Journal of Basic and Applied Engineering Research
Print ISSN: 2350-0077; Online ISSN: 2350-0255; Volume 1, Number 7; October, 2014 pp. 1-4

http://www.krishisanskriti.org/jbaer.html

Association Rules in Data Mining

Aayushi Maheshwari[1], Garima Kharbanda[2], Harsh Patel[3]

[1]*Institute of Technology, Nirma University, Ahmedabad, India*

Abstract: **Data mining is motivated by the decision support problem faced by most large retail organizations. Association rule mining is finding frequent patterns, associations, correlations or casual structures among sets of items or objects in transactional databases, relational databases and other information repositories. It has various applications including market-basket data, analysis, cross marketing, catalogue design, and loss-leader analysis. For example, 98% of customers that purchase tires and auto accessories also get automotive services done. Finding all such rules is valuable for cross-marketing and attached mailing applications. In this paper presentation we will analyses the various data association rules and develop an insight into the implementation of these rules for better sales of a company. Moreover in data mining association rules are useful for analyzing and predicting customer behavior. We will also throw a light on Apriori Algorithm, which is probably the best known algorithm for learning association rules.Apriori is designed to operate on databases containing transactions. For example: Collection of items bought by customers or detail of a website frequentation. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets, as long as those item sets appear sufficiently often in the database.**

Keywords: repositories, market based, crossmarketing, transitions.

1. INTRODUCTION

"We are drowning in data but starving for knowledge."

A large amount of data is being generated by many organizations on a day to day basis. Generally, **data mining** (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cut costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.Data mining is considered to deal with huge amounts of data which are kept in the database, to locate required information and facts. Data mining techniques can be categorized according to the objectives they follow and the results they offer, which obtains computer as a tool and makes use of the skill and knowledge significance to comprehend and explain the problem. Various data mining techniques such as, decision Trees, association rules, and neural networks are already presented and become the point of attention for several years. Association rule mining technique is the most efficient data mining technique to search hidden or desired pattern among the huge amount of data. It is responsible to get correlation relationships among various data attributes in a large set of items in a database.

Although data mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

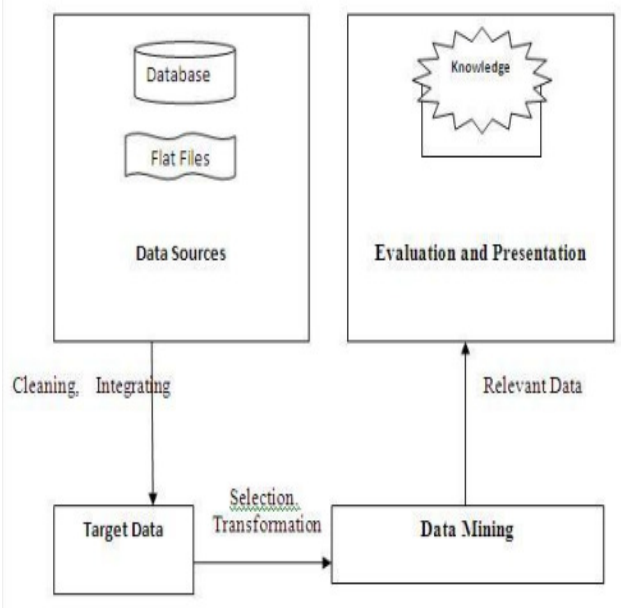

Fig.1. Data Mining Process

Association rule mining, one of the most important and well researched techniques of data mining, aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, and inventory control etc.

1). Apriori algorithm
2). Eclat algorithm
3). FP growth algorithm
4). Node set based algorithm
5). GUHA procedure ASSOC
6). OPUS search
7). Context based association rule mining algorithm.

It is by far the most important data mining algorithms for mining frequent item sets and associations. It opened new doors and created new modalities to mine the data. Since its inception, many scholars have improved and optimized the Apriori algorithm and have presented new Apriori-like algorithms. The authors became living legends in the data mining communities. Apriori uses a breadth-first search strategy to count the support of item sets and uses a candidate generation function which exploits the downward closure property of support.

Although technology provides numerous benefits to young people, it also has a 'dark side', as it can be used for harm, not only by some adults but also by the young people themselves. E-mail, texting, chat rooms, mobile phones, mobile phone cameras and web sites can and are being used by young people to bully peers. It is now a global problem with many incidents reported in the United States, Canada, Japan, Scandinavia and the United Kingdom, as well as in Australia and New Zealand. This growing problem has as yet not received the attention it deserves and remains virtually absent from the research literature.

Cyber bullying can be defined as an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him of herself. Hence we also throw light upon the possibility of relating cyber bullying with data association techniques and thereby identifying victims and perpetrators of cyber bullying.

NORA is a technology that can find incomprehensible hidden connections between people, connections between people or other entities by analyzing information from many different sources to correlate relationships.

Systems Research & Development (SRD) developed its Non-Obvious Relationship Awareness (NORA) technology to help casinos identify cheaters by correlating information from multiple sources about relationships and earlier transactions.

"It will tell you that the Drug Enforcement Agency's agent's college roommate's ex-wife's current husband is the drug lord, " says Jeff Jonas, chief technology officer at SRD. NORA can bridge up to 30 such links, he says.

The new NORA module uses streaming technology that scans data and extracts information in real time as it flows by. That would allow it to, for example, instantly discover that a man at an airline ticket counter shares a phone number with a known terrorist and then issue an alert before he can board his flight.

2. RELATED WORKS

The concept of association rules in data mining was introduced by Agrawal et al in 1993 with the objective of finding interesting and useful patterns in transactional database. It has acquired more than 6000 citations according to Google Scholar, as of March 2008, and is thus one of the most cited papers in the Data Mining field. It is however possible that what we now refer to as "association rules" is similar to what appears in the 1966 paperon GUHA, a general data mining method developed by Petr Hajek et al. The database contains transactions which consist of a set of items and a transaction identifier(eg. market basket).

Many algorithms for generating association rules have been presented over time. A few well known algorithms are Apriori Eclat and FP-Growth(proposed by Han in 2000).However Apriori algorithm submitted by Agrawal and R. Srikant in 1994 is the most effective algorithm as it opened new pathways for mining the data. Since its commencement, many scholars have improved and optimized the Apriori algorithm and have presented new Apriori-like algorithms. The authors have become living legends in the data mining communities. They both received masters and PhDs from University of Wisconsin, Madison and both worked for IBM. The IBM's Intelligent Miner was created mainly by them. Once colleagues, they now work for competing companies – Agrawal for Microsoft and Srikant for Google. Apriori uses a breadth-first search strategy to count the support of item sets and uses a candidate generation function which exploits the downward closure property of support.

3. HOW ASSOCIATION RULES WORK?

The usefulness of association rules to address unique data mining problems is best illustrated in a simple example. Suppose we are collecting data at the check-out cash registers at a large book store. Each customer transaction is logged in a database, and consists of the titles of the books purchased by the respective customer, perhaps additional magazine titles and other gift items that were purchased, and so on. Hence, each record in the database will represent one customer (transaction), and may consist of a single book purchased by that customer, or it may consist of many (perhaps hundreds of) different items that were purchased, arranged in an arbitrary order depending on the order in which the different items (books, magazines, and so on) came down the conveyor belt at the cash register. The purpose of the analysis is to find associations between the items that were purchased, i.e., to derive association rules that identify the items and co-

occurrences of different items that appear with the greatest frequencies. For example, we want to learn which books are likely to be purchased by a customer who we know already purchased (or is about to purchase) a particular book. This type of information could then quickly be used to suggest to the customer those additional titles. You may already be familiar with the results of these types of analyses if you are a customer of various on-line (Web-based) retail businesses; many times when making a purchase on-line, the vendor will suggest similar items (to the ones purchased by you) at the time of "check-out", based on some rules such as "customers who buy book title *A* are also likely to purchase book title *B*, " and so on.

4. DISCUSSIONS ON ASSOCIATION RULE WITH APRIORI ALGORITHM

BASIC-TERMINOLOGIES

A. Data Definition

Data mining is a technology with great potential to help companies focus on the important information in the data they have collected from the behavior of their customers. Data generally by organized in tables that hold a set of complete database environment. Tables contain set of information and especially different symbolizes which is coupled with a test case to analyses.

B. Association Rule

Association rule basically express how items or objects are related to each other and how they formed a group together. As example, if a customer buys a dozen eggs, he is 80% likely to also purchase milk.

- Definition – Main discovery of association rules from a particular transaction database. X={x1, x2… xn} be array of n different element called item-sets in that database.
- Support - It indicates frequency of the items in the database.
- Confidence – It indicates the number of times the statements have been found to be true.

The Apriori Algorithm is an influential algorithm for mining frequent item sets for Boolean association rules. Apriori simply scan the whole database and count the frequency for each item. Basic Indicators of association rules the support and confidence are the main factors in process of Apriori algorithm. The Apriori uses bottom-up breadth-first approach and hash tree to find the large item sets. The algorithm finds frequent item-sets with cardinality from 1 to k (k-item set). It generates candidate k-itemsets from frequent (k-1) – item sets, and prunes candidate itemsets. With the support counting, get candidate k-itemset then it generate frequent (k-1) – itemsets, so back and fourth, until the frequent item-set can not be produced.

C. Pseudo code For Apriori Algorithm

Join Step: C_k is generated by joining Lk-1with itself

Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

C_k: Candidate itemset of size k
L_k: frequent itemset of size k
L_1= {frequent items};
for(k= 1; L_k!=∅; k++) do begin
C_{k+1}= candidates generated from L_k;
for eachtransaction tin database do
increment the count of all candidates in C_{k+1}that are contained in t
L_{k+1}= candidates in $C_{k+1}$with min_support
end
return$\cup_k L_k$;

5. CYBERBULLYING DETECTION

Cyber bullying is a form of bullying which has in recent years become more apparent, as the use of electronic devices such as computers and mobile phones by young people has increased.

Electronic Technology comprising of devices and equipment such as cellphones, computers, tablets and communication tools including social media sites, text messages, chat, and websites have triggered cyber bullying. Few Examples are: obscene text messages or emails, demeaning rumors sent by emails or posted on social-networking sites, and embarrassing pictures, videos, websites, or fake profiles. We try to present a possibility of identifying the victims of cyber bullying through the use of association rules. The frequency of communication between two people has a direct relationship with the probability of them being related with each other closely. Hence if we can determine the contact with which the victim has communicated with the most and a contact which follows the victim constantly be it on social networking sites or over the cellular network we can chalk out a list of people who maybe involved in bullying the victim. Through the use of “Gephi” which is an interactive visualization and exploration platform for all kinds of networks and complex system. Dynamic and hierarchical graphs. We will be able to find out the most frequently contacted person. Unlike the case of Asharam bapu wherein NORA (Non-obvious-relationship-awareness) was used to identify people who were associated with him. The application of one of the association rules after listing out the most doubtful contacts will prove to be a better option and prevent us from putting the innocent behind the bars . For example, if a girl named Meghna has been the victim of cyber bullying the steps we can follow to identity the victim are:

1) Use the software package “Gephi” to from a list of contacts that follow Meghna as well as have contacted Meghna Mostfrequantly.

2) To back it up We can also perform a check on Meghna’s call log to identify if she has been talking to someone excessively over the past few months.

3) With the help of Apriori Algorithm and then association rule we can find out the frequencies against cach contact we can then obtain the contact with maximum chances of being guilty.

4) This presents a possibility of the Guilty being caught easily and with a better efficiency.

6. ACKNOWLEDGMENT

This research paper is made possible through the help and support from everyone, including: parents, teachers, family and friends. We would also like to thank our teacher Mr.Vivek Prasad for his utmost support and encouragement.

7. CONCLUSION

With the continual improvements in technology and rapid increase in data, it becomes important to design systems to effectively manage data. Data association techniques provide an organized means to determine a common link and interesting relations between massive amount of data.

Association rule mining helps uncover hidden patterns between seemingly unrelated data thus providing a useful means for determining predators in cyber bullying, a crime which is on an upswing in the current scenario. Collaborating the concept of data association with graphic software Gephi, which is used by applications like Facebook and Twitter to determine network traffic.

A wide range of applications in many areas of business practice and also research - from the analysis of consumer preferences or human resource management, to the history of language use these investigative techniques. These techniques enable analysts and researchers to discover hidden patterns in large data sets, such as "customers who order product *A* often also order product *B* or *C*" or "employees who said positive things about initiative *X* also frequently complain about issue *Y* but are happy with issue *Z*." Based on predefined threshold values for detection, the apriori algorithm allows us to rapidly process such enormous data sets. On delving deeper into data mining and association techniques we can curb cyber bullying to a large extent.

REFERENCES

[1] Jaiwei Han and Micheline Kamber, ”Data Mining Concepts and Techniques”, Second Edition, Morgan kaufmann Publishers.

[2] Association Rules, Apriori Algorithm – Wikipedia, the free encyclopedia.

[3] Cyberbullying, Gephi – Wikipedia, the free encyclopedia

[4] S.Suriya, Dr.S.P.Shantharajah, R.Deepalakshmi.

[5] International Journal of Advanced Scientific and Technical Research,

[6] Issue, Volume (February 2012), ISSN: 2249-9954 page no 2. Figure 1, 2.

[7] Rachna Somkunwar, International Journal of Advanced Research in Computer Science and Software Engineering.Volume 2, Issue 9, September 2012.page no 2.