Association Rules Mining for Business Intelligence

Rashmi Jha

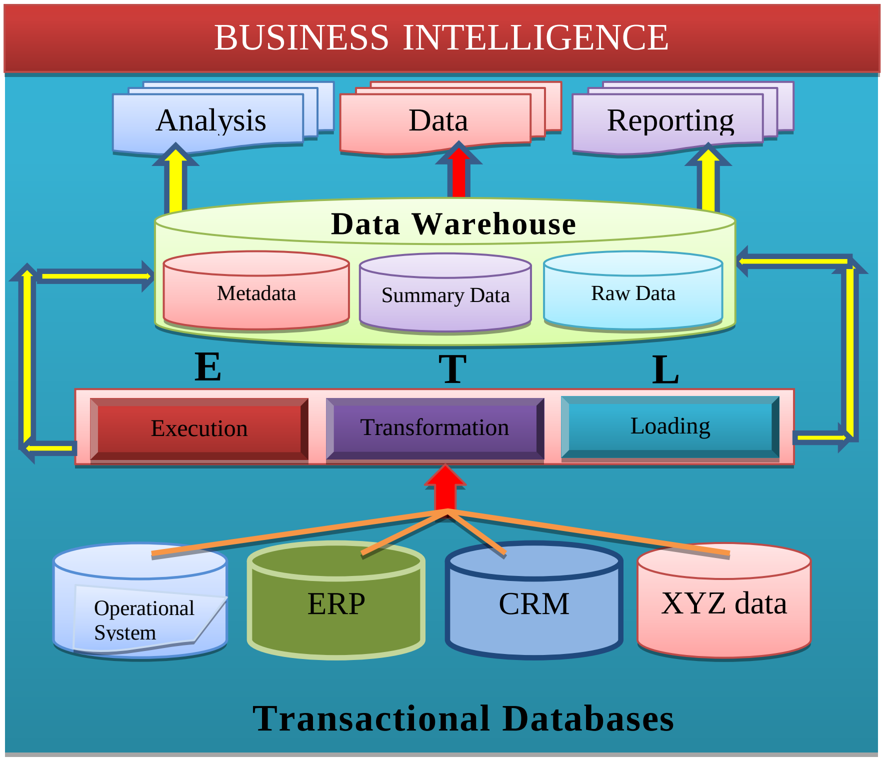
NIELIT Center, Under Ministry of IT, New Delhi, India

Abstract- *Business Intelligence (BI) is any information derived from analytics of existing data that can be used strategically in the organization. Data Mining is a subset of BI or a means/process of deriving BI from data using statistical modeling of the data. It can be used to find relationships/correlations between the various data elements captured which can be used to improve business performance or at least understand what is happening better. With the rapid exponential growth in size and number of available Databases in commercial, industrial, administrative and other applications, it is mandatory and important to examine how to extract knowledge from voluminous data. Mining Association rules in transactional or relational databases has recently attracted a lot of attention in database communities. The task is to derive a set of strong association rules in the form of “A1^....^ m=>B1^...^Bm” where Ai(for i €{1,2,.....m}) and Bi(for j € {1,2,.....,n}) are set of attribute-values, from the relevant data sets in a databases.*

Index Terms- Data Mining, Association, Knowledge Discovery, Algorithm, Integration, Pattern Discovery.

I. Introduction

Business Intelligence is having three essential parts as extraction, transformation and loading (ETL) software which extracts data from operational databases, then structures and organizes the data into Ia form (or data model) suitable for efficient analyses. The second is a set of databases and files (which may variously be data warehouses, data marts or data cubes) which store and make available the data in its analyzable form. The third is a set of end user tools that enable users to design, run and view reports, queries and analyses. Design of all these three elements for data mining (or data driven analysis) has to be very different to the design of conventional business intelligence systems which are intended for model based query and reporting. BI is about making intelligent business decision for an event that has not yet happen. Meaning based on a certain pattern happened from the past, you can pre-empt or predict what will likely to happen in the future. Decision Making is the re-organization of collected data from various customers' touch points, to facilitate analysis and diagnosis, and all these corresponding data-collection events were already past. DM is what actually has happened.



Business intelligence can be applied to the following business purposes, in order to drive business value.

1) **Measurement –** program that creates a hierarchy of performance metrics and benchmarking that informs business leaders about progress towards business goals.
2) **Analytics –** program that builds quantitative processes for a business to arrive at optimal decisions and to perform business knowledge discovery. Frequently involves: data mining, process mining, statistical analysis, predictive analytics, predictive modeling, business process modeling, complex event processing and prescriptive analytics.
3) **Reporting/enterprise reporting –** program that builds infrastructure for strategic reporting to serve the strategic management of a business, not operational reporting. Frequently involves data visualization, executive information system and OLAP.
4) **Collaboration/collaboration platform** – program that gets different areas (both inside and outside the business) to work together through data sharing and electronic data interchange.
5) **Knowledge management –** program to make the company data driven through strategies and practices to identify, create, represent, distribute, and enable adoption of insights and experiences that are true business knowledge. Knowledge management leads to learning management and regulatory compliance.

II. Requirements and challenges of Data Mining

In order of conduct effective Data mining, one needs to first examine what kind of features an applied knowledge discovery system is expected to have and what kind of challenges one may face at the development of Data Mining technique.

1) **Handling of different types of data**
Because there are many kinds of data and databases used in different applications, one may expect that knowledge discovery system should be able to perform effective data mining on different kinds of data.
Specific data mining system should be constructed for knowledge mining on specific kind of data, such as systems dedicated to knowledge mining in relational databases, transaction databases, spatial databases, multimedia databases, etc.

2) **Efficiency and reliability of Data Mining algorithms**
To efficiently extract information from a huge amount of data in databases, the knowledge discovery algorithms must be efficient and scalable to large databases. That is, the running time of a data mining algorithm must be predictable and acceptable in large databases. Algorithms with exponential or even medium order polynomial complexity will not be practical use.

3) **Usefulness, certainty and expressiveness of data mining results.**
The discovery knowledge should accurately portray the content of the database and useful for certain applications. The imperfectness should be expressed by measures of uncertainty, in the form of approximate rules or quantitative rules. Noise and exceptional data should be handled elegantly in data mining systems. This also motivates a systematic study of measuring the quality of the discovered knowledge, including interestingness and reliability, by contraction of statistical, analytical, and simulative models and tools.

4) **Expression of curious kinds of data mining requests and results**
Different kinds of knowledge can be discovered a huge amount of data. Also, one may lime to examine discovered knowledge from different views and preserve them in different forms. This requires us to express all the data mining requests and discovered knowledge in high level languages or graphical user interfaces so that the data mining task can be specified by no experts and the discover knowledge can be understandable and directly useable by users. This also requires the discovery system to adopt expressive knowledge representation techniques.

5) **Interactive mining knowledge at multiple abstraction levels.**
Since it is difficult to predict what exactly could be discovered from database, a high level data mining query should be treated as a probe which may disclose some interesting traces for further exploration. Interactive discovery should be encouraged, which allows a user to interactively refine a data request, dynamically change data focusing, progressively deepen a data mining process flexibly view the data mining results at multiple abstraction levels and from different stages.

6) **Mining Information From different sources of data**
The widely available local and wide area computer network, including internet, connect many sources of data and form huge distributed, heterogeneous databases. Mining knowledge from different sources of formatted or unformatted data with diverse data semantics poses new challenges to data mining. On the other hand, data mining may help disclose the high level data regularities in heterogeneous databases which can hardly be discovered by simple query systems.

7) **Protection of privacy and data security**
When data can be viewed from many different angles and at different abstractions levels, it is the goal of protecting data security and guarding against the invasion of privacy. It is important to study when knowledge discovery may lead to an invasion of privacy, and what security majors can be developed for preventing the disclosure of sensitive information.

III. Classifying Data Mining Techniques

Different classification schemes can be used on the kinds of databases to be studied, the kinds of knowledge to be discovered, and the kinds of techniques to be utilized as shown below;

A. *What Kind Of Databases To Work On*

A data mining system can be classified according to the kinds of databases on which the data mining is performed. For example, a system is a relation minor if it discovered knowledge from relational data, or an object oriented one if it mines knowledge from object-oriented databases. In general, a data minor can be classified according to its mining of knowledge from the following different kinds of databases: relational databases, transaction databases, object-oriented databases, deductive databases, temporal databases, multimedia databases, heterogeneous databases, active databases, legacy databases and the internet information-base.

B. *What Kind Of Knowledge To Be Mined*

Several typical kinds of knowledge can be discovered by data minors, including association rules, characteristics rules, classification rules, discriminant rules, clustering evolution and deviation analysis. Moreover, data miners can also be categorized according to the abstraction level of its discovered knowledge which may be classified into generalized knowledge, primitive-level knowledge, and multiple-level knowledge. A flexible data mining system may discover knowledge at multiple abstraction levels.

C. *What Kind of Techniques To Be Utilized*

Data Mining can also be categorized according to the underlying data mining techniques. For example, it can be categorized according to the driven method into autonomous knowledge miner, data driven miner, query driven miner, and interactive data mining. It can also be categorized according to its underlying data mining approach into generalization based mining, pattern based mining, mining based on statistics or mathematical theories, and integrated approached etc

IV. Mining Different Kind of Knowledge From Databases

Data Mining is an application dependent issue and different applications may require different mining technique. In general the kinds of knowledge which can be discovered in databases are categorized as follows:

Mining Association Rules

A huge amount of data is stored electronically in most enterprises. In particular, in all retail outlets the amount of data stored has grown enormously due to barcoding of all goods sold.

For example, Wal-Mart, with more than 4000 stores, collects about 20 million point-of-sale transaction data each day. Given this mountain of data, it is good business sense to try to analyze it to find information. This is an interesting example of analyzing a large database of supermarket transactions with the aim of finding association rule. This is called Association Rules Mining or Market Basket Analysis. It involves searching for interesting customer habits by looking at associations. Association Rule mining has many applications other than market basket analysis, including marketing, customer segmentation, medicine, electronic commerce, classification, clustering, web mining , bioinformatics and finance**s.**

A.1 The Naïve Algorithm

To describe the association rules mining task informally with Naïve algorithm, we have taken the example of a small shop assuming that the shop sells only a small variety of products:

Bread	**Cheese**	**Coffee**
Juice	**Milk**	**Tea**
Biscuits	**Newspaper**	**Sugar**

It is assumed that the shopkeeper keeps record of what each customer purchases

Transaction ID	**Items**
10	Bread, Cheese, Newspaper
20	Bread, Cheese, Juice
30	Bread, Milk
40	Cheese, Juice, Milk, Coffee
50	Sugar, Tea, Coffee, Biscuits, Newspaper
60	Sugar, Tea, Coffee, Newspaper, Milk, Biscuits, Juice
70	Bread, Cheese
80	Bread, Cheese, Juice, Coffee
90	Bread, Milk
100	Sugar, Tea, Coffee, Milk, Juice, Newspaper

A simple example of Transactions

The shopkeeper wants to know which items are sold together frequently. We assume that the number of items in the shop stock is n and these items are represented by **I{i_1, i_2,………i_n}**.We denote transaction by **T{t_1, t_2………t_N}** each with a unique identifier(TID) and each specifying a subset of items from the item set I purchased by one customer. Each transaction of m items be **{i_1, i_2, …..,i_m}** with **$m \leq n$**. Now Find association relationship, given a large no. of transactions, such that items that tend to occur together are identified. It should be noted that association rules mining does not take into account the quantities of items bought. ***Association rules are often written as X → Y. X is often referred to as the rule's antecedent and Y as rule's consequent***. It indicates that only X and Y have been found together frequently in the given data and does not show a causal relationship implying that buying of X by a customer causes him/her to buy Y. Suppose item X and Y appear together in only 10% of the transaction but whenever x appears there is 80% chance that y also appears. ***The 10% presence of X and Y together is called support or prevalence of the rule 80% chance is called confidence or predictability of the rule.***

Essentially, the support and confidence are measure of the interestingness of the rule. A high level of support indicated that the rule is frequent enough for the business to be interested in it. A high level of confidence shows that the rule is true often enough to justify a decision based on it.

A.2 The Apriori Algorithm

This algorithm may be considered to consist of two parts. In the first part, those item sets that exceed the minimum support requirement are found. Such item sets are called frequent item sets. In the second part, the association rules that meet the minimum confidence requirement are found from the frequent item sets. The second part is relatively straightforward.

Transaction ID	Items
100	Bread, Eggs, Juice, Cheese
200	Bread, Cheese, Juice
300	Bread, Milk, Yogurt
400	Bread, Juice, Milk
500	Cheese, Juice, Milk

Table (i)- Transactions example

In table 1 we can see that Bread appears 4 times, cheese 3 times, juice 4 times, Milk 3 times and eggs and yogurt only ones. We require 50% support and therefore catch frequent items must appear in at least three transactions.

Items	Frequency
Bread	4
Cheese	3
Juice	4
Milk	3

Table - (ii) - Frequent item

In Table 2 there are two frequent item pairs which are {Bread, Juice} and {Cheese, Juice}.

Items Pairs	Frequency
(Bread, Cheese)	2
(Bread, Juice)	3
(Bread, Milk)	2
(Cheese, Juice)	3
(Cheese, Milk)	1
(Juice, Milk)	2

Table - (iii) - Candidate item pairs

Above item set leads to following possible rules.

Bread	↔	**Juice**
Juice	↔	Bread
Cheese	↔	Juice
Juice	↔	Cheese

The confidence of these rules is obtained by dividing the support for both items in the rule by the support for the item on the left hand side of the rule. The confidence of the four rules therefore are ¾=75%, ¾=75%, 3/3=!00% and 3/=75% respectively. Since all of them have a minimum 75% confidence, they all qualify.

Clustering Analysis

The process of grouping physical or abstract objects into classes of similar objects is called clustering or unsupervised classification. Clustering analysis helps construct meaningful partitioning of a large set of objects based on a “divide and conquer” methodology which decomposes a large scale system into smaller components to simplify design and implementation.

Pattern Based Similarity Search

Example of this type of database include: financial database for stock price index, medical database, band multimedia databases. When searching for similar pattern in a temporal or spatial-temporal database, two types of queries are usually encountered in various data mining operations:

1) ***Object-relative similarity query*** (i.e., range query or similar query) in which a search is performed on a collection of objects to find the ones that are within a user-defined distance from the queried object.

2) ***All-Pair similarity query*** (i.e., spatial join) where the objective is to find all the pair of element that is within a user-specified distance from each other.

Characterization

Data characterization is a summarization of general features of objects in a target class, and produce what is called characteristics rules. The data relevant to a user specified class are normally retrieved by database query and run through a summarization module to extract the essence of the data at different levels of abstractions. For Example, one may want to characterize the Our Video Store customer who regularly rent more than 30 movies a year. With concept hierarchies on the attribute describing the target class, the attribute oriented induction method can be used, for example, to carry out data summarization. Note that with a data containing summarization of data, simple OLAP operation fit the purpose of data Characterization.

Discrimination

Data discrimination produces what are called discriminant rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class. For Example one may want to compare the general characteristics of the customers who rented more than 30 movies in the last two years with those whose rental account is lower than 5.The technique used for data discrimination is very similar to the techniques used for data characterization with the exception that data discrimination results include comparative measures

V. The Issues in Data Mining

Data mining algorithms embody techniques that have sometimes existed for many years, but have only lately been applied as reliable and scalable tools that time and again outperform older classical statistical methods. While data mining is still in its infancy, it is becoming a trend and ubiquitous. Before data mining develops into a conventional, mature and trusted discipline, many still pending issues have to be addressed. Some of these issues are addressed below.

1. **Security and social issues:** Security is an important issue with any data collection that is shared and/or is intended to be used for strategic decision making. In addition, when data is collected for customer profiling, user behavior understanding, correlating personal data with other information, etc., large amount of sensitive and private information about individual and companies is gathered and stored. Moreover, Data mining could disclose new implicit knowledge about individual or groups that could be against privacy policies, especially if there is potential dissemination of discovered information.
2. **User interface issues:** The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. Good data visualization eases the interpretation of data mining results, as well as helps user better understand their needs. There are many visualization ideas and proposals for effective data graphical presentation.
3. **Mining methodology issues:** These issues pertain to the data mining approaches, applied and their limitation. It is often desirable to have different data mining methods available since different approaches may perform differently depending upon data at hand. Moreover, different approaches may suit and solve user's needs differently. Most algorithms assume the data to be noise-free.
4. **Performance issue:** Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the very large data sets data mining is dealing with today. Terabyte sizes are common. This raises the issue if scalability and efficiency of the data mining methods when processing considerably large data.
5. **Data Source issues:** There are many issues related to the data sources. Some are practical such as diversity of data types, while others are philosophical like the data glut problem. We certainly have an excess of data since we already have more data that we can handle and that we are still collecting data at an ever higher rate.

VI. Conclusion

Business decision making based on facts that is taking decisions intelligently, by extracting information and knowledge from transactional data. Data mining, is a tool that extract information from a huge data ,puzzling to the human mind, without prior assumptions or model. and ideally this tool should be part of BI, although data mining needs a lot of memory space and thus can be done in parallel then extracted information and knowledge shall be integrated into BI. In addition to the above, business intelligence can provide a pro-active approach, such as alert functionality that immediately notifies the end-user if certain conditions are met. For example, if some business metric exceeds a pre-defined threshold, the metric will be highlighted in standard reports, and the business analyst may be alerted via email or another monitoring service.

References

[1] R.Agarwal and R Srikant: “fast algorithms for mining association rules”, proc of intl. Conf on VLDB,1994.

[2] Ming-syan chen, Jiawei Han, Philip S.Yu: “Data mining :a overview from Database perspective”, IEEE transactions on knowledge and data engineering, vol8, No 6, december 1996.

[3] Sergey Brin, Rajeev Motwani, Jeffrey D.Ullman, and Shalom Tsur, Dynamic Itemset Counting and Implication Rules for Market Basket Data, Proceedings of the ACM SIGMOD Conference, PP255-264,1997.

[4] J.S.Park, M.S.Chen, and P.S.Yu : “an effective Hash-Based Algorithm for mining association rules”, proc. ACM SIGMOD intl conf.management of data, May 1995.

[5] Han J, Pei J and Yin Y 2000: “ Mining frequent patterns without candidate generation” proc. ACM-Sigmod Intl conf.management of Data.

[6] Usma Fayyad, G.P Shapiro, P. Smith, “The KDD process for extracting useful knowledge from volumes of data”, communications of the ACM.

[7] Ramakrishnan Srikant and Rakesh Agarwal, Mining Quantative Association Rules in large Relational Tables, Proceedings of the !996 ACM SIGMOD International Conference on Management of Data, PP 1-12 Montreal, Quebec, Canada, $-6 June 1996.

[8] Ramakrishnan Srikant, Fast Algorithms for Mining Association Rules and sequential Patterns, Ph.D Dissertation, 1996, Universities of Wisconsin, Madison.

[9] Jong Soo Park, Phillip S. Yu, Ming- Syan Chen: “Mining Association Rules with Adjustable Accuracy”,IBM Research Report,1997.

[10] Ashoka Savasere, Edward Omiecinski, and Shamkant B. Navathe, An Efficient Algorithm for Mining Association Rules in Large databases, Proceedings of the 2nd International Conference on Very Large Databases, PP. 432-444, Zurich, Swizerland, 1995.

Authors

First Author – RASHMI JHA,
NIELIT Center, Under Ministry Of IT, New Delhi, India,
Email-jharashmi21@gmail.com

She was born in Sitamarhi, Bihar, India. She has done

•M.Phil. in Computer Science from Manav Bharti University, H.P in 2012.

- Worked as a lecturer in BRBA University in Computer Science Dept. She is currently associated with NIELIT here handling different type of Govt. projects under ministry of Information Technology. She is doing her research work in developing algorithm for data mining.