International Journal of Engineering Innovation & Research
Volume 3, Issue 3, ISSN: 2277 – 5668

Propound Algorithms for Association Rules Mining with Reference to Some Applications

Rashmi Jha
NIELIT Center, Under Ministry of IT New Delhi, India
Email: jharashmi21@gmail.com

***Abstract* – 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.. We have been collected heterogeneous type of data, from simple numerical measurements and text documents, to more information such as spatial data, multimedia channels, and hypertext documents.**

***Keywords* – Data Mining, Association, Knowledge Discovery, Algorithm, Implicit, Integration, Pattern Discovery, Repository.**

I. INTRODUCTION

With the enormous amount of data stored in files, database, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision –making . Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases.

The Knowledge Discovery in Database process comprises of the following steps leading from raw data collections to same from of new Knowledge. The iterative process consists of the following steps:

1) Data Preprocessing: also known as data cleaning, it is a phase in which noise data and irrelevant data are removed from the collection.

2) Data Integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

3) Data Selection: The data relevant to the analysis is decided on and retrieved from the data collection.

4) Data Transformation: Also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

5) Data Mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.

6) Pattern Evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.

7) Knowledge Representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the Data Mining results

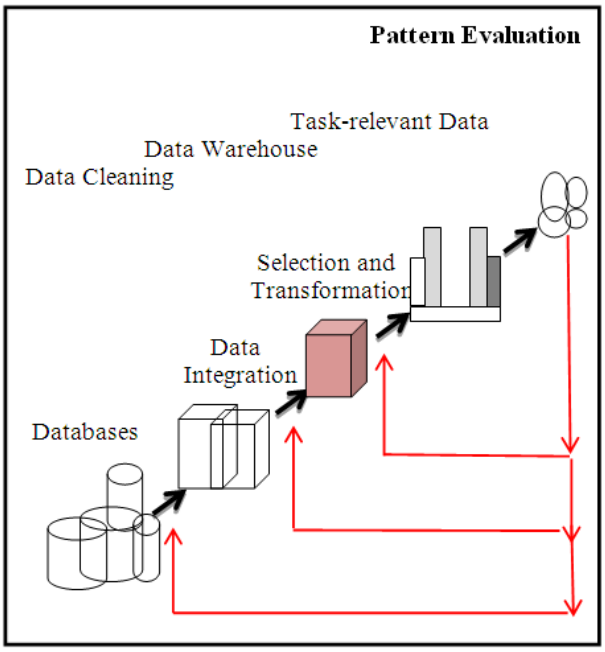

Fig.1. Data Mining is the core of Knowledge Discovery Process

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;

1) 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.

2) 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.

3) 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

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. For example, one may find from a large set of transaction data, such an association rule as if a customer buys(one brand of) milk, he/she usually buys(another brand of) bread in the same transaction. Since mining association rules may require to repeatedly scanning through a large transactions database to find different association patterns, the amount of processing could be huge, and performance improvement is an essential concern.

In general, the new association rule algorithm consists of four phases as follows:

1) Transforming the transaction database into the Boolean matrix.
2) Generating the set of frequent 1-itemsets L1.
3) Pruning the Boolean matrix.
4) Generating the set of frequent k-item sets Lk(k>1).

Transaction ID	Items
10	A, B, D
20	D, E, F
30	A, F
40	B, C, D
50	E, F
60	D, E, F
70	C, D, F
80	A, C, D, F

Fig.2 A simple representation of transaction as an item list

Transaction ID	A	B	C	D	E	F
10	1	1	0	1	0	0
20	0	0	0	1	1	1
30	1	0	0	0	0	1
40	0	1	1	1	0	0
50	0	0	0	0	1	1
60	0	0	0	1	1	1
70	0	0	1	1	0	1
80	1	0	1	1	0	1

Fig.3. Representing transaction as a binary item list

Data Classification

Data classification is to classify a set of data based on the values in certain attributes. For example, it is desirable for a car dealer to classify its customers according to their performance for cars so that the sales person will know when to approach, and catalogues of new models can be mailed directly to those customers with identical features so as to maximize the business opportunity. Data classification is the process which finds the common properties among a set of objects in a database and classifies, according to a classification model.

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

In this paper an idea is given to the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class level for some data. The latter is tried to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute value of the class. Prediction is however more often referred to the forecast of missing numerical values, or increase/decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Pages 207–216, Washington, DC, May 26- 28 1993.

[2] Agrawal, R., Imielinski, T., & Swami, A. (1993), "Mining Association Rules between Sets of Items in Large Databases", Proceedings of the ACM SICMOD Conference on Management of Data, pp.207-216, Washington, D.C.

[3] Han, J., Pei, J., & Yin, Y (2000), "Mining Frequent Patterns Candidate Generation" In Proc. 2000 ACM SIGMOD Int.Management of Data (SIGMOD'00), Dallas, TX.

[4] Berzal, F., Blanco, I., Sánchez, D. and Vila, M.A."Measuring the Accuracy and Importance of Association Rules: A New Framework" Intelligent Data Analysis, 6:221- 235, 2002.

[5] David A. Clausi, "An Analysis of Co-occurrenc Texture Statistics as a Function of Gray Level Quantization", Can. J. Remote Sensing, 28, No. 1, pp. 45-62, 2002.

[6] Bodon, F. "A Fast Apriori Implementation", In Proc. IEEEICDM Workshop on Frequent Item set Mining Implementations, 2003.

[7] Brijs, T. Vanhoof, K. and Wets, G., "Defining Interestingness for Association Rules", In Int. Journal Of Information Theories and Applications,10:4, 2003.

[8] Tung, A., Lu, H., Han, J., & Feng, L. (2003), "Efficient Mining of Inter transaction Association Rules", IEEE Transaction on Knowledge and Data Engineering, 15(1),43-56.

[9] Xu, Z. & Zhang, S. (2003), "An Optimization Algorithm Base on Apriori for Association Rules",Computer Engineering, 29(19), 83-84.

[10] 4th European Conference of the International Federation for Medical and Biological Engineering ECIFMBE 2008 23–27 November 2008 Antwerp, Belgium, 10.1007/978-3-540-89208-3_144, Jos Vander Sloten, Pascal Verdonck, Marc Nyssen and Jens Haueisen.

AUTHOR'S PROFILE

Rashmi Jha
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 in under ministry of Information Technology. She is doing her research work in developing algorithm for data mining.