IJCSE International Journal of Computer Sciences & Engineering Open Access
Review Paper Volume-3, Issue-6 E-ISSN: 2347-2693

A Comparative analysis of Association rule excavating in Big Data Mining Algorithms

Ahilandeeswari. G[1], DR.R.Manicka Chezian[2*]

[1,2*] *Department of Computer Science, NGM College, Pollachi, India*

ahilaplatinum@gmail.com, *chezian_r@yahoo.co.in*

www.ijcseonline.org

Received: May/26/2015 Revised: June/04/2015 Accepted: June/24/2015 Published: June/30/ 2015

Abstract— In Data Mining Research, Association rule mining plays a significant role in data mining. This paper presents the review of Association rule mining. The analysis of research survey would give the instruction concerning somewhat has been done previously in the same area, what is the present tendency and what are the other related areas. Big data is the word for a set of data sets which are enormous and convoluted, it holds structured and unstructured both varieties of data. Data comes from everywhere, sensors used to amass climate information, posts to social media sites, digital pictures and videos etc. This data is known as big data. Useful data can elicit from this big data with the help of data mining. In this paper, the association rule of data mining and advanced big data mining algorithms are scrutinized.

Keywords : Association rule, Apriori Algorithm Big data mining, Data mining

I. INTRODUCTION

In many applications in the real world, generated data are of huge concern to the stakeholder as it delivers essential information / knowledge that assists in making predictive analysis. This information helps in modifying positive decision parameters of the application that changes the overall outcome of a business process. The amount of data, collectively called datasets, generated by the function are very large. So, there is a need of processing large datasets efficiently. The dataset collected may be from heterogeneous sources and may be structured or unstructured data. Processing such data generates helpful patterns from which observation can be extracted. He simplest looms is to use this template and insert headings and text into it as appropriate.[1]

Data mining is the method of verdict correlations or patterns among fields in large datasets and building up the knowledge-base, based on the given constraints. The overall destination of data mining is, to my knowledge from an existing data set and remake it into a human-understandable structure for further use. This procedure is often referred to as Knowledge Discovery in datasets (KDD). The development has revolutionized the advance of solving the multipart real-world problems. The KDD process consists a series of tasks like selection, pre-preprocessing, transformation, data mining and interpretation.[3] The implied information

Within databases, and mainly the interesting association relationships among sets of objects, that guide to association rules, may disclose useful patterns for decision support, financial forecast, marketing policies, even medical diagnosis and many other applications. This fact attracted a lot of attention in current data mining research. As shown in, mining association rules may require iterative scanning of large databases, which is costly in processing. Many researchers have focused their work on capable mining of association rules in databases. A very influential association rule mining algorithm, Apriori has been developed for rule mining in large transaction databases. Many other algorithms developed are derivative and/or extensions of this algorithm.[4] A major step forward in improving the performances of these algorithms was made by the introduction of a novel, compact data structure, referred to as a frequent pattern tree, or FP-tree [2], and the associated mining algorithm, FP-growth.The main disparity between the two approaches is that the Apriori-like techniques are based on a bottom-up innovative of frequent itemset combinations and the FPtree based ones are partition-based, divide-and-conquer methods.We are living in the world of data. Whether it is data on any social networking site on which we are sharing our photos, videos regularly, or it is our favored shopping site, or the knowledge bank from where we obtain any in sequence at any moment such as Wikipedia. In recent 10 years, the progress in the Information skill makes easy to generate and update data continuously. For example, there are so many photos, videos are uploaded daily. [5][6]Even rapid growth in internet and cloud computing

The main disparity between the two approaches is that the Apriori-like techniques are based on a bottom-up innovative of frequent itemset combinations and the FPtree based ones are partition-based, divide-and-conquer methods.

We are living in the world of data. Whether it is data on any social networking site on which we are sharing our photos, videos regularly, or it is our favored shopping site,

Corresponding Author: *Ahilandeeswari.G, ahilaplatinum@gmail.com, Department of Computer science, NGM College, Pollachi India*

or the knowledge bank from where we get any information at any time such as Wikipedia. In recent 10 years, the progress in the Information skill makes easy to generate and update data continuously. For example, there are so many photos, videos are uploaded daily. [5][6]

Even rapid growth in internet and cloud computing also played a vast role in mounting data. Data is constantly generated by use of internet as well as by companies who have generated and updated a large amount information from sensors, computers, automated processes.A current study estimated that every minute, Google receives over 2 million queries, email users send over 200 million messages, YouTube users upload 48 hours of video, Facebook users share over 680,000 pieces of content, and Twitter users generate 100,000 tweets.[1]Dataset or Database includes large amounts of data. But sometimes not whole data are essential, sometimes we need to haul out only specific or in other words, the only useful information from the dataset as per our prerequisite or analysis purpose. Methodology, so that we can uncover veiled information and insights from dataset effectively. Such efficient processor or tactic called data mining.[6]

II. DATA MINING FOR BIG DATA

The Big Data is nothing but a data, available at heterogeneous, autonomous sources, in extreme large amount, which get updated in fractions of seconds. For example, the data stored on the server on Facebook, as most of us, daily use the Facebook; we upload various types of information, upload photos. Get all the data stored in the data warehouses at the server on Facebook. This data is nothing but the big data, which is so called due to its complexity. Also, another example is storage of photos at Flicker. These are the good real-time examples of the Big Data. Another good example of Big data would be, the readings taken from an electronic microscope of the universe. Now the term Data Mining, Finding for the exact useful information or knowledge from the collected data, for future actions, is nothing but the data mining. Revenue, cut costs, or both.. Data mining as a term used for the specific classes of six activities or tasks as follows:

TABLE 1 : STATE OF BEING ACTIVE AND OFTEN ASSIGNED IN DATAMINING ALGORITHMS

State Of Being Active	Often Assigned	Algorithms
Classification	Classification is a process of generalizing the data according to dissimilar instances.[1]	Decision tree, k-nearest neighbor classifier, Naive Bayes, Apriori and AdaBoost
Estimation	Assessment deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous Variables such as income, height or credit card balance	-------
Association Rules	An association rule is a rule which implies certain association relationships among a set of objects in the database	-------

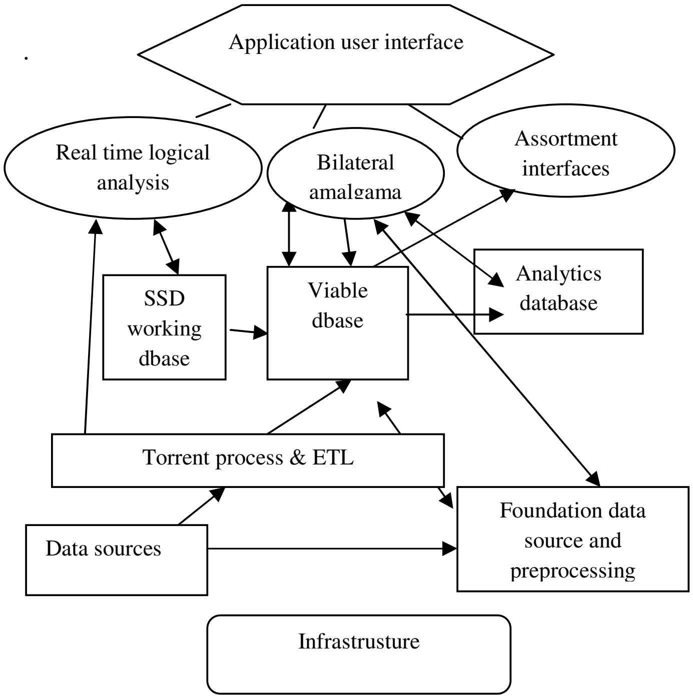

Figure 1: Deliberate Bigdata Reference Architecture

Big Data is carried out computing, on the PB (Petabyte) or even on EB (Exabyte) data with multipart computing process, so the lateral computing framework, programming language backing and software model utilizing to efficiently analyze and mine distributed data. MapReduce structure is suitable for large scale data mining task on clusters.[7]

TABLE 2:DATA MINING ACCOMBINIED BY BIG DATA

Data mining Scruntize	Big data Excavation
Data mining refers to the action going through large data set to gaze for related information	Big data is a tenure for large data set.
Data mining is the handler which provides beneficial results.	Big data is the asset
Data mining refers to the procedure that involve a relatively sophisticated look for operation[6]	Big data" varies depending on the capabilities of the association managing the site, and on the capabilities of the applications that are traditionally used to process and examine the data.

III ASSSOCIATION RULE IN DATA MINING AND BIG DATA MINING

Association rules are if/then statements that help to divulge relationships between unrelated data in a database, relational database or other information depository. Association rules are used to find the relationships between the objects which are frequently used together. Applications of association rules are basket data analysis, classification, cross-marketing, clustering, catalog design, and loss-leader analysis, etc. For example, if the customer buys bread then he may also buy butter. If the customer buys a laptop, then he may also buy a memory card. There are two basic norm that association rules uses, support and confidence.[5][6] It identifies the relationships and rules generated by analyzing data for frequently used if/then patterns

$$\text{Support} = \frac{\text{Frq (X, Y)}}{\text{N}}$$

$$\text{Confidence} = \frac{\text{Frq (X, Y)}}{\text{Frq (X)}}$$

A) AIS ALGORITHM

This algorithm is used to find frequent item sets. It uses candidate generation in order to detect them.[8].AIS algorithm consists of two phases.

Phase1:

- *The cohort of the frequent itemsets.*
- *This is followed by the creation of the confident and frequent*

Phase2:

- *Association rules in the second phase.*

One of the disadvantages of this algorithm includes the generation and counting of too many candidate item sets that turn out to be small. This was the first algorithm to propose the crisis of generation of association rules.[3][4] The flaw of the AIS algorithm is that it makes multiple passes in the glut of the database. Furthermore, it generates and counts too many candidate itemsets that twist out to be small, which requires more space and waste much effort that turned out to be useless

B) SETM ALGORITHM

Just like the AIS algorithm, this algorithm also does on the fly counting. It is based on the transaction read from the database. But SETM was created for SQL and uses relational transactions.[4]Whereas SETM uses the standard SQL join procedure for the generation of candidates and next separates candidate generation from counting. The first candidates are generated using equi-joins and then sorted and then the ones that don't meet the minimum support are removed.[5]

C) APRIORI ALGORITHM

This algorithm is very frequently used for mining of frequent item sets and to detect associations[3][8]

Approach of Apriori Algorithm

```
• Ck : Set of candidate itemsets of size k
• Fk : Set of frequent itemsets of size k
F1 = {large items}
for ( k=1; Fk
!= 0; k++) do {
Ck+1 = New candidates generated from Fk
for each transaction t in the database do
Increment the count of all candidates in Ck+1 that
are contained in t
Fk+1 = Candidates in Ck+1 with minimum support

}
```

D) APRIORI TID ALGORITHM

Just like the Apriori algorithm, AprioriTID algorithm uses the generation objective in order to resolve the candidate item sets.[1][2]

Method of Apriori TID

```
1.  L1 = ( large 1-itemsets);
2.  C1'= database D;
3.  for (k=2; Lk-1 ; k++) do begin
4.  Ck = apriori-gen(Lk-1); //New candidates
5.  Ck'= ;
6.  For all entries t Ck-1' do begin
7.  //Decide candidate itemsets in Ck limited in the
transaction with identifier t€TID
i.    Ct={c€C|(c-c[k])=1t.set-of-itemsets(c-c[k-1])=1t.set-
of-itemsets };
b. for all that candidates c€Ct do
i.    c.count++;
c. if (Ct ) then Ck'+= < t.TID, Ct>;
8. end;
a. Lk = {c | Ck c.count  minsup}
9. End
10. Answer = Uk Lk
```

Two Pass of AprioriTID

First pass

- *The database is not used at all for including the support of candidate itemsets behind the first pass.*
- *The candidate itemsets are generated the similar way as in Apriori algorithm.*

Second pass

- *An additional set C' is generated of which every member has the TID of both transaction and the large itemsets present in this transaction. This set is used to count the support of all candidate itemset.*
- *The benefit is that the figure of entries in C' may be smaller than the integer of transactions in the database, particularly in the later phases.*

E) APRIORI HYBRID ALGORITHM

It is not necessary to use the same algorithm in all the passes over the data.

Initial pass:Apriori performs better

Later pass:AprioriTId performs better

Figure 2 shows the execution times for Apriori and AprioriTid for different passes over the dataset Apriori does better than AprioriTid.[5][10].

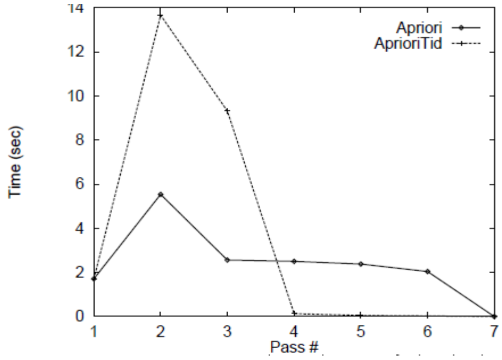

Figure 2: Per pass Execution Times of Apriori and AprioriTid (Minsup = 0.75 seconds)

F) ECLAT ALGORITHM

Eclat application represents the set of connections as a bit matrix and intersects rows give the support of item sets. It follows a depth first traversal of a prefix tree.[6]

G) RECURSIVE ELIMINATION ALGORITHM

Recursive eliminations are based on a step by step elimination of items as of the deal database composed with a recursive dispensation of transaction subsets.[4][5] The basic operations of the RElim algorithm.The rightmost list is traversed and reassigned: once in an initially empty list array (conditional database for the prefix, see top right) and once in the original directory array (eliminating itemsets, see bottom left). These two databases are then both processed recursively.[4]

STEPS:

1. Load transactions (in memory).
2. Count item frequencies.
3. Delete all rare items from the transactions.
4. Sort each transaction according the items frequency.
5. Create recursive elimination data structure.

H) FP GROWTH ALGORITHM

FP-Tree frequent pattern mining is used in the evolution of association rule mining. FP-Tree algorithm conquered the dispute found in Apriori algorithm. By avoiding the candidate generation approach and less passes over the database, FP-Tree was found to be faster than the Apriori algorithm [9]. An FP-Tree is a prefix tree for transactions. Every node in the tree represents one item and each aisle represents the set of transactions that involve with the appropriate item.[5][6]

The method of FP-growth (Tree T, A)

```
If Tree T contains a single path P,
Then for each combination of the nodes in the
Path P do Generate pattern B U A with
Support = minimum support of nodes in B
Else for each Hi in the header of the
Tree T do
{
Generate pattern B=Hi U A with
Support = Hi. Support;
Construct bs conditional pattern base
And bs conditional FPTree
That is B;
If Tree B ≠Ø
Then call FP-growth (Tree B, B)
}
```

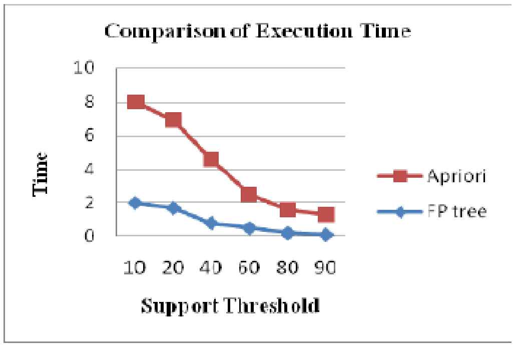

Figure 3: Comparison of Execution Time

I) COMPARISON TABLE

Table 3 shows the comparative study of assorted association rule mining algorithms. The classification is based on the features such as the uses, merits and demerits of each. The. The main purpose of the table is to highlight the application of all the above stated algorithms.[9]

TABLE 3 : REPRESENTS A PROPORTIONAL STUDY OF ASSOCIATION RULE MINING ALGORITHMS WITH MERITS AND DEMERITS

Author	Algorithm	Application	Merits	Demerits
Aggarwal, R., and Srikant	AIS	Not frequently used, but when used is used for small problems.[1]	1. Better than SETM. 2. Easy to use	1. The Candidate sets generated on the fly. 2. Size of candidate sets large
Aggarwal, R., and Srikant	SETM	Not frequently used. [1]	1. Separates generation from counting.	1. Very large execution time. 2. Size of candidate sets large
Borgelt, C.	Apriori	Best for closed item sets[3]	1. Fast 2. Less candidate sets.	1. Fast 2. Less candidate sets.
Khurana, K., and Sharma	AprioriTID	Used for less significant problems [2]	1. Doesn't use the whole catalog to count candidate sets. 2. Better than SETM. 3. Better than Apriori for small databases. 4.Time saving.	----
Khurana, K., and Sharma	Apriori Hybrid	Worn where Apriori and AprioriTID used. [2]	Better than both Apriori and AprioriTID.	------
Thieme, S.L	Eclat	Best use of free item sets.	1. Less memory procedure. 2. Lower minimum support. [4]	Apriori wins in cases where candidate sets are more
Borgelt, C.	Recursive Elimination	---------	1. Better than Apriori in all cases. [3]	Less than éclat in all cases
Hunyadi, D., Borgelt, C.	FP Growth	Used in cases of large problems as it doesn't require generation of candidate sets [3][5]	1. Only 2 passes of the dataset. 2. Compresses data set. 3. No candidate sets generation required so better than éclat, Apriori	Using tree the structure creates complexity

TABLE 4 : EVALUATION OF ASSOCIATION RULE MINING ALGORITHMS

Characteristics	Ais	Setm	Apriori	Aprioritid	Apriori Hybrid	Fp Growth
Data support	Less	Less	Limited	Often Suppose Large	Very Large	Very Large
Speed in the initial phase	Slow	Slow	High	Slow	High	High
Speed in the later phase	Slow	Slow	Slow	High	High	High
Accuracy	Very Less	Less	Less	Most Accurate	More Accurate than Apriori	More Accurate

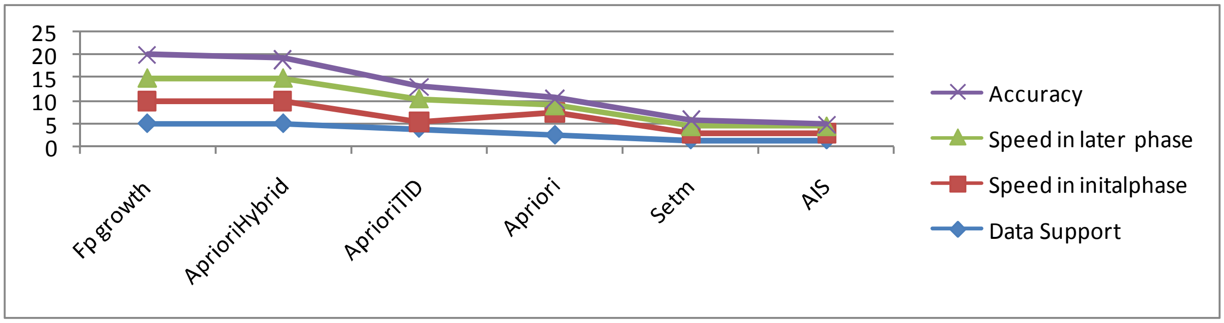

Figure 4: Comparison of Association rule excavation Algorithms

IV. ADVANCED TECHNIQUES IN BIG DATA MINING ALGORITHMS

Big data mining is used to resolve knowledge or extract patterns from the huge data set or Big data set which means that the size of the data set is very big so that it afford the practical in sort the distant algorithms are used for big data mining.

In this paper the algorithms surveyed are Two-Phase Top-Down Specialization(TPTDS)approach, Tree-Based Association Rules(TARs), Fuzzy C-Means (FCM) algorithm and Associate Rule Mining (ARM) algorithm, and correlation of these algorithms each algorithm is used for mining from offbeat sources.[9][10]

TABLE 5: COMPARISON OF BIG DATA MINING ALGORITHMS

Algorithm/Approach	Performance Criteria	Usage
Two-phase Top-Down Specialization (TPTDS) approach	Execution time and Information Loss	Privacy Preservation of data
Tree-Based Association Rules (TAR) approach	Extraction time and Answer time	Mining from semistructured (XML) document
Fuzzy c-Means (FCM) algorithm	Run time	Clustering of data
Association Rule Mining (ARM)	Comorbidity	Mining of ICU data

The table shows the algorithm comparison.Size of data includes Megabyte, Gigabyte, Terabyte etc…., and the content of data includes both structured and as well as unstructured data.[5][6]

Each of the algorithms satisfies the different purpose, the algorithm done the mining process from the different data sources their performance criteria or factors and usage are listed in the table

V. COMPARISON BETWEEN TRADITIONAL DATA MINING ALGORITHMS WITH ADVANCED BIG DATA MINING ALGORITHMS FORMING BIGDATA.

In this segment, we have introduced scope and new directions for delving into in BIG data mining as per the data mining challenges with BIG data.[3]

Data mining Algorithms	Problems with BIG data	Big data mining algorithms	Result
Apriori	Memory rations	MREclat	Improves routine, sufficient memory
FP growth	Inadequate computational Capabilities	DISEclat	High scalability and good accelerate
Eclat	Unbiased data Sharing,inter communication cost	BigFIM	Improves performance , high speed up, competent pattern mining for BIG data
SVM	Memory necessities	SVM with Map Reduce	Reduce training time and computation time increases the recital
C4.5	A sub-tree can be imitation numerous times	C4.5 with Map Reduce	Minimize the communications cost, reduce the execution time
KNN	High computation fee	KNN with Map Reduce	Diminish communications cost and increase performance
Naïve Bays	Strong characteristic autonomy assumptions, low performance in large dataset	Naïve Bays with MapReduce	Reduce time complexity, capability to process Large data

TABLE 6 : DATA MINING TECHNIQUES WITH BIG DATA

VI. COMPARISON BETWEEN CONVENTIONAL DATA MINING ALGORITHMS AND BIG DATA MINING ALGORITHMS

This table represents some main data mining techniques to their purposes and problems occurred while dealing with BIG data, assessment between traditional data mining algorithms and BIG data mining algorithms.[4]

TABLE 7: COMPARISON FLANKED BY CONVENTIONAL DATA MINING AND BIG DATA MINING ALGORITHMS

Data Mining Techniques, And Purpose	For Normal Dataset	Problems With Big Data
Frequent pattern mining: Mining patterns that occur frequently in the dataset	Produce competent results in short time	Reminiscence requirements and Minimum frequency threshold & speed
Classification: Classify the dataset using known class labels for effective data analysis	Categorize the dataset accurately	Preprocessing, facet extraction, training or learning, accuracy, Time complexity
Clustering: To make group of objects from based on similarity or dissimilarity	Efficiently detachment and categorize dataset	Heterogeneity of data and preprocessing
Outlier Analysis: To classify rare and abnormal data from the dataset	Detects outliers in fast and competent way	Volume, Veracity and Value

VII CONCLUSION

The overall goal of this mining process helps to form the association rules for further use. Association rules prove to be the most successful method for frequent pattern matching over a decade. This paper gives a brief survey of association rule mining algorithms and advanced techniques of Big data mining algorithms

REFERENCES

[1] R. Agarwal and R. Srikant,"Fast Algorithms for Mining Association Rules.", International Conference on very large Databases, proc 20th, pp 487-499, June 1994.

[2] Khurana K and Sharma S, —A comparative analysis of association rule mining algorithms, International Journal of Scientific and Research Publications, Volume 3, Issue 5, pp 38-45, May 2013.

[3] Borgelt, C. **"**Efficient Implementations of Apriori and Eclat". Workshop of frequent item set mining implementations (FIMI 2003, Melbourne, FL, USA).

[4] Hunyadi, D."Performance comparison of Apriori and FP-Growth Algorithms in Generating Association Rules".Proceedings of the European Computing Conference ISBN: 978-960-474-297-4.

[5] Thieme, S.L. "Algorithmic Features of Eclat". FIMI, Volume 126 of CEUR Workshop Proceedings, CEUR-WS.org, 2004.

[6] Ms. Dhamdhere Jyoti L., Prof. Deshpande Kiran B. "An Effective Algorithm for Frequent Itemset Mining on Hadoop.", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 8, August 2014.

[7] Agrawal, R., Shafer, J.C., "Parallel mining of association rules.", IEEE Transactions on Knowledge and Data Engineering, Volume.8, no.6, pp 962-969, Dec 1996.

[8] A. Swami, T. Imielienski, R. Agrawal," Mining Association Rules between Sets of Items in Large databases.", ACM Press, pp 207–216, July 1993

[9] Brain.s Motwani R,Ullman.J.D and S. Tsur,"Dynamic itemsets counting and implication rules for market basket analysis.", ACM-SIGMOD ,pages 255-264, May 1997.

.

[10] Ferenc Kovacs and Janos Illes "Frequent Itemset Mining on Hadoop.",IEEE 9th International conference on Computational Cybernetics, Volume 2 Issue 4, June 2013.

AUTHORS BIOGRAPHY:

MS Ahilandeeswari.G received an MCA degree from Anna University,Chennai. In 2006 and 2009 respectively,currently a research scholar at Department of Computer Science,NGM College, Pollachi.Her research interest lies in the area of Data Mining and Big Data Mining

DR. R.Manicka chezian received his M.Sc., degree in Applied Science from P.S.G College of Technology, Coimbatore, India in 1987. He completed his M.S. degree in Software Systems from Birla Institute of Technology and Science, Pilani, Rajasthan, India and Ph D degree in Computer Science from School of Computer Science and Engineering, Bharathiar University,Coimbatore, India. He served as a Faculty of Maths and Computer Applications at P.S.G College of Technology, Coimbatore from 1987 to 1989. Presently, he has been working as an Associate Professor in N G M College (Autonomous), Pollachi under Bharathiar University, Coimbatore, India since 1989. He has published more than 120 papers in various International Journals / Conferences. His research focuses on Network Databases, Data Mining, Distributed Computing, Mobile Computing, Real Time Systems and Bio-Informatics.