ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

Using vertical association knowledge with correlation for retrieval in case based reasoning

Ms.Aparna V.Mote, **Mr. Madhav D. Ingle**
Department of Computer Engineering Savitribai Phule University of Pune

Manuscript Info

Manuscript History:

Received: 14 May 2015
Final Accepted: 26 June 2015
Published Online: July 2015

Key words:

Association Rules, case based reasoning (CBR), association knowledge (AK), correlation

****Corresponding Author***

Ms.Aparna V.Mote

Abstract

Case-based reasoning (CBR), is the process of solving new problems based on the solutions of similar past problems. The most important step in CBR is reteieval. For retrieving relevant data the CBR systems mainly uses the similarity knowledge. Most of the retrieving systems use similarity knowledge and association rules for retrieving the required cases. But the existing algorithms strongly rely on similarity knowledge and ignore the other forms of knowledge that can be used to improve the retrieval performance. In this paper the well known algorithm that is Apriori algorithm is used to extract desired relevant cases based on the knowledge system of the association rules with the efficient correlation methods. The goal of this paper is to provide detailed review about retrieving useful cases by using different methods and showing the effectiveness of each algorithm.

INTRODUCTION

The Case-based reasoning (CBR) methods are basically used to solve new problem by using the previous available solutions. All the previous cases stored are called as experience and every experience is known as case. All the cases are stored at the location called as case-base. Usually, every case is expressed with the help of two factors i.e. the detail description of the problem and its solution. Basically, there are four phases in CBR as follows:

A) Retrieve:
Given a target problem, retrieve from memory cases relevant to solving it. A case consists of a problem, its solution, and, typically, annotations about how the solution was derived.

B) *Reuse:*
Map the solution from the previous case to the target problem. This may involve adapting the solution as needed to fit the new situation.

C) *Revise:*
Having mapped the previous solution to the target situation, test the new solution in the real world (or a simulation) and, if necessary, revise.

D) *Retain:*
After the solution has been successfully adapted to the target problem, store the resulting experience as a new case in memory.

But, retrieval is the most important phase in CBR because the performance of CBR is dependent on this phase [2].The main aim of this phase is to obtain similar cases or somewhat relevant cases to get the solution for the target problem. CBR retrieves similar past cases from the case base, reusing solutions from similar past cases to infer a proper solution to the current problem revising the proposed solution if necessary and retaining the new solution by incorporating it into the existing case base for future problem solving. The main goal of the CBR is to retrieve relevant and useful cases which can be used to solve the target problems. If CBR fails to retrieve useful cases, these systems will not be able to generate suitable cases to the given problem.

I. RELATED WORK

Normally, similarity knowledge (SK) is used in the retrieval process which is known as similarity-based retrieval (SBR)[2]. In this type of retrieval, SK is used to obtain the previous cases related to the target problem. With the help of measures and ranking, SBR obtains the cases related to the problem and with the help of these solutions the target problem is solved.

But, there are two disadvantages of SBR, first is for defining the SK practically, domain experts are required which makes this dependent on domain experts [3] and there is no any specific methodology available. Also, for defining SK, time required is more and it is very complicated process. Due to which the performance of SK is poor and results obtained are sometimes inaccurate. Second disadvantage is static definition of similarity measure. This means the definition is applicable consistently to all the target problems. This creates problem because the defined criterion is applicable to some target problems and not to all. So, the performance of the SBR varies based on the target problem even in the same domain [4].

In [11], a new hybrid data mining method TSFCR was introduced which dynamically applies the appropriate classifier between CBR and RI. But, the criterion to select the classifier is based on the appropriateness of the CBR rather than RI so it is unable to guarantee the appropriateness. In [4], ELEM2–CBR hybrid method was introduced which integrates RI and CBR but, this paper gives results for only specific data and not for all. Also the performance is dependent on the properties of data.

II. IMPLIMENTATION DETAILS

A) BLOCK DIAGRAM:

The Fig.1shows the proposed data flow architecture of retrieval process for case based reasoning by using vertical association knowledge with correlation.

The proposed system's different modules are communicating with one another on the following scenarios:

1. From User problem entering module to pre-processing module
2. From pre-processing module to TF-IDF module
3. From TF-IDF module to Info gain module
4. From Info gain Module to association rule mining module
5. Association rule mining module to correlation module
6. Correlation module to relevant case extraction module

Basically, the proposed system operates in four steps:

1. PREPROCESSING:

This is the step where all the XML data stored in DB are pre-processing by the following four main activities: Sentence Segmentation, Tokenization, Removing Stop Word, and Word Stemming.

2. INFO GAIN:

In order to summarize each of documents in an IR result, we use Shannon's term weighting based on formation Gain Ratio (IGR).This method extracts the similarity structure among a set of documents through a hierarchical clustering, then gives higher weights to words that contribute to forming the structure. Thus, by the using the vertical intersection of the words system identifies the most obvious words for rule mining using power set Where all these words are extracting by the comparative recursion of the combination of the words.

3. ASSOCIATION:

Then after fetching the important words from all the documents system will perform association rule using Apriori Algorithm.

4. PEARSON CORRELATION:

In the final step proposed system will perform vertical frequent pattern mining using éclat algorithm as shown below.

B) MATHEMTICAL MODEL AND ALGORITHM

Set Theory:

1. Let S={ } be as system for CBR
2. Identify Input as Q={ Q_1 , Q_2 ,……………….. Q_n}
   Where Q_n=User Problem
   S= {Q}
3. Identify R as Output i.e. RELEVANT CASES
   S= {Q, R}
4. Identify Process P
   S= {Q, R, P}
   P= {P_r, T, I_g , A_s ,P_c }
   Where P_r =Preprocessing
   T =Tf-IDF
   I_g=Info-Gain
   A_s =Association
   P_c =Pearson Correlation
5. S = { Q, R, P_r, T, I_g , A_s ,P_c }

Mathematical model for proposed system:

1. PREPROCESSING:
   Set P_r:
   P_{r0} =Get User Comments in String
   P_{r1}=split in Words
   P_{r2} =Remove Special Symbols
   P_{r3} =Identify Stopwords
   P_{r4} =Remove Stopwords
   P_{r5} =Identify Stemming Substring
   P_{r6} =Replace Substring to desire String
   P_{r7} =Concatenate Strings
2. TF-IDF:
   Set T:
   T_0 =calculate Term Weight of each term
   T_1 =Check for frequency in other document
   T_2 =Calculate inverse document frequency
3. INFO GAIN:
   Set I_g:
   I_{g0} =Count positive possibilities of a term
   I_{g1} =Count negative possibilities of a term
   I_{g2}= Calculate true ratio
   I_{g3} =Calculate logarithm of true ratio
   I_{g4} = Find info gain ratio
4. ASSOCIATION:
   Set A_s
   A_{s0} =Get important words
   A_{s1} = Apply power set
   A_{s2} =Check power set for combination of rules

A_{s3} =Check for threshold Confidance
A_{s4} =Check for Threshold support
A_{s5} =Collect rules

5. PEARSON CO_RELATION
   Set P_c:
   P_{c0} = Get rules
   P_{c1} =get user query problem
   P_{c2} = Co-Relation Coefficients
   P_{c3} = Covariance Calculations
   P_{c4} =Variance Calculation
   P_{c5} =Pearson Score

Steps of association rule using Apriori Algorithm:

Let *T* be the training data with *n* attributes *A1*, *A2*, …, *An* and *C* is a list of class labels. A particular value for attribute *Ai* will be denoted *ai*, and the class labels of *C* are denoted *cj*.

- An item is defined by the association of an attribute and its value (*A*i, *ai*), or a combination of between 1 and n different attributes values, e.g. < (A1, a1)>, < (A1, a1), (A2, a2)>, (A1, a1), (A2, a2), (A3, a3)>, … etc.
- A rule *r* for multi-label classification is represented in the form: $(A_{i1}\ ,\ a_{i1}\ )\ \wedge\ (A_{i2}\ ,\ a_{i2}\ )\wedge\ldots\wedge(A_{1m}\ ,\ a_{im}\ )\rightarrow c_{i1}\ldots.c_{im}$ where the condition of the rule is an item and the consequent is a list of ranked class labels.
- The actual occurrence (*ActOccr*) of a rule *r* in *T* is the number of cases in *T* that match *r's* condition.
- The support count (*SuppCount*) of *r* is the number of cases in *T* that matches *r's* condition, and belong to a class *ci*. When the item is associated with multiple labels, there should be a different *SuppCount* for each label.
- A rule *r* passes the minimum support threshold (*MinSupp*) if for *r*, the $SuppCount(r)/\ |T| \geq MinSupp$, where $|T|$ is the number of instances in *T*.
- A rule *r* passes the minimum confidence threshold (*MinConf*) if $SuppCount(r)/ActOccr(r) \geq\ MinConf$.
- Any item in *T* that passes the *MinSupp* is said to be a frequent item.

Eclat Algorithm:

Input: Alphabet A with ordering ≤ multiset T ⊆ P(A) of sets of Items , Minimum support value minsup Є ℕ.
Output: Set F of frequent Itemsets and their support counts.

```
1. F:={(Ø,|T|) }.
2. CØ:= {(x,T({x}))| x Є A}.
3. C'Ø:= freq (C $_Ø$):= {(x,$T_x$)|(x,$T_x$) Є C $_Ø$, |$T_x$|≥ minsup }
4. F:= { Ø }.
5. Add frequent supersets (Ø, C'$_Ø$).
```

Function add frequent Supersets():
Input: frequent Itemsets p Є P(A) called prefix, incidence matrix C of frequent 1-item-extentions of p.
Output: add all frequent extensions of p to global variable F.

```
for (x, $T_x$) Є C do
q:= p U {X}.
$C_q$:={(y,$T_x \cap T_y$) | (y,$T_y$) Є C, y > x}.
C'q := freq($C_q$) := {(y,$T_y$) | (y, $T_y$) Є $C_q$, |$T_y$| ≥ minsup }
If C'$_q$ ≠ Ø then
Add frequent supersets (q,C'$_q$).
End if
```

```
8.  F := F U {(q, |T_x|)}
9.  End for
```

III. RESULTS AND DISCUSSIONS

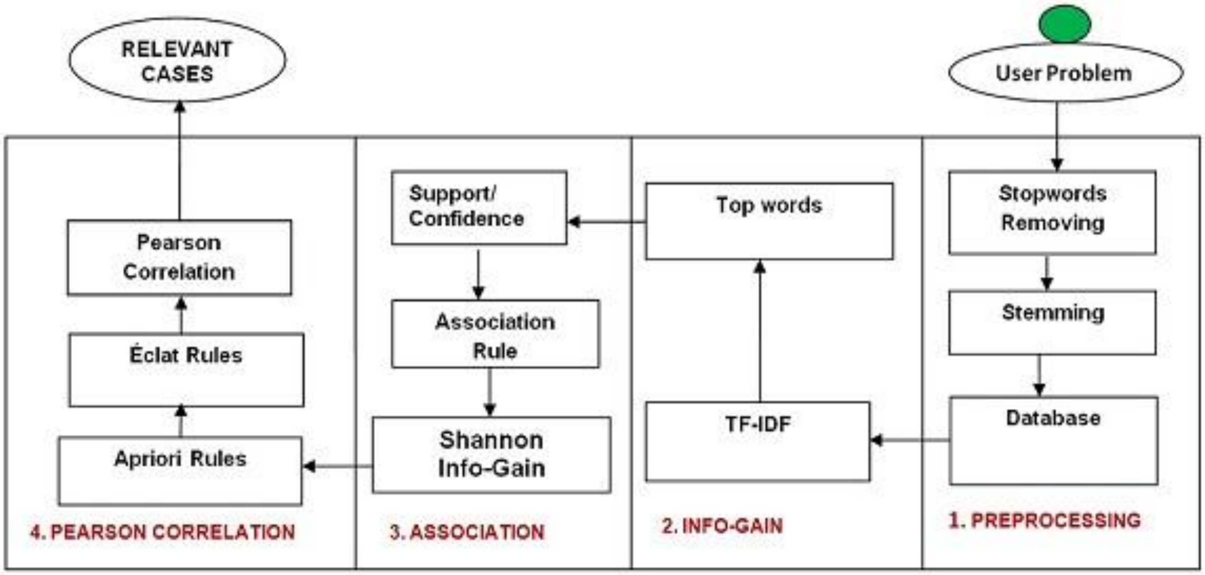


Fig1: Proposed system's architecture

The evaluation performance of CBR using vertical association knowledge with correlation approach, a series of experiments on Excel data and all experiments were performed on Windows machine having configuration dual core processor of 2.2 GHz, 100 GB hard disk and 2GB RAM.

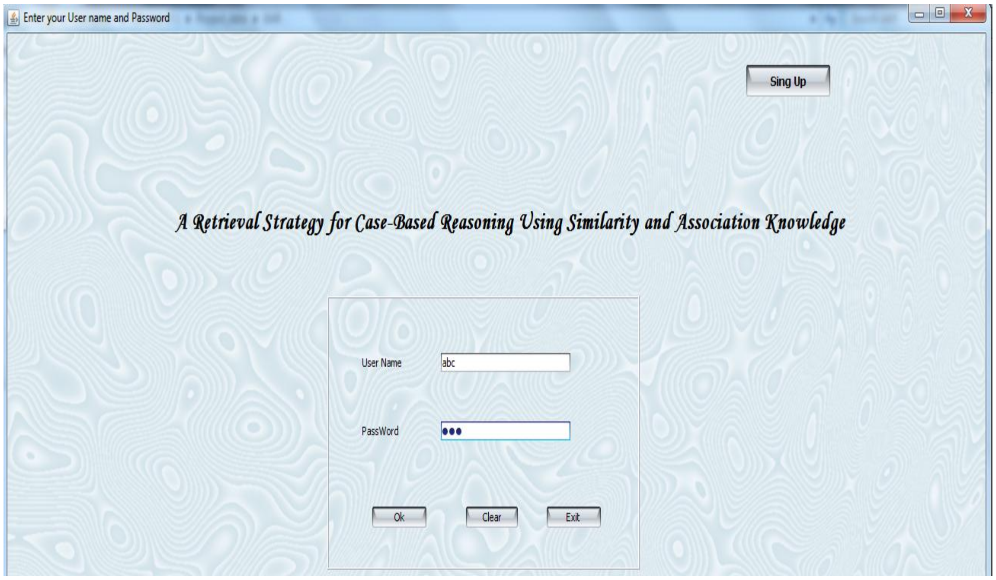
To show the effectiveness of the proposed system, some experiments are reported. Selecting a suitable dataset is a critical and important step in designing rule mining system.

There is no condition in data mining for the usage of the specific dataset for the research. Any huge data set can be serving for this purpose. So to perform experiment on our system we use most generalized data set from the Reuters which are in the xml structure. As this data set is huge and having great versatility it provide a good challenge to our task.

A) Practicability of System Demonstration

In our proposed system the user selects the XML dataset and extracts the needed data using XQuery to store in database. After that user need to enter minimum support and confident on the basis of which he wants to extract the rules from Eclat algorithm. Then System performs the series of feature extraction methods like tf-idf and Shannon information gain system. Then by applying a powerset for the intersection of the transaction data system generates the frequent item sets. Then generated frequent item sets will be tested for the minimum support and confidence to get the efficient rule.

B) Screenshots



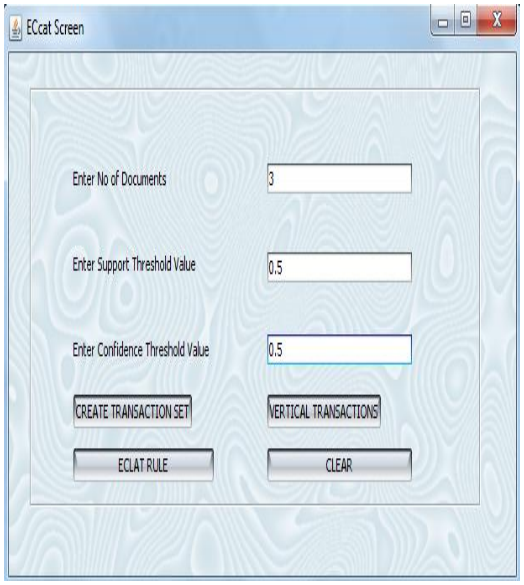
Enter your User name and Password

A Retrieval Strategy for Case-Based Reasoning Using Similarity and Association Knowledge

User Name abc

PassWord •••

Fig 2: Login Form

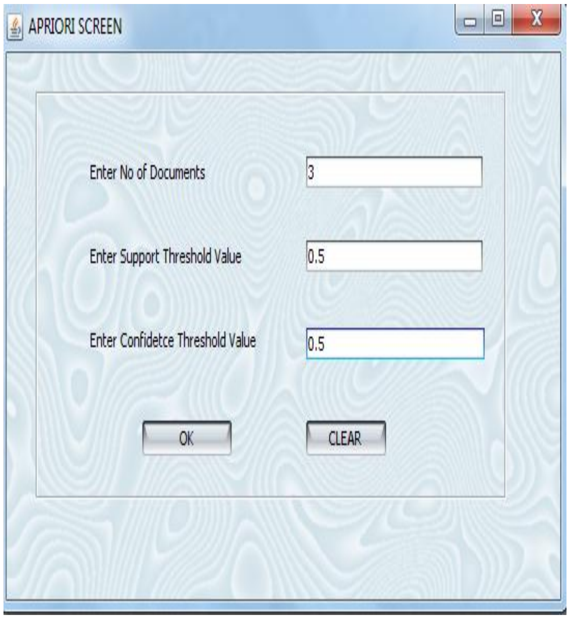


ECcat Screen

Enter No of Documents 3

Enter Support Threshold Value 0.5

Enter Confidence Threshold Value 0.5



APRIORI SCREEN

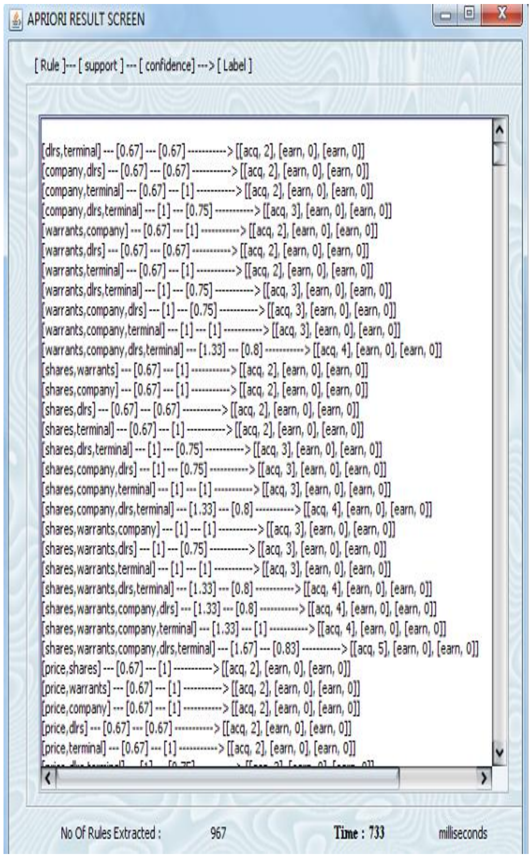
Enter No of Documents 3

Enter Support Threshold Value 0.5

Enter Confidetce Threshold Value 0.5

Fig 3: Eclat Screen

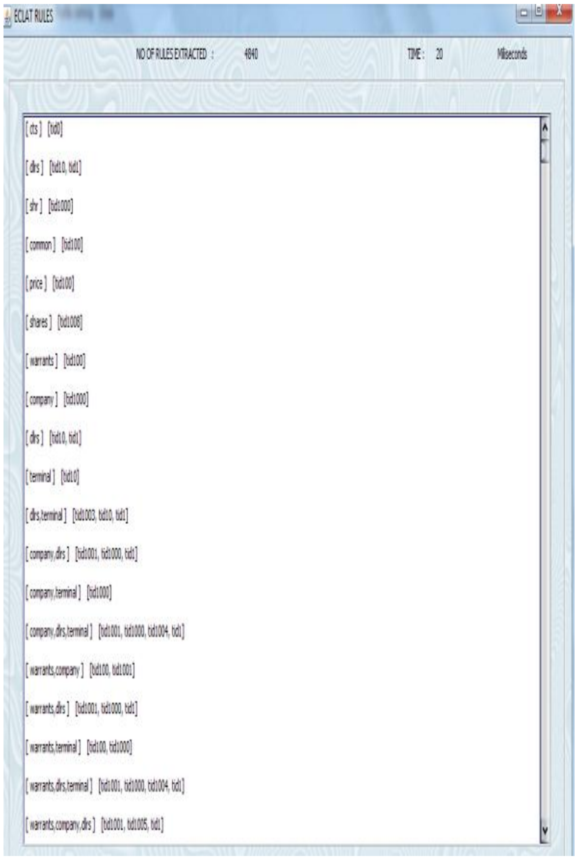
Fig 4: Apriori Screen



```
[ Rule ]--- [ support ] --- [ confidence] ---> [ Label ]

[dlrs,terminal] --- [0.67] --- [0.67] -----------> [[acq, 2], [earn, 0], [earn, 0]]
[company,dlrs] --- [0.67] --- [0.67] -----------> [[acq, 2], [earn, 0], [earn, 0]]
[company,terminal] --- [0.67] --- [1] -----------> [[acq, 2], [earn, 0], [earn, 0]]
[company,dlrs,terminal] --- [1] --- [0.75] -----------> [[acq, 3], [earn, 0], [earn, 0]]
[warrants,company] --- [0.67] --- [1] -----------> [[acq, 2], [earn, 0], [earn, 0]]
[warrants,dlrs] --- [0.67] --- [0.67] -----------> [[acq, 2], [earn, 0], [earn, 0]]
[warrants,terminal] --- [0.67] --- [1] -----------> [[acq, 2], [earn, 0], [earn, 0]]
[warrants,dlrs,terminal] --- [1] --- [0.75] -----------> [[acq, 3], [earn, 0], [earn, 0]]
[warrants,company,dlrs] --- [1] --- [0.75] -----------> [[acq, 3], [earn, 0], [earn, 0]]
[warrants,company,terminal] --- [1] --- [1] -----------> [[acq, 3], [earn, 0], [earn, 0]]
[warrants,company,dlrs,terminal] --- [1.33] --- [0.8] -----------> [[acq, 4], [earn, 0], [earn, 0]]
[shares,warrants] --- [0.67] --- [1] -----------> [[acq, 2], [earn, 0], [earn, 0]]
[shares,company] --- [0.67] --- [1] -----------> [[acq, 2], [earn, 0], [earn, 0]]
[shares,dlrs] --- [0.67] --- [0.67] -----------> [[acq, 2], [earn, 0], [earn, 0]]
[shares,terminal] --- [0.67] --- [1] -----------> [[acq, 2], [earn, 0], [earn, 0]]
[shares,dlrs,terminal] --- [1] --- [0.75] -----------> [[acq, 3], [earn, 0], [earn, 0]]
[shares,company,dlrs] --- [1] --- [0.75] -----------> [[acq, 3], [earn, 0], [earn, 0]]
[shares,company,terminal] --- [1] --- [1] -----------> [[acq, 3], [earn, 0], [earn, 0]]
[shares,company,dlrs,terminal] --- [1.33] --- [0.8] -----------> [[acq, 4], [earn, 0], [earn, 0]]
[shares,warrants,company] --- [1] --- [1] -----------> [[acq, 3], [earn, 0], [earn, 0]]
[shares,warrants,dlrs] --- [1] --- [0.75] -----------> [[acq, 3], [earn, 0], [earn, 0]]
[shares,warrants,terminal] --- [1] --- [1] -----------> [[acq, 3], [earn, 0], [earn, 0]]
[shares,warrants,dlrs,terminal] --- [1.33] --- [0.8] -----------> [[acq, 4], [earn, 0], [earn, 0]]
[shares,warrants,company,dlrs] --- [1.33] --- [0.8] -----------> [[acq, 4], [earn, 0], [earn, 0]]
[shares,warrants,company,terminal] --- [1.33] --- [1] -----------> [[acq, 4], [earn, 0], [earn, 0]]
[shares,warrants,company,dlrs,terminal] --- [1.67] --- [0.83] -----------> [[acq, 5], [earn, 0], [earn, 0]]
[price,shares] --- [0.67] --- [1] -----------> [[acq, 2], [earn, 0], [earn, 0]]
[price,warrants] --- [0.67] --- [1] -----------> [[acq, 2], [earn, 0], [earn, 0]]
[price,company] --- [0.67] --- [1] -----------> [[acq, 2], [earn, 0], [earn, 0]]
[price,dlrs] --- [0.67] --- [0.67] -----------> [[acq, 2], [earn, 0], [earn, 0]]
[price,terminal] --- [0.67] --- [1] -----------> [[acq, 2], [earn, 0], [earn, 0]]

No Of Rules Extracted :      967          Time : 733          milliseconds
```



```
NO OF RULES EXTRACTED :    4840          TIME :  20          Miliseconds

[ cts ]  [tid0]
[ dlrs ]  [tid10, tid1]
[ shr ]  [tid1000]
[ common ]  [tid100]
[ price ]  [tid100]
[ shares ]  [tid1008]
[ warrants ]  [tid100]
[ company ]  [tid1000]
[ dlrs ]  [tid10, tid1]
[ terminal ]  [tid10]
[ dlrs,terminal ]  [tid1003, tid10, tid1]
[ company,dlrs ]  [tid1001, tid1000, tid1]
[ company,terminal ]  [tid1000]
[ company,dlrs,terminal ]  [tid1001, tid1000, tid1004, tid1]
[ warrants,company ]  [tid100, tid1001]
[ warrants,dlrs ]  [tid1001, tid1000, tid1]
[ warrants,terminal ]  [tid100, tid1000]
[ warrants,dlrs,terminal ]  [tid1001, tid1000, tid1004, tid1]
[ warrants,company,dlrs ]  [tid1001, tid1005, tid1]
```

Fig 5: Apriori result Screen

Fig 6: Eclat result Screen

c) Relevant Comparisons

Author [7] proposes a method of extracting the rules using Apriori over the XML data using XQuery. For maintaining balance and similarity for the comparison proposed system also uses a dataset which contains about 20 files and average of 6 transactions in each files. And each file is containing more than 12 items.

Then system was tested for various support values to check its feasibility with the Apriori algorithm whose results can be shown in below figure 7.

It is clearly observed from the figure 7 that as the support increases the processed time of both the algorithms leaps for same value. The proposed system of Eclat has achieved better precision as compared to system proposed by the author [7] which uses Apriori as the mining algorithm. This shows frequent items fetched by intersection of transaction perform well in time and also gives good quality of rules.

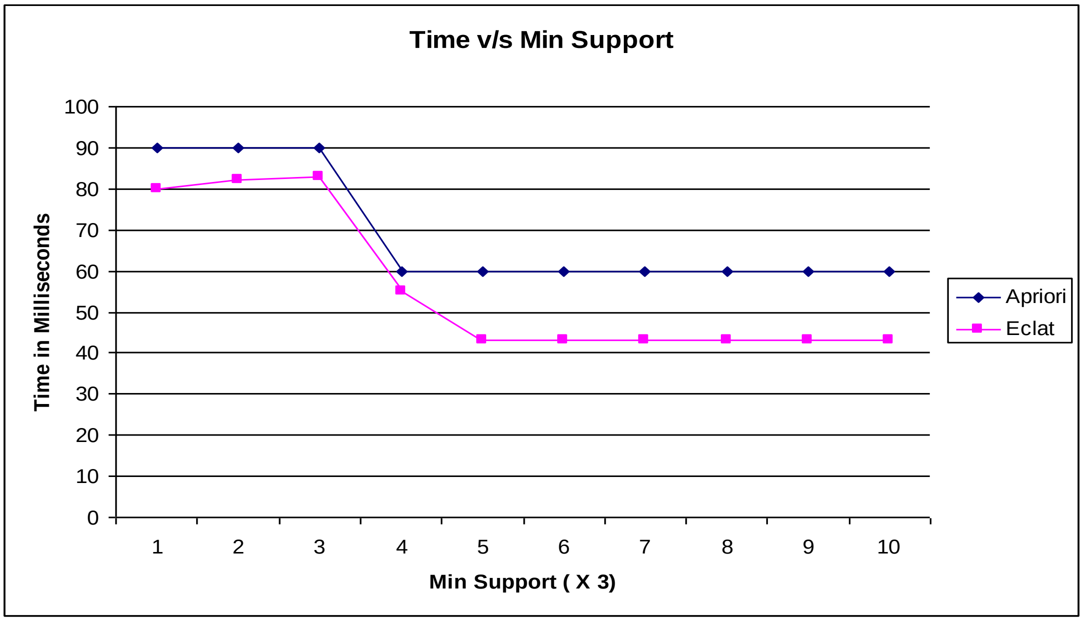


Fig 7: Time comparison of Apriori and Eclat Algorithms

IV. CONCLUSION

In the proposed approach of mining association rules system efficiently enhance the feature of Eclat algorithm with comparative power set. Comparative power set extract the maximum frequent itemsets from important words which are been decided by tf-idf and Shannon information gain. Proposed system enforces the powerset with multi recursion methodology to get as maximum as possible of intersection transactions. This method actually enhances the Eclat algorithm to create frequent itemsets on intersection and thereby to reduce the space and time complexity efficiently.

System efficiently takes comparatively less processing time to get the rules for the given minimum support than the other mining algorithms like Apriori. Which are creating more frequent items on each run even on small datasets; this actually doubts the selection of Apriori algorithm for huge datasets. The comparison of both algorithms were discussed in the last section, where éclat is over coming Apriori clearly in all possible given minimum support, This justifies Eclat over Apriori for huge datasets.

As the feature work of this proposed method, frequent itemsets can be extracting on the basis of group of distinct terms with recursive multithreading methodology to enhance the time complexity to perform the rule mining in exponentially less time.

ACKNOWLEDGMENT

I would like to give my sincere gratitude to my guide Prof.M.D.Ingle who encouraged and guided me throughout this paper. I am especially grateful to H.O.D Prof. Hingoliwala H.A. for their valuable guidance and encouragement.

REFERENCES

[1] Yong-Bin Kang, Shonali Krishnaswamy, and Arkady Zaslavsky, "A Reteieval Strategy For Case-Based Reasoning Using Similarity And Association Knowledge ,"IEEE transactions on cybernetics,Vol 44,No.4,April 2014.

[2] R.Lopez De Mantras, D.McSherry,D.Bridge,D.Leake,B.Smyth, B.Faltings, M.L.Maher,M.T.Cox,K.Forbus,M.Keane,A.Amodt, and I.watson , "Retrieval,reuse,revisionand retention in case-based reasoning," Knowl.Eng.Rev.,vol.20,n0.3,pp.215-240,2005

[3] Y. Guo, J. Hu, and Y. Peng, "Research on CBR system based on data mining," *Appl. Soft Comput.*, vol. 11, no. 8, pp. 5006–5014, 2011.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[4] Y.-J. Park, E. Choi, and S.-H. Park, "Two-step filtering datamining method integrating case-based reasoning and rule induction," *ExpertSyst. Appl.*, vol. 36, no. 1, pp. 861–871, 2009.

[5] B. Smyth and P. McClave, "Similarity vs. diversity," in *Case-Based* Reasoning Research and Development. Berlin, Germany: Springer-Verlag,2001,pp.347–361.

[6] J. L. Castro, M. Navarro, J. M. S´anchez, and J. M. Zurita, "Loss and gain functions for CBR retrieval," *Inf. Sci.*, vol. 179, no. 11, pp. 1738–1750,2009.

[7] R.Porkordi,V Bhuvaneshwari, R. Rajesh and T. Amudha "An Improved Association Rule Mining Technique for Xml Data Using Xquery and Apriori Algorithm ", IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009

[8] Y.-B. Kang, A. Zaslavsky, S. Krishnaswamy, and C. Bartolini, "A knowledge-rich similarity measure for improving IT incident resolution process," in *Proc. ACM Symp. Appl. Comput.*, 2010, pp. 1781–1788.

[9] A. Stahl, "Learning of knowledge-intensive similarity measures in casebased reasoning," Ph.D. dissertation, Artificial Intelligence Knowledge- Based Systems Research Group, Tech. Univ. Kaiserslautern, Kaiserslautern,Germany, 2003.

[10] P. Gautam and K. R. Pardasani, "Algorithm for efficient multilevel association rule mining," *Int. J. Comput. Sci. Eng.*, vol. 2, no. 5, pp. 1700–1704, 2010.

[11] Y. Guo, J. Hu, and Y. Peng, "Research on CBR system based on data mining," *Appl. Soft Comput.*, vol. 11, no. 8, pp. 5006–5014, 2011.