2018 International Conference on Information, Communication, Engineering and Technology (ICICET)
Zeal College of Engineering and Research, Narhe, Pune, India. Aug 29-31, 2018

Information Retrieval in Case Based Reasoning Using Vertical Association Knowledge and Shannon Information Gain

Aparna Vinayak Mote
Computer Engineering Department
Zeal college of Engineering & Research
Pune, India
aparna.mote23@gmail.com

Pratima Patil
Computer Engineering Department
Trinity Academy of Engineering,
Pune, India
pratima_dk@yahoo.com

Tejaswini Mane
Computer Engineering Department
Zeal college of Engineering & Research
Pune, India
tejaswinimane18@gmail.com

***Abstract*—In case based reasoning the new problems occurred are solved with the help of the solutions used for similar problems occurred in the past. Information retrieval is very important step in case based reasoning. Similarity knowledge is used to retrieve the data in CBR. Many of the existing systems use similarity knowledge as well as the association rules to retrieve the information. But many existing algorithms are mainly dependent on the similarity knowledge and they don't consider the other available forms of which are helpful for the information retrieval process. This paper uses Apriori algorithm for extracting the expected relevant cases which are dependent on association rules and also on the correlation methods. The main goal of this paper is to provide details of information retrieval in CBR with the help of different methods and also to show the efficiency of the Eclat algorithms.**

Keywords—Case Based Reasoning (CBR), Association Knowledge (AK), Association Rule Mining (ARM)

I. INTRODUCTION

In case based reasoning the new problems occurred are solved with the help of the solutions used for similar problems occurred in the past.The previous cases are known as experience where as every experience is considered as a case. Generally, case is represented with the help of two factors namely the detail description of the problem and its solution. CBR mainly has four phases namely Retrieve, Reuse, Revise, Retain. They are described as follows:

A. Retrieve: In this phase the similar relevant cases are retrieved to solve the given problem. A case contains the problem, solution of the problem also information about how the solution was derived.

B. Reuse: This phase is used for mapping the solutions of the previous problems to the target problem.

C. Revise: the real world solution or simulation is tested in this phase and revised if required.

D. Retain: once the final solution is obtained, the results are referred as a new case and they are stored for future use.

Following diagram shows the different phases of CBR:

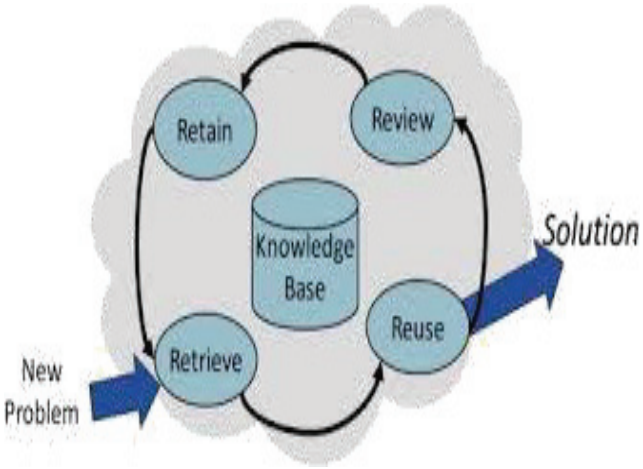

Fig. 1. Phases of CBR

But, as the performance of the CBR depends upon the retrieval phase [2] which is very important phase in CBR. The main aim of this phase is to get the exact similar or close to relevant cases for obtaining the solution for the given problem. The CBR uses these past similar cases which are stored in the case base, the solutions of past similar cases are reused to obtain proper solution for the current problem. The previously available solution can be revised if required and the new solution can also be retained by incorporating it into the available case base for future use.

The main aim of CBR is to retrieve the most relevant and useful cases those can be used to obtain the solution for the target problem and if CBR fails to obtain useful cases fail to generate suitable results for the given problem.

II. 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. By using 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 not applicable to all the problems so the performance of these systems varies depending upon the target problem even if problem belongs to the same domain [4].

In [11], a new hybrid data mining method TSFCR was introduced which dynamically applies specific classifier between CBR and RI. But, the criterion to select the classifier is dependent on the correctness of the CBR and not on the RI so it is unable to guarantee about the correctness. In [4], ELEM2–CBR hybrid method was introduced which integrates RI and CBR but, system gives results only for some specific problems and not to all the problems. Also the performance varies as it is dependent on the properties of data.

III. PROPOSED SYSTEM

A. Block Diagram:

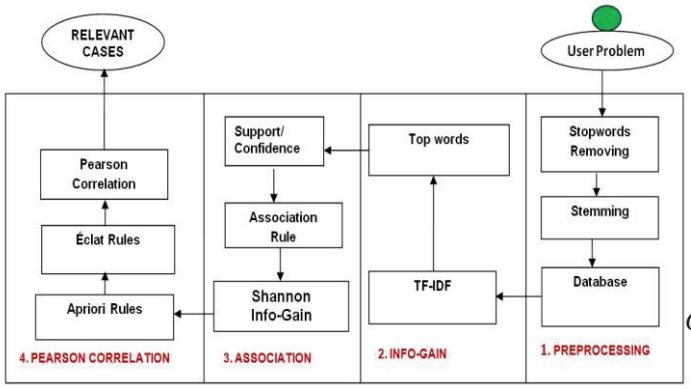

Fig. 2. Proposed system's architecture

The Fig.2 shows 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 operates in four steps as described below:

1) Preprocessing:

In this is step the preprocessing is done on the XML data which is stored in database. The preprocessing is done by segmentation of sentences, Tokenization, stop word removing and stemming.

2) Info Gain:

To summarize the documents in results of IR, Shannon's term weighing is used which is based on formation.

Information Gain Ratio (IGR) is a method which with the help of hierarchical clustering extracts the structures which are similar and then based on the contribution of words for forming the structure weights are assigned.

Thus, with the help of vertical intersection of words, the obvious words for rule mining are identified using power set.

3) Association:

In this phase, system performs association rules using Apriori algorithm using important words which are fetched from all the documents.

4) Pearson Correlation:

In this final step, with the help of Eclat algorithm, vertical frequent pattern mining is done.

B. Mathematical Model: Set Theory:

1. Let S={ } be as system for CBR
2. Identify Input as Q={ Q1 , Q2 ,…… Qn}
   Where Qn= User Problem and S= {Q}
3. Identify R as Output i.e. RELEVANT CASES={Q,R}
4. Identify Process P
   S= {Q, R, P}
   P= {Pr, T, Ig , As ,Pc }
   Where Pr =Preprocessing
   T =Tf-IDF
   Ig=Info-Gain
   As =Association
   Pc =Pearson Correlation
5. S = { Q, R, Pr, T, Ig , As ,Pc }

C. Mathematical model for proposed system:

1) Preprocessing:

Set P_r:
Pr0 =Get User Comments in String
Pr1=Split in Words
Pr2=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:
T0 =calculate Term Weight of each term
T1 =Check for frequency in other document
T2 =Calculate inverse document frequency

3) Info Gain:

Set I_g:
Ig0 =Count positive possibilities of a term
Ig1 =Count negative possibilities of a term
Ig2 = Calculate true ratio
Ig3 =Calculate logarithm of true ratio
Ig4 = Find info gain ratio

4) Association:

Set As:
As0 = Get important words

As1 = Apply power set
As2 =Check power set for combination of rules
As3 =Check for threshold Confidence
As4 =Check for Threshold support
As5 =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

IV. RESULTS AND DISCUSSIONS

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.

The important and critical step in designing rule mining system is selecting suitable dataset where as in data mining there are no restrictions for usage of datasets. Any dataset can be used for research. So, to check the performance of our system, we have used most generalized dataset from Reuters. The datasets used are in XML structure.

A. Practicability of System Demonstration:

In our proposed system XML dataset is selected by user and then the required data is extracted to store it in database using XQuery. Then user has to enter the minimum support and confidence. After that feature extraction is done by the system using methods like tf-idf and Shannon information gain. Then the system generates frequent item sets by applying power sets. Then these obtained frequent item sets are tested for minimum support and confidence to obtain the efficient rule.

B. Relevant Comparisons:

Author [8] proposed a method for extracting rules using Apriori method. For maintaining the similarity for comparison, we have used data set which has around 20 files and each file contains average 6 transactions. Also every file has more than 12 items.

The testing of the system is done for various support values to check the feasibility with Apriori algorithm. The results are as shown below in fig3.

From fig3, it is observed that as the support value increases, both the algorithms leaps for same value of processed time. The proposed system which uses Eclat algorithm achieves better precision than the system proposed by author [8] which uses Apriori algorithm. This shows that frequent items fetched by intersection of transaction does well in time and also good quality of rules are obtained.

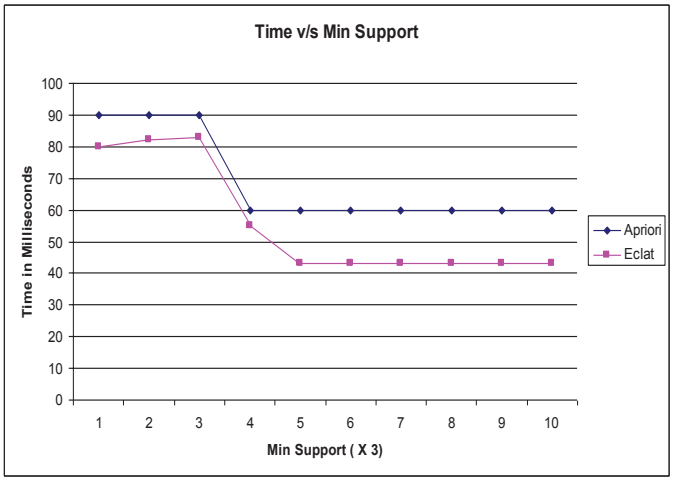

Fig. 3. Time comparison of Apriori and Eclat Algorithms

V. CONCLUSION

The proposed system of mining association rules efficiently enhances the Elcat algorithm with the help of comparative power sets. The comparative power sets are used for extracting maximum frequent item sets decided by tf-idf and Shannon information gain. To get maximum possible intersection of transactions, the proposed system uses multi recursion methodology. This method enhances the Eclat algorithm also reduces the time and space complexity effectively. This system comparatively takes less processing time than other mining algorithms like Apriori. So, it is clear that Eclat is overcoming Apriori in all possible given minimum support. So thus justifies Eclat over Apriori over huge datasets.

In future work, the extraction of frequent item sets can be done on the basis distinct groups using recursive multithreading methodology which will enhance the time complexity to perform the rule mining in exponentially less time.

REFERENCES

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

[2] M.L.Maher,M.T.Cox,K.Forbus,M.Keane,A.Amodt,"Retrieval,reuse,revision and retentionin CBR,"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] 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.

[8] 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

[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.