Performance Measure of Similis and FP-Growth Algorithm

Archana Singh
Amity University
AIIT, I1 Block, 3rd floor
Sec-125, Amity University

Jyoti Agarwal
Amity University
ASET, E-2 Block
Sec-125, Amity University

Ajay Rana, PhD.
Amity University
ASET, E-2 Block, Ground floor
Sec-125, Amity University

ABSTRACT
Exploration, analysis of data and to know patterns from large data repository has become the need of an hour. Data Mining Technology provides the solution to meet the market trends. Mining association rule is one of the main application areas of Data Mining. It gives a set of customer transactions on items; the aim is to find correlations between the sales of items. At present there are various Association Rules Algorithms are in market. This paper define the survey done on various algorithms of Association Rules of Data Mining and also compare two main algorithms-Similis Algorithm and FP-Growth Algorithm depending upon the different criteria

General Terms
Data Mining Association Rules Algorithms

Keywords
Association Rule Mining, Data Mining, Market Basket, Adjancey Matrix, Graph etc.

1. INTRODUCTION
Data Mining is also called as Knowledge discovery in database (KDD).It helps to explore, analyze and then finally retrieve the data. Data Mining has many application areas. The Most important areas are Neural Networks, Association rules in Market Basket Analysis, Data Visualization, Rule Induction, Logistic Regression etc. Here the main focus is on Market Basket Analysis Association Rules using Data Mining technique. The term Market Basket or commodity bundle refers to a fixed list of items used specifically to track the progress of inflation in an economy or specific market.

1.1 Market basket Analysis Method
There are several methods to do market basket analysis-

1.1.1 Association Rules Mining-

Association Rules was introduced by Agarwal et al in 1993. ***Association Rule Mining*** *is a popular and well researched method for discovering interesting relations between variables in large databases. Mining association rule is one of the main application areas of Data Mining. It gives a set of customer transactions on items; the aim is to find correlations between the sales of items.*

Association rule is often written as X->Y meaning that whenever X appears Y also tends to appear.

1.1.2 *Association Rule strength Measures-*
Association Rules Mining measures are-

1.1.2.1 Support(X): Supp(X) **o**f an item set X is defined as the proportion of transactions in the data set which contain the item set.

Support(X) = (Numbers of times X appears)/N=P(X)

Support (XY) =

(Numbers of times X and Y appear together)/N=P (X∩Y)

1.1.2.2 Confidence: The confidence rule is defined by

Conf(X->Y) = P (X∩Y)/□P(X) =P (Y|X)

*1.1.2.3 Lift:-*Lift is used to measure the power of association between items that are purchased together. Lift must be above 1.0 for the association to be of interest.

Lift is defined as-P (Y|X)/P(Y)

1.1.2.4 Conviction: The conviction of a rule is defined as-.

$$\mathrm{conv}(X \Rightarrow Y) = \frac{1 - \mathrm{supp}(Y)}{1 - \mathrm{conf}(X \Rightarrow Y)}$$

It defines the ratio of the expected frequency that X occurs without Y.

1.1.3 Goals of Association Rule Mining-
The goal of Association rule Mining is to find all rules having-

Support ≥ minsup threshold
Confidence ≥ minconf threshold

Rules that satisfy both minsup and minconf are called strong rules.

1.1.4 Process for Association rules generation
Find all frequent item sets- Generate all item sets whose support >= minsup and then Generate strong association rules from the frequent item sets-Generate strong rules from each frequent item set. These rules must satisfy minsup and minconf.

2. LITERATURE REVIEW
Various research papers from different journals and conferences like ACM, Springer, IEEE etc related to Association Rule Algorithms have been collected and studied. The literature survey has two main parts. In the first part two important algorithms(Simlis and FP-Growth) is discussed and in the second part a survey table (table 1)is created which help

in understanding the work done by different authors on Association Algorithm. Table is used to make understanding easier. The table contains various fields like Title name, journal name, objective (Focus) and result analysis.

2.1 Algorithm Description

2.1.1 Similis Algorithm [3]: The name Similis [3] is derived from Latin word which meaning is similar. Similis algorithm is presented in two steps

(a) Transformation step

(b)Search step

STEP 1 – Data Transformation

Input: support measure, table T (transaction, item);

Output: weighted graph G (V, E);

- Discard the infrequent 1-itemset using a filter;
- Generate graph G(V,E) using the 2-itemset frequency;

STEP 2 – Finding the Maximum-Weighted Cliques

Input: weighted graph G (V, E) and size k;

Output: weighted clique S of size k that corresponds to the most frequent k-item set;

- Find in G (V, E) the clique S with k vertexes with the maximum weight, using the Primal- Tabu Meta-heuristic shown in Fig.1

Fig.1 Primal-Tabu Meta-Heuristic for the Maximum-Weighted Clique

Primal-Tabu Meta-Heuristic for the Maximum-Weighted Clique:-

Input: weighted graph G(V,E), size k;

Output: maximum-weighted clique S* with size k;

Initialize S, S* e Tabu;

while not end condition

if N+(S)\Tabu ¹ Æ and |S| < k choose the best S´;

else if N0(S)\Tabu ¹ Æ choose the best S´;

else choose the best S´ in N–(S) and update Tabu;

Update S¬S';

if Clique weight(S) > Clique_ weight(S*) then S*¬S;

end while;

return S*:

2.1.2 FP-Growth Algorithm: This algorithm mines the complete item sets without generating candidate set and uses and uses divide and conquer technique. It works on two steps-

- Build FP-Tree
- Mining of the FP-Tree to find the frequent item sets

Algorithm: FP growth-

Input: D, a transaction database;

Min sup, the minimum support count threshold.

Output: The complete set of frequent patterns.

Method:

- Construct the FP-tree using following steps-

(a) Scan the transaction database D once. Collect F, the set of frequent items, and their support counts. Then sort the F in descending order of support count.

(b) Create the root of an FP-tree, and label it as "null." For each transaction Trans in D do the following-if there is a transaction {I1, I5, I9} then make I1 child of root,I5 child of I1 and I9 child of I5.

- The FP-tree is mined by calling FP growth (FP tree, null), which is implemented in Fig.2

Fig.2. procedure for Mining FP-Tree

,

Procedure FP growth (Tree, a)

(1) If Tree contains a single path P then

(2) For each combination (denoted as b) of the nodes in the path P

(3) Generate pattern b [a with support count = minimum support count o f nodes in b;

(4) Else for each a_i in the header of Tree f

(5) Generate pattern $\beta = a_i$ [a with support count = a_i. Support count;

(6) Construct b's conditional pattern base and then b's conditional FP tree ;

(7) If Tree $_\beta \neq Ø$

(8) Call FP growth (Tree $_\beta$, β) ;}

2.2 Survey Table

After doing study of survey papers on Association Rules Mining a table is prepared to analyze all papers easily. This analysis is shown in Table 1.

Table 1: Comparison Study of Various Research papers

S.No	Title& Author Name	Published In	Focus On	Result Analysis
1	Research of Commonly Used Association Rules Mining Algorithm in Data Mining By, Ruowu Zhong, Huiping Wang	IEEE-International Conference on Internet Computing and Information Services-2011	Detailed study of Association algorithm and their analyses	It becomes easy to understand various algorithms.
2	Online Mining of data to Generate Association Rule Mining in Large Databases By, Archana Singh, Megha Chaudhary, Dr (Prof.) Ajay Rana Gaurav Dubey	IEEE-International Conference on Recent Trends in Information Systems-2011	A new and more optimized algorithm has been proposed for online rule generation. The advantage of this algorithm is that the graph generated in our algorithm has less edge as compared to the lattice used in the existing Algorithm.	Helps to remove redundant rules and in compact representation of association rules.
3	The Association Rule Mining on a Survey Data for Culture Industry By, Zhengui Li, Renshou Zhang	IEEE-2012 International Conference on Systems and Informatics (ICSAI 2012)	Apply Apriori Algorithm on culture industry.	The result shows that main affecting factors for a culture industry are participation, recognition, income.occupati-on, age and education.
4	An Improved Association Rule Mining Technique for Xml Data Using Xquery and Apriori Algorithm By,R.Porkodi,V.Bhuvanes wari,R.Rajesh,T.Amudha	IEEE International Advance Computing Conference (IACC 2009)	Improved framework for mining association rules from XML data using XQUERY and .NET based implementation of Apriori algorithm.	Apriori algorithm is used to extract association rules from Xml data.
5	A New Approach to Online Generation of Association Rules By, Charu C. Aggarwal	IEEE Transactions on knowledge and data Engineering, VOL. 13, NO. 4, JULY/AUGUST 2001	An online algorithm which is independent of the size of the transactional data and the size of the pre processed data.	The algorithm supports technique for quickly finding out association rules from large data item sets also present the association rules in compact form and reduce redundancy also.
6	An Effectual Algorithm For Frequent Item set Generation In Generalized Data Set Using Parallel Mesh Transposition By, Gurudatta Verma, Vinti Nanda	IEEE International Conference On Advances In Engineering, Science And Management (ICAESM -2012) March 30, 31, 2012	An algorithm that uses database in transpose form and the transpose of database is done by using Parallel transposition algorithm.	This algorithm is compared with Apriori algorithm for frequent items et generation and find faster than Apriori.
7	A Fast Algorithm for Mining Association Rules in Medical Image Data By, Adepele Olukunle and Sylvanus Ehikioya,	IEEE Canadian Conference on Electrical & Computer Engineering-2000	Fast association rule mining algorithm suitable for medical data sets.	FP-Growth algorithm is used to mine medical database.
8	Scalable Parallel Data Mining for Association Rules By, Eui-Hong (Sam) Han, George Karypis	IEEE Transactions on Knowledge and Data Engineering,12(3),Ma y-June 2000	Two new parallel formulations of the Apriori Algorithm(Intelligent Data Distribution and Hybrid Distribution algorithm) for computing Association Rules	Intelligent Data Distribution Algorithm creates the hash tree more effectively and scalable with respect to the candidate size. Hybrid Distribution algorithm achieves more load balancing than IDD because candidate set is partitioned

				into buckets.
09	Parallel Data Mining for Association Rules on Shared-Memory Systems By, S. Parthasarathy1, M. J. Zaki2, M. Ogihara3, W. Li4	Knowledge and Information Systems (2001) 3: 1-29 2001 Springer-Verlag London	New parallel algorithm for data mining of association rules on shared memory multiprocessor	This algorithm is used to balance the candidate generation and hash tree balancing and also provide good speed up for parallel- ization
10	Mining Approximate Frequency Item sets over Data Streams based on D-Hash Table By, Chunhua Ju, Gang You	10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing 2009	Hashed table is introduced to represent the synoptic data structure and an algorithm of frequent item set mining based on D-hashed table is proposed.	Proposed algorithm is more effective in time and space than Lossy counting Algorithm.
11	A New Perfect Hashing and Pruning Algorithm for Mining Association Rule By, Hassan Najadat, Amani Shatnawi and Ghadeer Obiedat	IBIMA Publishing Communications of the IBIMA Vol. 2011 (2011), Article ID 652178	A new hashing algorithm in discovering association rules among large data item sets.	The analysis shows that the new algorithm does not suffer from the collisions, which lead to high accuracy.
12	Data Structure for Association Rule Mining: T-Trees and P-Trees By, Frans Coenen, Paul Leng, and Shakil Ahmed	IEEE Transaction Knowledge and data Engineering, VOL. 16, NO. 6, JUNE 2004	Two new data structures T-tree and P-tree for association rules	T-tree provides better time and space requirement than hash tree and P-tree offers better processing in terms of storage and time requirement than FP-tree
13	An Efficient Decision Tree Classification Method Based on Extended Hash Table for Data Streams Mining By, Zhenzheng Ouyang, Quanyuan Wu, Tao Wang	IEEE Fifth International Conference on Fuzzy Systems and Knowledge Discovery-2008	Focuses on continuous attributes handling for mining data stream with concept drift and implemented a system Hash CVFDT on top of CVFDT.	Extended hash Table (EHT) is fast as hash table when it inserting, deleting or seeking examples. It calculates best split point efficiently in advantage of the list structure.
14	Research of Tax Inspection Cases-Choice Based On Association Rules In Data Mining By, Qing-Xiang Zhu,Li-Juan Guo,JingLiu,NanXu,Wei-Xu Li	IEEE-Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009	Set up inspection indexes for enterprise income tax, and then applies the Apriori algorithm of association rules of data mining into tax inspection cases-choice**.**	This method accurately finds the dishonest enterprise in order to improve the efficiency and effectiveness of inspection**.**
15	Algorithms for Association Rule Mining – A General Survey and Comparison By, Jochen Hipp, Ulrich G¨untzer, Gholamreza	ACM SIKKDD VOL-2,July 2000	Focuses on continuous attributes handling for mining data stream with concept drift and implemented a system Hash CVFDT on top of CVFDT.	The experiments were performed on SUNULTRA SPARC-II workstation clocked at 248 MHz and analyzed that all algorithms shows quite similar runtime behavior.

3. CASE STUDY

This paper performs a case study on an Academic Database to find out frequent elective sets chosen by M.Tech (C.S.E) students in their fourth and fifth semester and comparison is done between Similis Algorithm and FP-Growth Algorithm. Electives in fourth semester are-Advanced Computer Architecture (I1), Data Warehousing & Data Mining (I2), Digital Signal Processing (I3), Network Security (I4).First Elective set in fifth semester are-Compiler Construction(I5),Digital Image Processing(I6),Parallel Computing(I7). Second electives set in fifth semester are-Enterprise Resolution Planning (I8), Genetic Algorithms (I9), Total Quality Management (I10).The original Database for Electives of all the three semesters is shown in Table 2.

Table2: Database

Student_id	Elective Sets
01	I1,I5,I8
02	I2,I5,I8
03	I3,I6,I9
04	I4,I7,I9
05	I1,I5,I8
06	I2,I5,I8
07	I2,I6,I10
08	I1,I7,I8
09	I3,I5,I10
10	I2,I5,I8
11	I4,I5,I9
12	I3,I6,I9
13	I4,I6,I8
14	I2,I5,I8
15	I3,I5,I8
16	I1,I7,I9
17	I2,I5,I8
18	I3,I7,I10
19	I3,I6,I9
20	I2,I6,I9
21	I2,I5,I8
22	I4,I5,I9
23	I1,I5,10
24	I3,I7,I8
25	I2,I5,I8
26	I2,I7,I8
27	I3,I6,I9
28	I1,I5,I0
29	I3,I6,I9
30	I2,I5,I8

2.1 Similis Algorithm

First the database is solved using Similis algorithm. The minimum support count is 3.

Step 1-In the first step an Adjancey Matrix (Table-3) is created, which counts the frequency between nodes (electives).

Table-3: Adjancey Matrix

G(V,E)	I2	I3	I4	I5	I6	I7	I8	I9	I10
I1	-	-	-	4	-	2	3	1	2
I2	-	-	-	8	2	1	9	1	1
I3	-	-	-	2	5	2	2	5	2
I4	-	-	-	2	1	1	1	3	-
I5	-	-	-	-	-	-	11	2	3
I6	-	-	-	-	-	-	1	6	1
I7	-	-	-	-	-	-	3	2	1
I8	-	-	-	-	-	-	-	-	-
I9	-	-	-	-	-	-	-	-	-

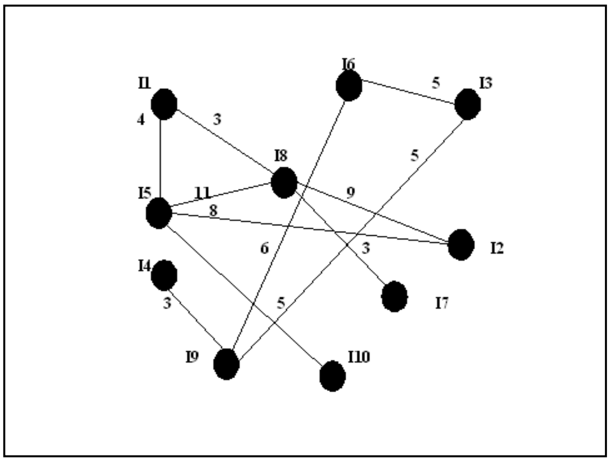
Step 2-The second step contains removal of entries from Adjancey Matrix which are smaller than minimum support Count (Table-4)

Table-4: Adjancey Matrix after Step-2

G(V,E)	I5	I6	I7	I8	I9	I10
I1	4			3		
I2	8			9		
I3		5			5	
I4					3	
I5				11		3
I6					6	
I7				3		

Step 3- In the third step a weighted graph G (Fig.3) in which all nodes represents electives.

Fig 3: Weighted Graph G



Step 4-In fourth step the maximum weight clique is found. The electives which are the part of maximum weight cycle are the most frequent elective sets.

{I2**, I5, I8}** is the maximum weighted clique. So **{Data warehousing and Data Mining, Compiler Construction, Enterprise Resource Management**} are most frequent elective subjects.

3.2 FP-Growth Algorithm

Step 1: Arrange the electives in descending order of frequencies.

Table 5: Electives in descending order

Electives	**Count**
I5	16
I8	15
I2	11
I9	10
I3	9
I6	8
I1	6
I7	6
I10	5
I4	3

Step2: Fig.4 shows the final FP-Tree after scanning the complete database.

Step3**:** After mining **I5:16, I8:11, I2:7** is the most frequent pattern so **{Data warehousing and Data Mining, Compiler Construction, Enterprise Resource Management}** are most frequent elective subjects.

4. RESULT ANALYSIS

After doing case study the same result for both algorithm (Similis algorithm and FP-Growth algorithm) is found. The result is that **{Data warehousing and Data Mining, Compiler Construction, Enterprise Resource Management}** are the subjects that students frequently preferred as electives. In other words if a student is choosing **Data warehousing and Data Mining** in 4th semester then the highest priority is that the student will choose **Compiler Construction** in 5th semester and **Enterprise Resource Management** in 6the semester.

A comparison of the performance of both algorithms is also done and a table (Table-6) is constructed that differentiates between performance of Similis algorithm and FP-Growth algorithm.

Fig.4 FP-Tree

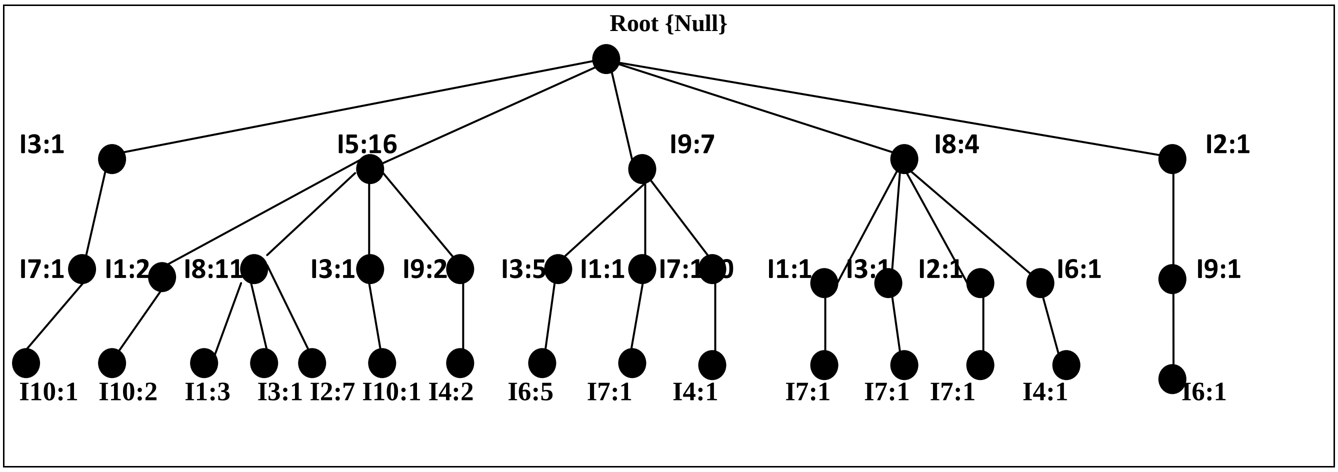


Table 6: Comparison between Performance of Similis Algorithm and FP-Growth Algorithm

S.No	Characteristics	Similis Algorithm	FP-Growth Algorithm
1	Measure Criteria	Clique weight	Support measure
2	Computational Time	Steady when item set increases	Takes more time if more items are there
3	Scalability	Good scalability	Less scalable than Similis algorithm
4	Time Complexity	$O(N^3)$	O(header-count^2*depth of tree)
5	Data Structure	Graphs	Tree

5. CONCLUSION AND FUTURE SCOPE

This paper finds out the most frequent electives for thirty students. In future the results can be derived from large datasets. Lot of data in huge databases enforces to bring out knowledge from the patterns. Association rule mining is one of the techniques in data mining which helps to know about the patterns and trends from the database. In this paper, FP-Tree and Primal-Tabu Meta-Heuristic for the Maximum-Weighted association rule mining algorithms have been studied and compared with the help of case study of post graduate students in knowing the patterns of Elective paper they choose. Further the same experiment and result analysis can be done on large datasets.

6. REFERENCES

[1] Adepele Olukunle and Sylvanus Ehikioya “A Fast Algorithm for Mining Association Rules in Medical Image Data” IEEE Canadian Conference on Electrical & Computer Engineering-2000

[2] Archana Singh, Megha Chaudhary, Dr (Prof.) Ajay Rana Gaurav Dubey ”Online Mining of data to Generate Association Rule Mining in Large Databases” IEEE-International Conference on Recent Trends in Information Systems-2011

[3] Cavique, Luís,” A Scalable Algorithm for the Market Basket Analysis”, Elsevier 2007

[4] Charu C. Aggarwal, “A New Approach to Online Generation of Association Rules” IEEE Transactions on knowledge and data Engineering, JULY/AUGUST 2001, VOL. 13, NO. 4

[5] Chunhua Ju, Gang You “Mining Approximate Frequency Item sets over Data Streams based on D-Hash Table”, 10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing 2009

[6] Eui-Hong (Sam) Han, George Karypis ,”Scalable Parallel Data Mining for Association Rules “ , IEEE Transactions on Knowledge and Data Engineering,12(3)

[7] Frans Coenen, Paul Leng, and Shakil Ahmed,”Data Structure for Association Rule Mining: T-Trees and P-Trees” IEEE Transaction Knowledge and data Engineering. 6, JUNE 2004, VOL. 16

[8] Gurudatta Verma, Vinti Nanda” An Effectual Algorithm For Frequent Item set Generation In Generalized Data Set Using Parallel Mesh Transposition “IEEE International Conference On Advances In Engineering, Science And Management (ICAESM -2012) March 30, 31, 2012

[9] Hassan Najadat, Amani Shatnawi and Ghadeer Obiedat “A New Perfect Hashing and Pruning Algorithm for Mining Association Rule” IBIMA Publishing Communications of the IBIMA Vol. 2011 (2011), Article ID 652178

[10] Jiawei Han, Jian Pei,”Mining Frequent Patterns by Pattern-Growth: Methodology and Implications”, ACM SIGKDD December 2000, Volume 2, Issue 2-Page 31

[11] Qin Ding, Qiang Ding, and William Perrizo,”PARM-An Efficient Algorithm to Mine Association Rules From Spatial Data”, IEEE transactions On Systems, Man and Cybernetics-partB: Cybernetics, Vol.38, NO. 6, DECEMBER 2008

[12] Qing-Xiang,Zhu,Li-Zuan Guo, Jing Liu,NanXu,Wei-Xu Li ,”Research of Tax Inspection Cases-Choice Based On Association Rules In Data Mining” IEEE- Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009

[13] Ratchadaporn Amornchewin,”Probability- based Incremental Association Rules Discovery Algorithm with Hashing technique” International Journal of Machine Learning and Computing, Vol.1, No. 1, April 2011

[14] Ruowu Zhong, Huiping Wang “Research of Commonly Used Association Rules Mining Algorithm in Data Mining “IEEE-International Conference on Internet Computing and Information Services-2011

[15] S. Parthasarathy1, M. J. Zaki2, M. Ogihara3, W. Li4,“Parallel Data Mining for Association Rules on Shared-Memory Systems”, Knowledge and Information Systems Springer-Verlag London-2001

[16] WANG Pei-ji,SHI Lin1,BAI Jin-niu, ZHAO Yu-lin,” Mining Association Rules Based on Apriori Algorithm And Application” 2009 International Forum on Computer Science-Technology and Applications

[17] Wei Zhang, Hongzhi Liao, Na Zhao,” Research on the FP Growth Algorithm about Association Rule Mining”, IEEE 2008 International Seminar on Business and Information Management,

[18] Zhengui Li, Renshou Zhang”The Association Rule Mining on a Survey Data for Culture Industry “IEEE-2012 International Conference on Systems and Informatics (ICSAI 2012)

[19] Zhenzheng Ouyang, Quanyuan Wu, Tao Wang

“An Efficient Decision Tree Classification Method Based on Extended Hash Table for Data Streams Mining”, IEEE Fifth International Conference on Fuzzy Systems and Knowledge Discovery-2008.