INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

An Evolutionary Approach for Discovering Changing Frequent Pattern in Data Mining

TARUN DHAR DIWAN[1], KAMLESH LEHRE [2], VERTIKA KASHYAP [3]

[1]Dr.C.V.RAMAN UNIVERSITY, BILASPUR, INDIA
ASSISTANT PROFESSOR
DEPTT.OF ENGINEERING (CSE)
taruncsit@gmail.com

[2]DR.C.V.RAMAN UNIVERSITY, BILASPUR, INDIA,
ASSISTANT PROFESSOR
DEPTT.OF ENGINEERING (CSE)
lahrekamlesh@gmail.com

[3]DR.C.V.RAMAN UNIVERSITY, BILASPUR, INDIA,
M. TECH SCHOLAR

Abstract: *Data Mining is a study which is totally based on the historical database and the collection of huge data base is called as Data warehouse. In this study in the base of historical database we have to predict the future expectation and according that any we can take some decision. But there are so many difference type of data are available in the market and for to do frequent pattern data's there are so many association rule are available and according so many may algorithms are also available. But all algorithms are not suitable for all type of data. So there are some of the problems are raised for the decisions marking .After this study we are able to find that witch algorithm food for witch type of data. in the base this study the can pretend the future expectation and according able to take the decision related to the business. So the rezone behind the study to get the perfect algorithm for the particular type of data so we can take fast and perfect decision.*

Keywords: *Data Cleaning, Extracting Patterns, Verification, Machine Learning, Rule Association, Data Analysts.*

1. INTRODUCTION

The term data mining has mostly been used by statisticians, data analysts, and the Management Information Systems (MIS) communities. The phrase knowledge discovery in databases was coined at the first KDD) to emphasize that knowledge is the end product of a data-driven discovery. It has been popularized in the AI and machine-learning fields. In our view, KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data [1].The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper Interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data. Database techniques for gaining efficient data access, grouping and ordering operations.When data is accessed and optimized queries constitute the basics for scaling algorithms to larger data sets. Most data-mining algorithms from statistics, pattern recognition, and machine learning assume data are the main memory and they also pay no attention to how the algorithm breaks down if only limited views of the data are possible [2]. A related field evolving from databases is data warehousing, which refers to the popular business trend of collecting and cleaning transactional data to make them available for online analysis and decision support. Data warehousing helps set the stage for KDD in two important ways:

(1) Data Cleaning
(2) Data Access.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

2. RELETED WORK

Knowledge discovery in databases or data mining is an important research area in Computer Science.Since the number and the size of databases are rapidly growing. Data analysis underlies many computing applications, either in a design phase or as part of their on-line operations. As a primary tool of data mining, Association rule mining, one of the most important and well researched techniques of data mining, was first introduced[3,4]. It aims at extracting interesting correlations, frequent patterns, associated or casual structured among sets of items in the transaction databases or other data repositories.

The Interdisciplinary Nature of KDD

KDD has evolved, and continuously evolving from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, AI, knowledge acquisition for expert systems, data visualization, and high-performance computing [5].

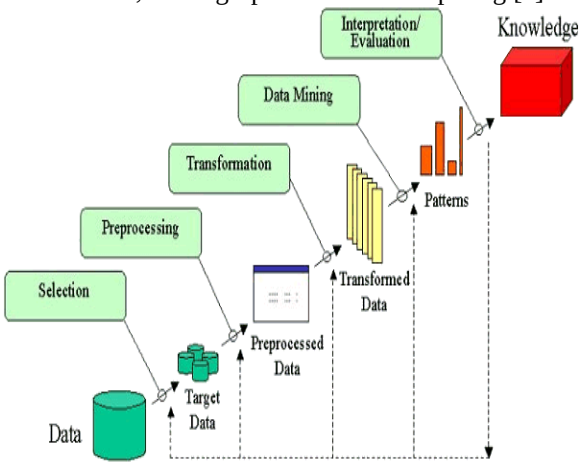


Figure 1: The Data-Mining Steps of the KDD Process

The data-mining component of the KDD process often involves repeated iterative application of particular data-mining methods. This section presents an overview of the primary goals of data mining, a description of the methods used to address these goals, and a brief description of the data-mining algorithms that incorporate these methods. The knowledge discovery goals are defined by the intended use of the system [6]. We can divide them into two types of goals:

1. Verification
2. Discovery

2.1 PURPOSE OF RESEARCH

Research and Application Challenges

Some of the current primary research and application challenges for KDD are as follows:

1. Larger databases
2. High dimensionality
3. Over fitting
4. Assessing of statistical significance
5. Integration with other systems

3. EXPERIMENT.DESIGN SPECIFICATION

Association Rule Problem

Definition 1: Let I = {I1, I2, … , Im} be a set of m distinct attributes, also called literals. Let D be a database, where each record (tuple) T has a unique identifier, and contains a set of items such that $T \subseteq I$ An association rule is an implication of the form $X \Rightarrow Y$, where X, $Y \subset I$, are sets of items called itemsets, and $XY = \phi$. Here, X is called antecedent, and Y consequent [7]. Two important measures for association rules, support (s) and confidence (α), can be defined as follows

Definition 2: The support (s) of an association rule is the ratio (in percent) of the records that contain XY to the total number of records in the database [8, 9].

Definition 3: For a given number of records, confidence (α) is the ratio (in percent) of the number of records that contain XY to the number of records that contain X.

C1	S1		S3	
C2		S2		
C3				S4
C4		S2	S3	S4
C5		S2	S3	
C6		S2	S3	
C7	S1	S2	S3	S4
C8	S1		S3	
C9	S1	S2	S3	
C10	S1	S2	S3	

Figure 2: Interpreting and Comparing Results

When comparing the results of applying association rules to those from simple frequency or cross-tabulation tables, we may noticed that in some cases

Vol. 1, Issue VII, Aug 2013
ISSN 2320-6802

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

very high-frequency codes or text values (items) are not the part of any association rule. This can sometimes be perplexing. To illustrate how this pattern of findings can occur, consider this example: Suppose we analyzed data from a survey of insurance rates for different automobiles manufacturers in America [10]. Simple tabulation would very likely show that many people drive automobiles manufacture by Ford, GM, and Chrysler; however, none of these may be associated with particular patterns in insurance rates, none of these brands may be involved in high-confidence, high-correlation association rules.Now linking them to particular categories of insurance rates.

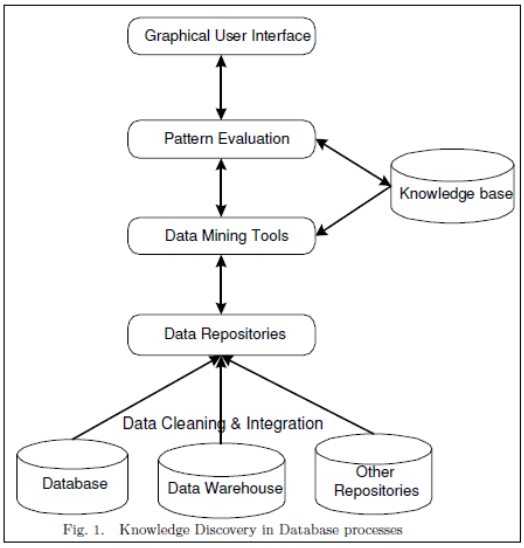


Figure 3: Mining Frequent Patterns

Mining frequent patterns is probably one of the most important concepts in data mining. A lot of other data mining tasks and theories stem from this concept [11, 12]. It should be the beginning of any data mining technical training because, on one hand, it gives a very well shaped idea about what data mining is and, on the other, it is not extremely technical. In this article we'll talk only about frequent patterns and specifically, about frequent item sets [13].

3.1 Patterns and Item sets

As we know that the minimal required information formulates a data mining problem. A very simple one for now. And we definitely won't try to solve it now. Our purpose for now is just to become accustomed to the main important concepts in data mining. Let's suppose we're conducting a data mining project for an insurance company that sells life insurance policies [14].

The supplemental benefits that this company offers are the following:

S1 Waiverof premium for disability benefit
S2 Disability income benefit
S3 Dismemberment benefit
S4 Accidental death benefit
S5 Waiver of premium for pay or benefit
S6 Terminal illness benefit
S7 Dread disease benefit
S8 Long term care
S9 Spouse and Children's Insurance Rider
S10 Children's Insurance Rider
S11 Second Insured Rider
S12 Guaranteed insurability benefit
S13 Paid-up additions option benefit

C1	S1		S3	
C2		S2		
C3				S4
C4		S2	S3	S4
C5		S2	S3	
C6		S2	S3	
C7	S1	S2	S3	S4
C8	S1		S3	
C9	S1	S2	S3	
C10	S1	S2	S3	

Figure 4: Customers with the Riders That They Bought

3.2 Frequent Item Sets

As you all have noticed, I didn't define the term of frequent. I didn't leave it; I did it on purpose because I didn't have enough information. But I'm going to do that now. A frequent itself is a set that occurs frequently. Don't smile, I know it sounds silly. But here comes the big question "How frequent is enough frequent?" How do we know what "frequent" means? 10 occurrences? 20? 100? Well, actually this is parameters that we have to set. If only 1 customer bought S1, S2, S3, S4 this fact isn't worth any consideration. We call it non frequent [15].

3.3 Closed Frequent Item sets

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

The only one downside of a maximal frequent itemset is that, even though we know that all the sub-item sets are frequent, we don't know the actual support of those sub-item sets. And we'll see how important this is when we'll try to find the association rules within the item sets. Keep that in mind for now and let's think of how to get all the frequent item sets that have the same support along with their subsets. This is how the Closed Frequent Itemsets came into picture: an item set is closed if none of its immediate supersets has the same support as the item set.

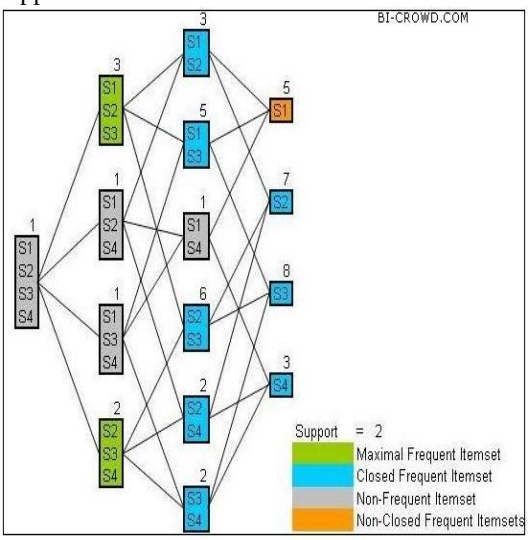


Figure 5: Finding These Closed Frequent Item Sets Can Be Of A Great Importance.

4. IMPLEMENTATION AND RESULT

This is a study which is based on basically three algorithms FP-Tree, Apriori, P-T Tree in this we changed the confidence and support and in the base of different confidence and support we write the reading and accordingly there farm graph.

Data set: - Pima Indian

Support	**20**	**30**	**40**	**50**	**60**	**70**	**80**
FP-Tree	2.060	1.660	1.280	0.930	0.770	0.650	0.620
Apriori	99.290	68.8	58.7	2.79	1.07	0.64	0.56
P-T Tree	0.511	0.51	0.500	0.480	0.480	0.470	0.470

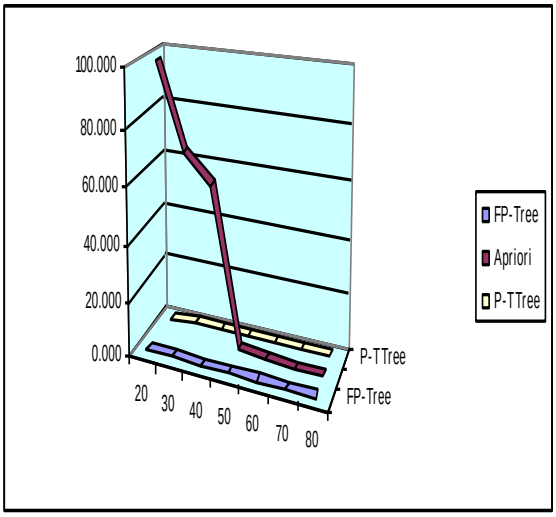


Figure 6: The Dance data we have use either FP Tree or P-T Tree.

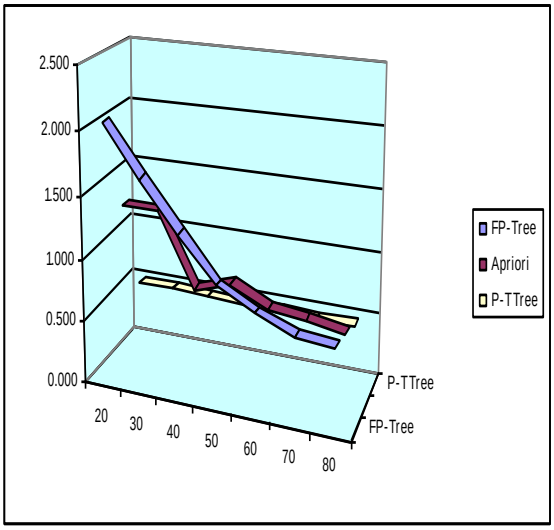
Here two tables two graphs are available respectively, in the first table we have take Pima Indian data set which is Dance data and the Adult is a Sparse data second table is farms in the base of that sparse data.As per research work we find that Apriori algorithm can perform well with the sparse data and FP-tree and P-T Tree both will perform well with the Dance data.The conclusion is that whenever we have sparse data at that time we have to use Apriori Algorithm

5. CONCLUSION

Here two tables and two graphs are present.In the first table we have taken Pima Indian data set which is Dance data and the Adult is a Sparse data second

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

table is farms in the base of that sparse data as per this research work we find that Apriori algorithm can perform well with the sparse data and FP-tree and P-T Tree both will perform good with the Dance data.The conclusion is that whenever- we have sparse data we have to use Apriori Algorithm and in the Dance data we have use either FP Tree or P-T Tree.



6. FUTURE WORK

There are so many Data's available in the market but here we are using two type of datasets pima Indian (Danaces data) and other is Adult (Sparse Data) in the future we can use another sets also which can be related from the Bio and science stream data sets. There are some database and so many more algorithms. This is a very vast filed for the study. So our target is to find some algorithm for the exact data. This can help us to take fast decision in future prediction which can be helpful for the business also.

REFERENCES

[1] Aggarwal, C.C., Procopiuc, C., Wolf, J.L., Yu, P.S., Park, J. S.: Fast Algorithms for Projected Clustering. In: Proc. 1999 ACM SIGMOD Int'l. Conf. on Management of Data (SIGMOD 1999), Philadelphia, Pennsylvania (June 1999)

[2] Blockeel, H., De Raedt, L., Ramon, J.: Top-down induction of logical decision trees. In: Proc. Fifteenth Int'l. Conf. on Machine Learning (ICML 1998), Madison, WI (July 1998)

[3]Dzeroski, S.: Inductive logic programming and knowledge discovery in databases.In: Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park (1996)

[4]Dzeroski, S.: Multi-relational data mining: an introduction. ACM SIGKDD Explorations Newsletter 5(1), 1–16 (2003)

[5] Fogaras, D., R´acz, B.: Scaling link-base similarity search. In: Proc. 14th Int'l. Conf. World Wide Web, China, Japan (May 2005)

[6] Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Computing Surveys 31, 264–323 (1999)

[7]Jeh, G., Widom, J.: SimRank: A measure of structural-context similarity. In: Proc. Eighth Int'l. Conf. on Knowledge Discovery and Data Mining (KDD 2002), Edmonton, Canada (July 2002)

[8] Kirsten, M., Wrobel, S.: Relational Distance-Based Clustering. In: Page, D.L. (ed.)ILP 1998. LNCS, vol. 1446, Springer, Heidelberg (1998)

[9] Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46(5), 604–632 (1999)

[10] Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking.bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998)

[11] Quinlan, J.R., Cameron-Jones, R.M.: FOIL: A midterm report. In: Brazdil, P.B. (ed.) ECML 1993. LNCS, vol. 667, Springer, Heidelberg (1993)

[12] Wang, J.D., Zeng, H.J., Chen, Z., Lu, H.J., Tao, L., Ma, W.Y.: ReCoM Reinforcement clustering of multi-type interrelated data objects. In: Proc. 26th Int'l. Conf.on Research and Development in Information Retrieval, Toronto, Canada (July 2003) Exploring the Power of Heuristics and Links 27

[13] Yin, X., Han, J., Yu, P.S.: LinkClus: Efficient Clustering via Heterogeneous Semantic Links. In: Proc. 32nd Int'l. Conf. on Very Large Data Bases (VLDB 2006),Seoul, Korea (September 2006)

[14]Yin, X., Han, J., Yu, P.S.: CrossClus: User-guided multi-relational clustering. Data Mining and Knowledge Discovery 15(3), 321–348 (2007)

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

[15] Yin, X., Han, J., Yu, P.S.: Truth Discovery with Multiple Conflicting Information Providers on the Web. In: Proc. 13th Intl. Conf. on Knowledge Discovery and Data Mining, San Jose, CA (August 2007).