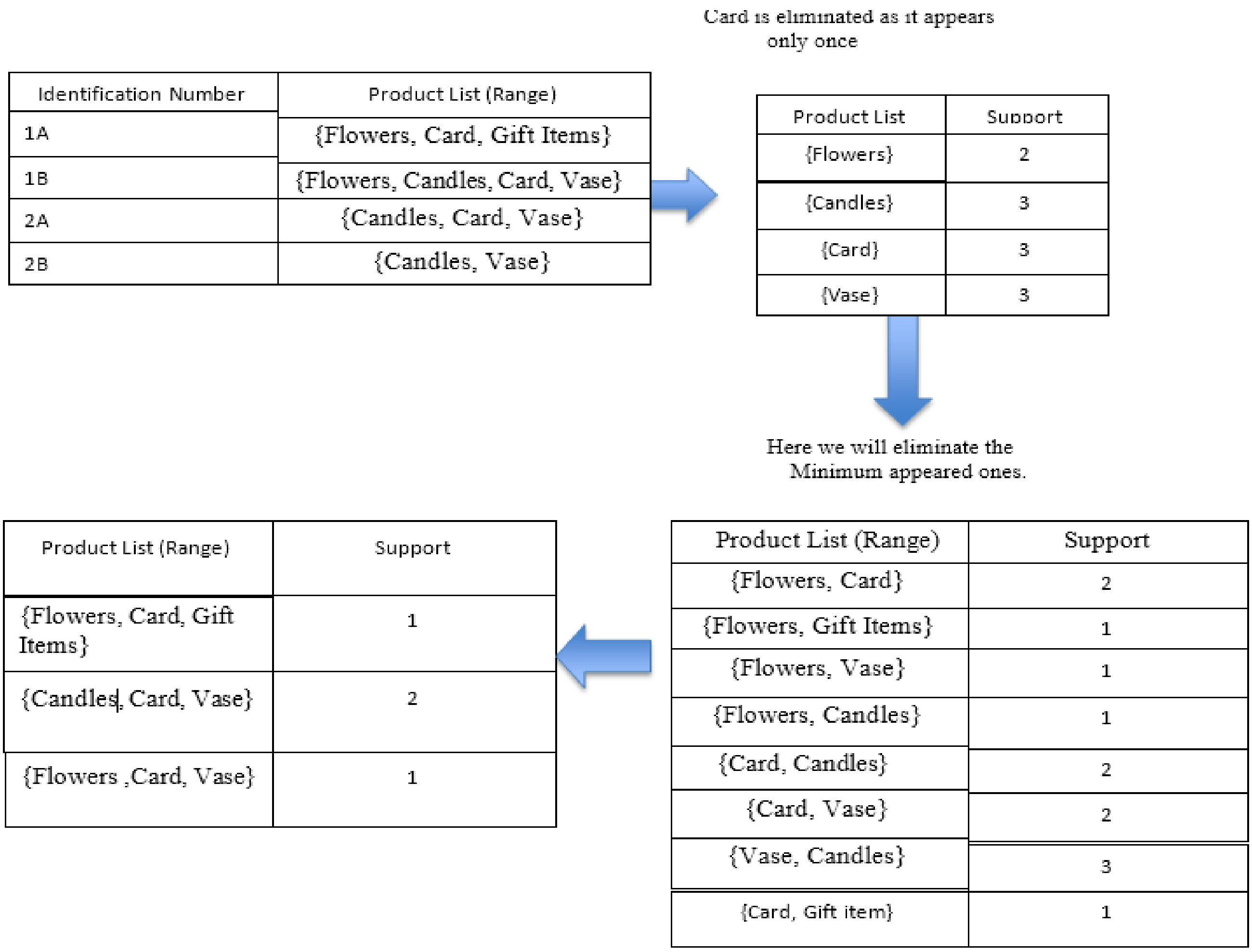
Support: And product range that contains the fraction of data items

For example, Support {Candles, Gift items} =2/4=

Support: -Support (sometimes referred as frequency) is simply a probability that a randomly chosen transaction T in data base contains both items sets Flowers and Candles.

Mathematically, support (Flowers ⇒ Candles) T = P (Flowers ⊂ T ∧ Candles ⊂ T) = total # of transactions # of transactions containing both Flowers and Candles pacifies a simplified notation that support (Flowers ⇒ Candles) = P (Flowers ∧ Candles)

Confidence: - Confidence (sometimes called accuracy) is simply a probability that an item set Candle is purchased in a randomly chosen transaction T in Database given that the item set Flowers are also bought. Mathematically, confidence (Flowers ⇒Candles) transaction T= P (Candles ⊂ T | Flowers⊂ T) total # of transactions containing Flowers # of transactions T database containing both Flowers and Candles shall use a simplified notation that confidence (Flowers ⇒ Candles) = P (Candles | Flowers.



Card is eliminated as it appears only once

Identification Number	Product List (Range)
1A	{Flowers, Card, Gift Items}
1B	{Flowers, Candles, Card, Vase}
2A	{Candles, Card, Vase}
2B	{Candles, Vase}

Product List	Support
{Flowers}	2
{Candles}	3
{Card}	3
{Vase}	3

Here we will eliminate the Minimum appeared ones.

Product List (Range)	Support
{Flowers, Card}	2
{Flowers, Gift Items}	1
{Flowers, Vase}	1
{Flowers, Candles}	1
{Card, Candles}	2
{Card, Vase}	2
{Vase, Candles}	3
{Card, Gift item}	1

Product List (Range)	Support
{Flowers, Card, Gift Items}	1
{Candles, Card, Vase}	2
{Flowers ,Card, Vase}	1

a) AIS Algorithm: When the resultant of association rule gives rise to one item only is represented by AIS algorithm. To explain rules like P ∩ Q ⇒ R can be generated but not the rules like P⇒ Q ∩ R. This system runs over the entire database. It scans with multiple passes on the entire database for all records. While scanning the first pass those data items are counted which are large in number and are present most of the time. They are categorized as large data sets. Every large dataset with every pass generates candidate data sets. The product in product range whose support for the calculate is less than its minimum value is deleted from the transaction record. Candidate item sets generation and frequent item sets generation process iterate until any one of them becomes empty. After all the candidate product in product range and frequent product in product range are presumed to be reserved in the main memory, memory management is also proposed for AIS when memory is not enough. This algorithm uses NoSQL database. The above diagram explains the concept of AIS algorithm.

Merit:-

This algorithm gives an idea of relationship between different categories of a client's buying

Characteristics.

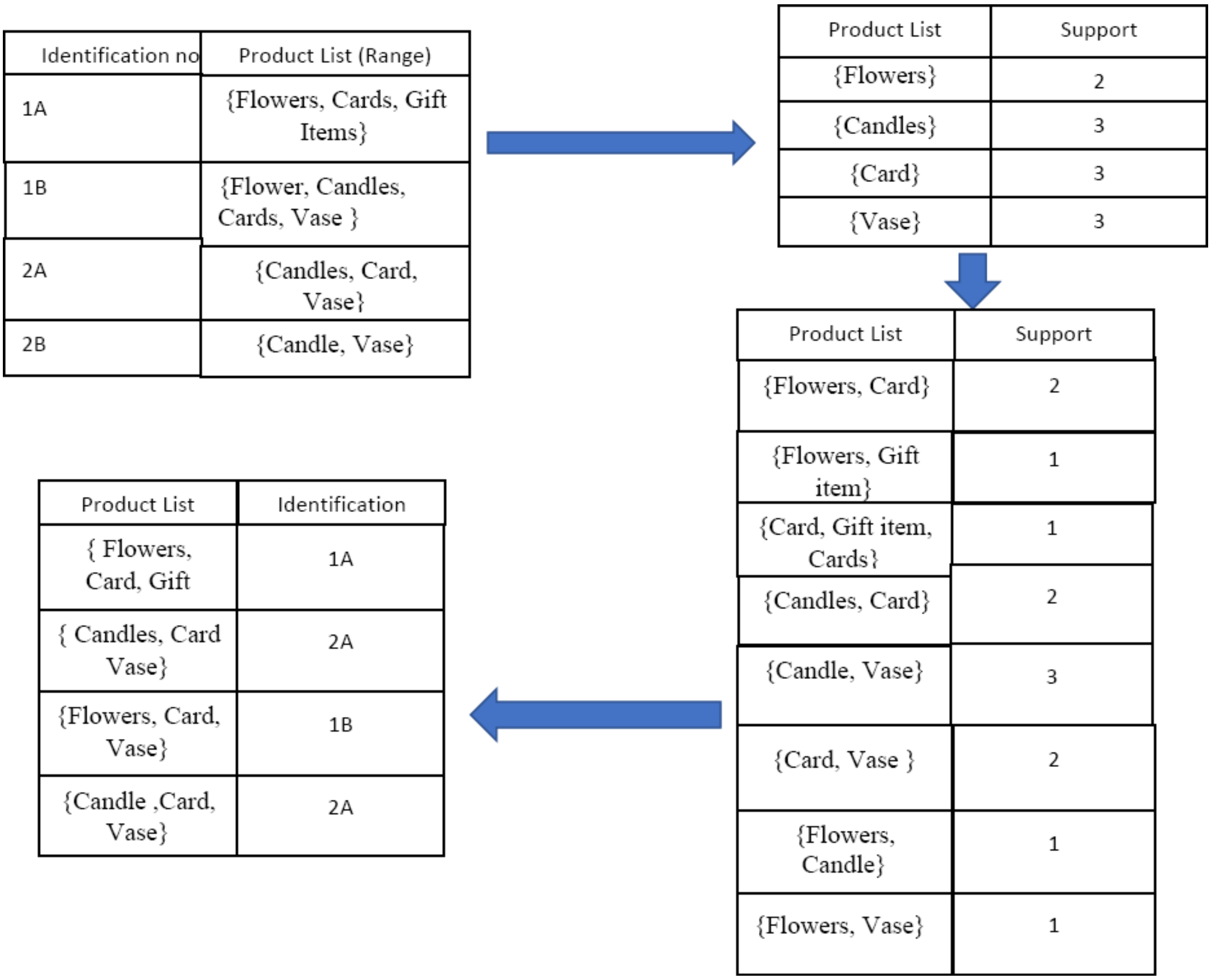
Demerit:-

1. Many candidate item sets arise and as the data base scanning process continues reduces from smaller to smallest and the no specified data structures are used to maintain the candidate item sets.

Steps to perform AIS Algorithm

```
1.  Big1 = {Big1- Product in Product Range };
2.  for ( i= 2; Bigi-1 not equal to Null i++ ) do begin
3.  Counti = Null;
4.  for all transactions T belongs to the data base DB do begin
5.  BigT= Subset{Bigi-1,T)\\ bigger product in product range contained on transaction T
6.  for all Product in product range MaT belongs to BigT do begin
7.  CountT = 1-extensions of MaT belongs to CountT do
8.  If (Candidate belongs to Counti) then
9.  Add 1 to the count of Candidate in the corresponding entry in Counti else
    Add candidate to Counti with a Count of 1
10. end
11. Bigi={Candidate belongs to Counti |Candidate.count >=Minimum support };
12. End
13. Answer = Upperi Loweri
```

Here as the database is scanned, gives rise to candidate product in product range and number of candidate product in product range are counted at the end of each scanning.



Identification no	Product List (Range)
1A	{Flowers, Cards, Gift Items}
1B	{Flower, Candles, Cards, Vase }
2A	{Candles, Card, Vase}
2B	{Candle, Vase}

Product List	Support
{Flowers}	2
{Candles}	3
{Card}	3
{Vase}	3

Product List	Support
{Flowers, Card}	2
{Flowers, Gift item}	1
{Card, Gift item, Cards}	1
{Candles, Card}	2
{Candle, Vase}	3
{Card, Vase }	2
{Flowers, Candle}	1
{Flowers, Vase}	1

Product List	Identification
{ Flowers, Card, Gift	1A
{ Candles, Card Vase}	2A
{Flowers, Card, Vase}	1B
{Candle ,Card, Vase}	2A

Figure 1.5 Pictorial Representation of SETM Algorithm

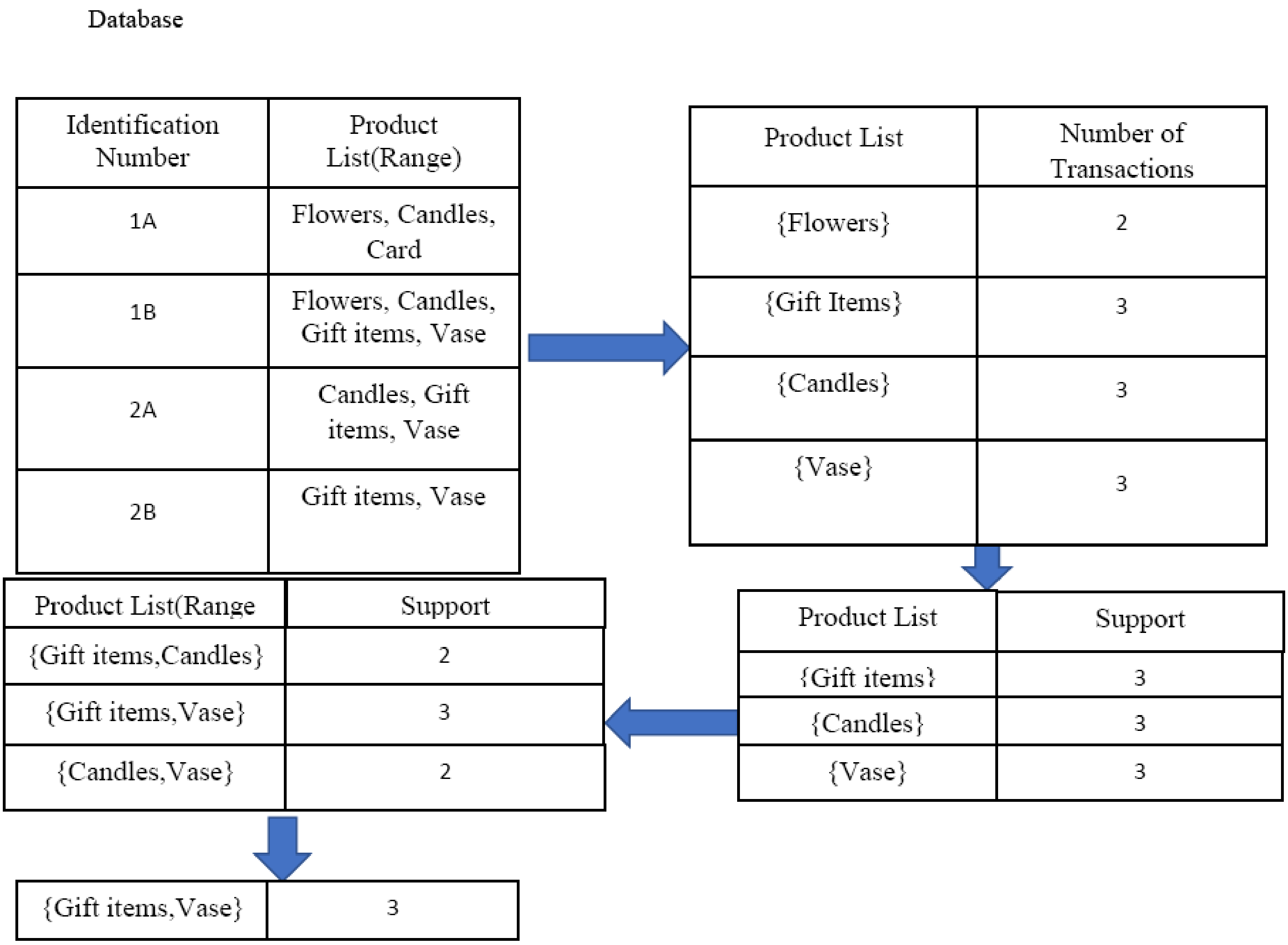
b) SETM Algorithm:- This algorithm works more like AIS algorithm, the SETM algorithm makes multiple passes over the database. In the first pass, it counts the number of every product which appears and determines which of them are large or more visible in the database. Then, it generates the candidate product from the range extending large Product sets from the previous pass. In addition, the SETM tallies the identification number of product range which are available at transaction database which in turn, generate the candidate product from product range which are counted at the exit of each pass. In the table the identification number of Transaction product range is saved along with candidate product range in a sequential manner. The candidate product in product range is not counted regularly but counts the end of each pass by finding the total number of candidate product ranges. In relational database JOIN command is used to generate candidate product range. This algorithm is used for relational database using SQL commands. The above diagram explains the working of SETM algorithm.

Merit: - 1. As Identification number of transaction record keeps a copy of candidate item thus both are kept together in a sequential presentation.

Demerit: - 1. This algorithm keeps both the identification item and candidate item of the transaction record thus needs more storage space to store transaction records. 2.It is not very convenient to use for relational Database. 3. For every candidate item set, there are as many operations as each operation is associated with support value and occupies the maximum memory.

Steps to perform SETM algorithm

1) Big1 = {Big1- Product in Product Range};
2) Big1 = {Big1 -Product in Product range together with the identification number of transaction tid data base DB and products appear, on sorted manner);
3) for (i= 2; Bigi-1 not equal to Null i++) do begin
4) Counti = Null;
5) for all transactions T belongs to the data base DB do begin
6) BigT= {Ma belongs to Bigi-1| Ma.tid= T.tid } // Big (i-1)product contained in product range contained in T
7) for all Product in product range MaT belongs to BigT do begin
8) CountT = 1-extensions of MaT contained in T; // Candidates in T
9) Counti += ;{T.tid,candidate>| candidate belongs to CountT
10) end
11) end
12) Sort Counti on Product given in Product Range;
13) Delete all Product available in the range for candidate Counti for which Candidate.count < minimum support giving Bigi;
14) Bigi = {Ma.product, count of Ma in Bigi>| Ma belongs to Bigi };//combined with step 13
15) Sort Bigi on identification number in Transaction tid DB;
16) end
17) Answer = Upperi Loweri



Database

Identification Number	Product List(Range)
1A	Flowers, Candles, Card
1B	Flowers, Candles, Gift items, Vase
2A	Candles, Gift items, Vase
2B	Gift items, Vase

Product List	Number of Transactions
{Flowers}	2
{Gift Items}	3
{Candles}	3
{Vase}	3

Product List	Support
{Gift items}	3
{Candles}	3
{Vase}	3

Product List(Range	Support
{Gift items,Candles}	2
{Gift items,Vase}	3
{Candles,Vase}	2

{Gift items,Vase}	3

Figure 1.6 Pictorial Representation of APRIORI Algorithim

c) APRIORI Algorithm: Apriori algorithm is the most prominent and used association rule algorithm. In Apriori algorithm the candidate product in product range arise by scanning the previous passes where large product range appear. All the large product range are joined together of the previous passes and others are deleted whose existence is scanty among the transaction records of the database. Then putting the large product in the range and the same product from the previous pass and deleting those subsets which are small in the previous pass without considering the transactions in the database. The main consideration is of large product range of the previous pass, the number of candidate large product range is significantly reduced. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation). This algorithm forces to generate the frequent products from the product range which can be used to the database and associated with association rules to generate a pattern. The above picture explains the working of Apriori algorithm.

Merits:-

1. Large product in product range property usage is predominant.
2. Easily parallelized
3. Easy to implement
4. The Apriori algorithm implements level-wise search using frequent item property

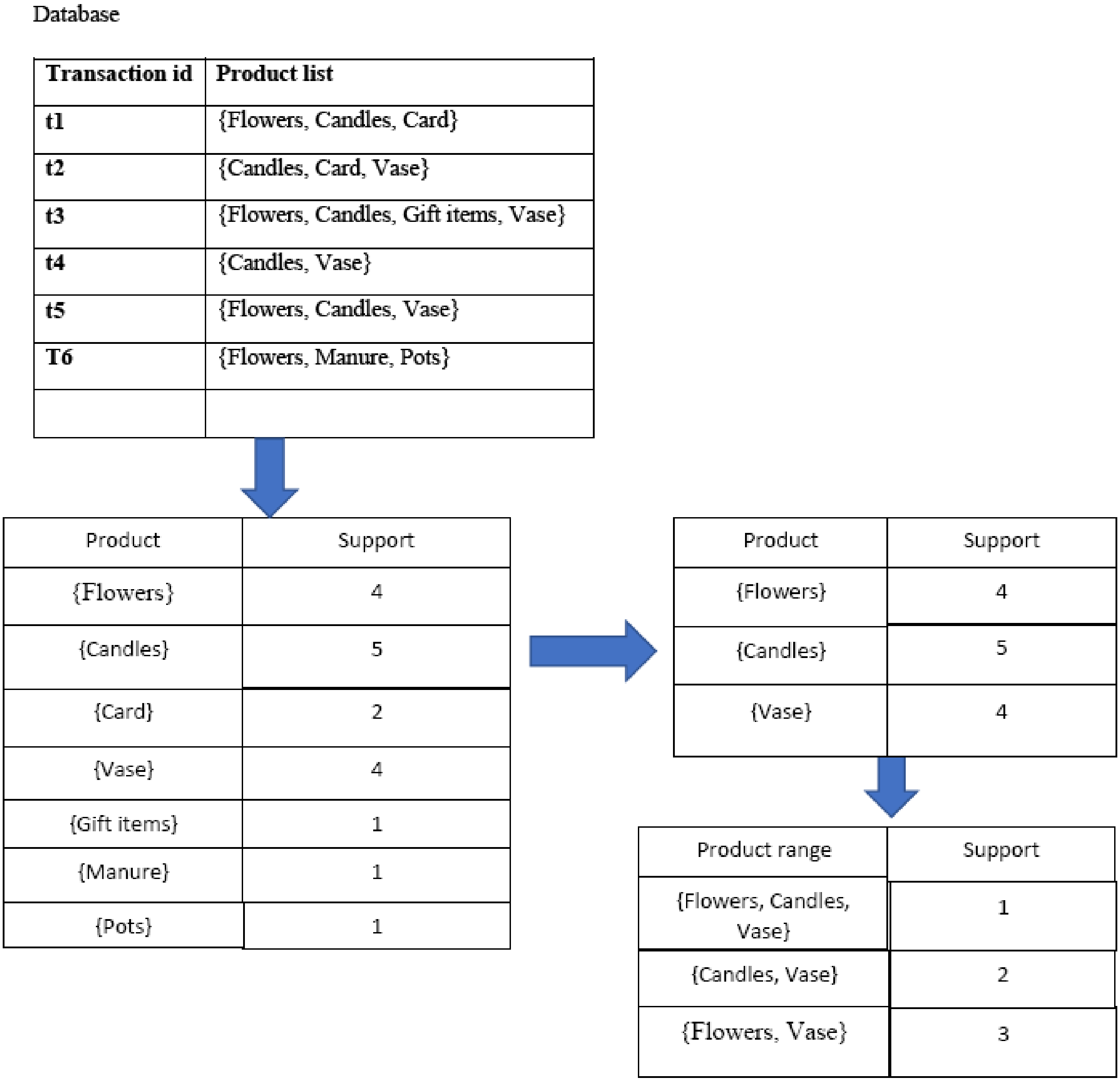
Demerits:-

1. Database scanning takes place many times for calculating periodic product (reduce performance)
2. It assumes that transaction database is memory resident
3. Generation of candidate item to different sets is expensive (in both space and time)
4. Support counting is expensive • Subset checking (computationally expensive)

Steps to execute Apriori algorithm

1) Big1=Check_Frequent_1_Product in Product range (Transaction T)
2) For (I=2; Li-1 not equal to NULL; I++)
{ // generate the CountI from the BigI-1
3) CountI=candidate generated from BigI-1;//get the product PK with minimum support in CountI Using Big1(1<=K<=I)

```
4)  P=Get_Product_minimum_support(CountI,Big1);//get the target product with identification number that contain product P
5)  Tgt=Get_transaction_identification_number(P)
6)  For each transaction T in Tgt do
7)  Increment the count of all products in CountI that are found in Tgt;
8)  BigI=Products in CountI>=minimum_support
9)  end;
10) }
    Return Upper K Lower K
```



Database

Transaction id	Product list
t1	{Flowers, Candles, Card}
t2	{Candles, Card, Vase}
t3	{Flowers, Candles, Gift items, Vase}
t4	{Candles, Vase}
t5	{Flowers, Candles, Vase}
T6	{Flowers, Manure, Pots}

Product	Support
{Flowers}	4
{Candles}	5
{Card}	2
{Vase}	4
{Gift items}	1
{Manure}	1
{Pots}	1

Product	Support
{Flowers}	4
{Candles}	5
{Vase}	4

Product range	Support
{Flowers, Candles, Vase}	1
{Candles, Vase}	2
{Flowers, Vase}	3

d) FP-Growth Algorithm: It is used for multiple database scanning. Here first we take the database. Then we offer partition into the database for various partitions. FP-growth method searches the least frequent items as a suffix and then by joining the suffixes finds the long frequent patterns. Finding the long frequent patterns through least frequent item is considered the best selection procedure.

The diagram above explains the working of FP growth algorithm.

This handles the problem of finding the most common existing patterns whose longevity is more than other products. It also looks or shorter ones recurrently and then joins the suffix. It uses the least frequent items as a suffix, offering good selectivity. The method substantially reduces the search costs. When the database is large, it is sometimes unrealistic to construct a main memory-based FP tree. An interesting alternative is to first partition the database into a set of projected databases, and then construct an FP-tree and mine it in each projected database. Such a process can be recursively applied to any projected database if its FP-tree still cannot fit in main memory

Merits:-

1. Faster than Apriori algorithm
2. No candidate generation
3. Only two passes over dataset
4. Uses productive ,scalable for mining both long and short frequent patterns.

Demerits:-

1. FP algorithm tree may not fit in memory
2. FP algorithm tree is expensive to build

Steps to follow for simple FP Growth

Begin

```
1) Frequent Pattern {};
2) Insert len1 regular pattern in Frequent Pattern
3) Until all regular patterns in frequent pattern are
   scanned do begin
4) Generate a can product from the product range from
   one or more frequent patterns in Frequent pattern
5) If (support {product, Transaction in DB}>=support
6) Add Product to frequent pattern range to set new
   Frequent Pattern
7) end
8) end
```

Application of Association:

The applications include:

- bioinformatics, natural disaster prediction
- image classification, Medical diagnosis, Protein Sequences
- network traffic analysis, Population census calculation
- analyzing customer reviews,
- activity monitoring, malware detection,
- E-learning…

CONCLUSION

The success and progress occurances in database technology is accelerated and the requisition and solicitation of database management system in every field is highly in demand. More the collection of data better is the future prediction activating the necessity of discovery of Association rule.

Data Mining is briskly developed to meet a request like design identification through recurrent product in product range, finding a potential value for the decision making , predicting a future situation etc. Data mining is to snippet implied information, previously unrevealed knowledge and rules from large database or data warehouse, and then utilize these evolved knowledge and rules to solve the problem.

The discovery and study of all the algorithms for the association rule is a most successful and most vital role in the data mining, is a very active research area in current data mining. Its goal is to discover all frequent models in the product used in product range.

The current research work carrying on are mostly `focused on the development of effective algorithm. This paper portraits the different applications of Association rules signifying data mining and widely used to analyze retail basket or day to day data, and are calculated to figure out strong rules used in transaction data for estimating investitive concern based on the concept of strong rules like Joint promotions, co-op advertising, bundled offerings etc.

We have also discussed in this paper AIS, SETM, Apriori, FP-growth four association rule mining algorithms with their example: Comparison is done based on the above performance criteria. Each algorithm has some advantages and disadvantages. It is also strongly visible that there is an improvements on number of scanning ,more storage utilization,optimal design acceptance for effective and decisive mining.From the above comparison we can conclude that, FP-growth performs better than all other algorithms.

ACKNOWLEDGEMENTS

I would like to thank my guide Ms.Prinima Gupta and co-guide Prof.Felix Musua for their expert advice and encouragement. I would also like to thank Faithful Wachira, Carolyne Kiarie and Maryanne Gichuhi for your honest support and co-operation.

REFERENCES

[1] W. Lin, S.A. Alvarez, and C. Ruiz. Efficient adaptive–support association rule mining for recommender systems. Data Mining and Knowledge Discovery, 6(1):83–105, January 2002.

[2] Yi - Dond Shen , Qiang Yang and Zhong Zhang (2002),“Objective - oriented utility-based associationmining “,Proceedings of the 2002 IEEE International conference on Data Mining.

[3] P.R. Pal, R. C. Jain, CAARMSAD: “Combinatorial Approach of Association Rule Mining for Sparsely Associated Databases”. Journal of Computer Science, Tamilnadu India, Vol. 2, No 5, pp 717, July 2008.

[4] Xianneng Li, Shingo Mabu, Huiyu Zhou, Kaoru Shimada and Kotaro Hirasawa, “Analysis of Various Interestingness Measures in Classification Rule Mining for Trac Prediction”, SICE Annual Conference 2010, August 18-21, 2010, pp. 1969 – 197

[5] V.Umarani and M.Punithavalli , April 2010 ”On Developing an Effectual Progressive Sampling Based Approach for Association Rule Discovery”, In the proceedings of 2nd IEEE International Conference on Information and data Engineering (2nd IEEE ICIME 2010), Chengdu ,China .

[6] V.Umarani et al., 2010 ”A Study on Effective Mining of Association Rules from Huge Databases”, International journal of computer science and research,Vol .1,issue 1

[7] B.Ramasubbareddy, Dr.A.Govardhan, Dr.A.Ramamohanreddy, Nov 2010 ” An Approach for Mining Positive and Negative Association Rules ”, International Journal of Recent Trends in Engineering and Technology, Vol.4,No.1.

[8] Ogunde A O, Folorunso O, Sodiya A S Oguntuase J A, and Ogunleye G O, 2011 “ Improved Cost Models For Agent Based Association Rule Mining In Distributed Databases”, Anale. Seria Informatica. Vol. IX. [56] Christopher.T, 2010 “ Character Based Weighted Support Threshold Algorithm Using Multi criteria Decision Making Technique “, International Journal On Computer science And Engineering Vol. 02, No. 04, 2010, pp. 965-971.

[9] Claudia Marinica and Fabrice Guillet, June 2010, ” Knowledge–Based Interactive Postmining of Association Rules Using Ontologies” IEEE Transactions On Knowledge And Data Engineering Vol.22 NO.6

[10] Fadi Thabtah, Hussein Abdel-jaber, Mofleh Aldiabat. Rule Pruning Methods in Associative Classification Text Mining . Journal of Intelligent Computing Volume 1 Number 1 March 2010.

[11] Rashmi Shikhariya, Nitin Shukla.. An improved association rule mining with fp tree using positive and negative integration. Journal of Global Research in Computer Science. Volume 3, No. 10, October 2012.

[12] D. Magdalene Delighta Angeline, . Association Rule Generation for Student Performance Analysis using Apriori Algorithm. The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 1, No. 1, March-April 2013

[13] Larose, D. T. “Discovering knowledge in data: an introduction to data mining”. John Wiley & Sons,2014.

[14] Oweis, N. E., Owais, S. S., George, W., Suliman, M. G., & Snášel, V. “A Survey on Big Data, Mining: (Tools, Techniques, Applications and Notable Uses)”. In Intelligent Data Analysis and Applications Springer International Publishing, pp. 109-119, 2015

[15] Data Mining: Practical machine learning tools and techniques IH Witten, E Frank, MA Hall, CJ Pal - 2016 - Morgan Kaufmann .