Combining the Attribute Oriented Induction and Graph Visualization to Enhancement Association Rules Interpretation

Safaa O. Al-Mamory
College of Business Informatics
University of Information Technology and Communications
Iraq
salmamory@uoitc.edu.iq

Zahraa Najim Abdullah
Department of computer science
University of Karbala
Iraq
zahraa_alanbary@itnet.uobabylon.edu.iq

***Abstract*— The important methods of data mining is large and from these methods is mining of association rule. The mining of association rule gives huge number of the rules. These huge rules make analyst consuming more time when searching through the large rules for finding the interesting rules. One of the solutions for this problem is combing between one of the Association rules visualization method and generalization method. Association rules visualization method is graph-based method. Generalization method is Attribute Oriented Induction algorithm (*AOI*). AOI after combing calls Modified AOI because it removes and changes in the steps of the traditional AOI. The graph technique after combing also calls grouped graph method because it displays the aggregated that results rules from AOI. The results of this paper are ratio of compression that gives clarity of visualization. These results provide the ability for test and drill down in the rules or understand and roll up.**

***Index Terms* —Data mining, Association rules, Visualization, AOI.**

I. INTRODUCTION

Data mining has a number of common methods; one of such methods is the association rules mining. Apriori is an example of the association rules algorithms .The mining of association rule gives huge number of the rules that resulted from apriori. These huge rules make analyst consuming more time when searching through the large rules for finding the interesting rules, interpreting and evaluation these rules.

Therefore, the problem of dealing with these rules is the basis of the idea of this paper. One of the solutions for this problem is the visualization that makes Audience in interactive with the rules .The visualization of association rules makes the analyst Focus on the main components of the association rules like items in the rules, the relation between the items and the interesting measures that Ingredients of evaluation of the association rules. Many researchers introduced many visualization techniques. This paper dealt one of these techniques is graph-based visualization. This technique Characterized by view way that make easy interpretation the rules by the user. This technique combined with AOI to view large rules.

In this paper, number of step of AOI remove and the others modified. After the combing process, AOI reduces the huge number of the rules to produce the aggregated rules, the graph visualization takes the results of AOI to visualize, AOI is called Modified AOI and the graph technique is also called grouped graph method.

The results of this thesis are ratio of compression that gives clarity of visualization. These results provide the ability for test and drill down in the rules or understand and roll up.

This paper contains into six sections. So far, there is an introduction. In section two, a survey of the literature related to the subject is given. In section three, we introduce a preliminary of the method. In section four, we present how rule sets could be grouped by the new modified AOI and then visualize the rules in new grouped graph visualization technique. In section five, results are discussed, while the conclusions are given in section six.

II. RELATED WORKS

This paper revolves around two classes of topics: the first class is the visualization techniques. Scatter plot visualization technique uses support and confidence measures for axes and lift measure for point shading [1] , while two-key plot uses the order measure for point shading [2]. Double decker is used for displaying one rule [3]. Parallel coordinate uses the items and its position in the rules for axes and arrow for the rules [4]. Then matrix-based visualization technique uses antecedent and consequent for axes and interest measures for colored rectangle [5], while (matrix3D) uses the 3D bar instead of colored rectangle.

Hahsler et al. [6] proposed grouped matrix-based visualization technique to enhance matrix-based by grouping the antecedent of the rules. Other techniques like Graph-based visualization technique uses vertex for items or item sets and edges for relationships [7];[8];[9].

The second class is the clustering of association rules techniques. In this context, Gupta et al. [10] proposed a new measure that takes the distance between Association rules based on a conditional probability estimate, as in Eq. (1)

$$d_{i,j} = P\left(\overline{BS_i} \vee \overline{BS_j} | BS_i \vee BS_j\right) \quad (1)$$

Where the set BS is the union of items in the left and right hand sides of rule *i*, we call $d_{i,j}$ the Conditional Market-Basket Probability (CMPB) distance and CMPB measure is used by Agglomerative Chain Clustering algorithm to find the clusters.

M. Klemettinen et al. [11] pruned the rules by extracting a subset that is called a rule covers from the original set of rules, a method for reducing the number of rules by eliminating of redundancy is applied.

Lent et al. [12] introduced a clustered association rule as a rule that is formed by combining similar "adjacent" association rules to form few general rules instead of a set of (attribute = value) equalities. For clustered rules, they had a set of value ranges using inequalities and he considered clustered association rules as in Eq. (2) the association rule is clustered in a two-dimensional space, where each axis represents one attribute on antecedent or Left Hand Side (LHS) and the consequent or Right Hand Side (RHS) that satisfies our segmentation criteria.

$$A_1 \wedge A_2 \wedge \cdots \wedge A_n \Rightarrow G \quad (2)$$

III. PRELIMINARY

Klemettinen et al., Ertek et al. and Rainsford et al [7];[8];[9] introduced graph-based method that represent the items or the itemsets by vertices and the relationship in rules by edges for visualization of association rule. In Fig. 1a, the vertex uses for the itemsets and directed edges between the itemsets for the rules. In Fig. 1b, the vertex uses for the items and rules share those items. This method selects as basis for the proposed visualization method in the next section.

IV. THE PROPOSED SYSTEM

Figure. 2 illustrate the proposed system flow chart, where we enhance the association rules interpretability by the following steps:

1. The system takes the large number of rules from the Apriori algorithm with lift, confidence and support interest measures.
2. The proposed modified AOI algorithm performs in Fig. 3. To produce aggregated rules that produce less number of rules than the original rules.
3. The aggregated rules views in one of the subjective approaches like the visualization to determine the interesting rules. In particular, graph method.
4. The proposed visualization method calls grouped graph method because it views the aggregated rules.
5. The analyst can evaluate the system by using the measure that will be some measures.

Now, the proposed system divides into two main stages: modified AOI algorithm and grouped graph visualization method. This two stage discuss in next sections.

A. Modified Attribute Oriented Induction

AOI technique [13] is used to produce general rules or pattern from large set of rules or patterns. By two steps, attribute removal and attribute generalization perform this induction technique [14].

AOI algorithm has number of steps : generalization of smallest attributes, distinct attribute removal When it lacks top-level, concept tree ascension, vote is accumulative when merging identical tuples in generalization, threshold control on maximum number of distinct attribute , generalization threshold controls on distinct tuples of generalized relation in target class, tuple is convert to conjunction formula and set of tuples are convert to disjunction formula.

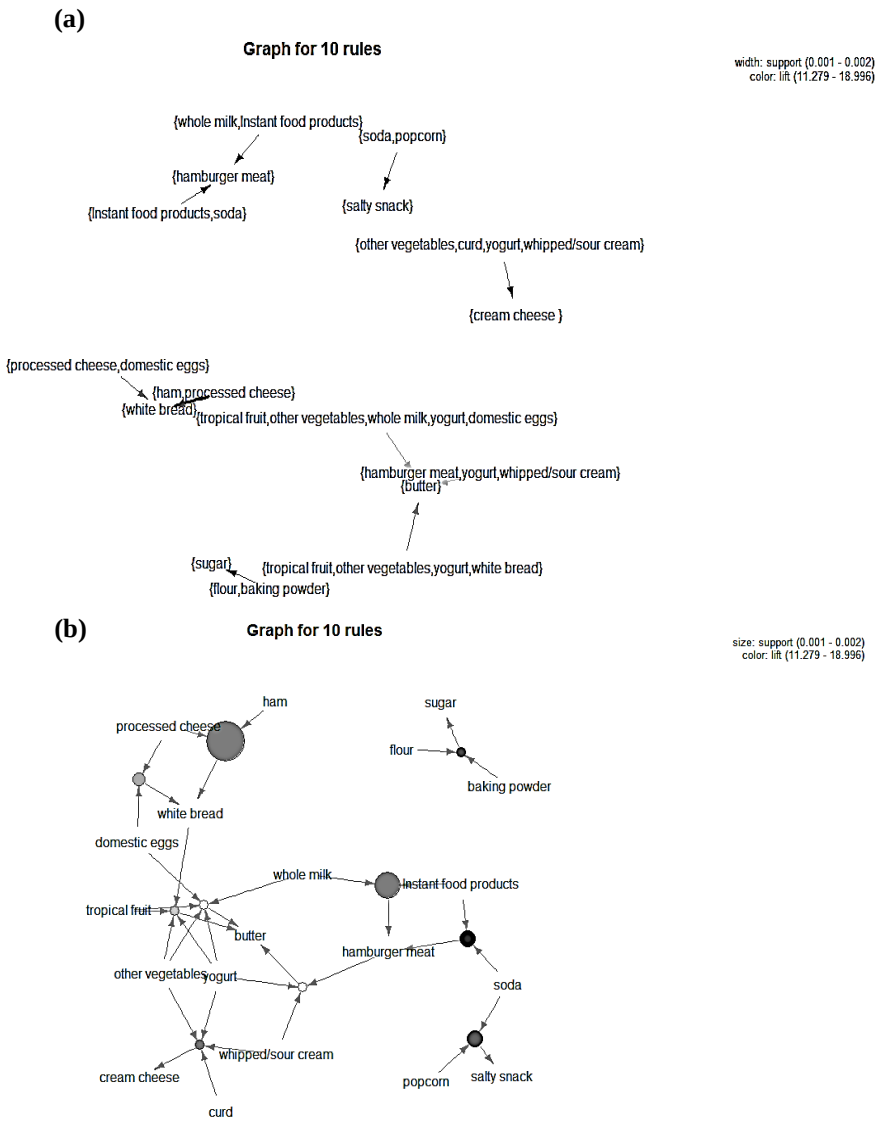

Fig. 1 Graph-based visualization with itemsets as vertices or with items and rules

In this paper, the modified on these steps is satisfied the paper goal in graph visualization method and satisfy new idea as in the following algorithm. In Fig. 3, the inputs for this algorithm are hierarchy trees that are built before the generalization step that contains a number of levels that are defined before any step and a number of levels represented by Generalization Level that entered to the proposed algorithm. In addition, the large number of rules are taken from the Apriori algorithm as input to our algorithm.

Rules reduced by aggregating them, that aggregation will represent the output of this algorithm, and that is the focus of this paper. The resulted algorithm calls **Modified AOI.**

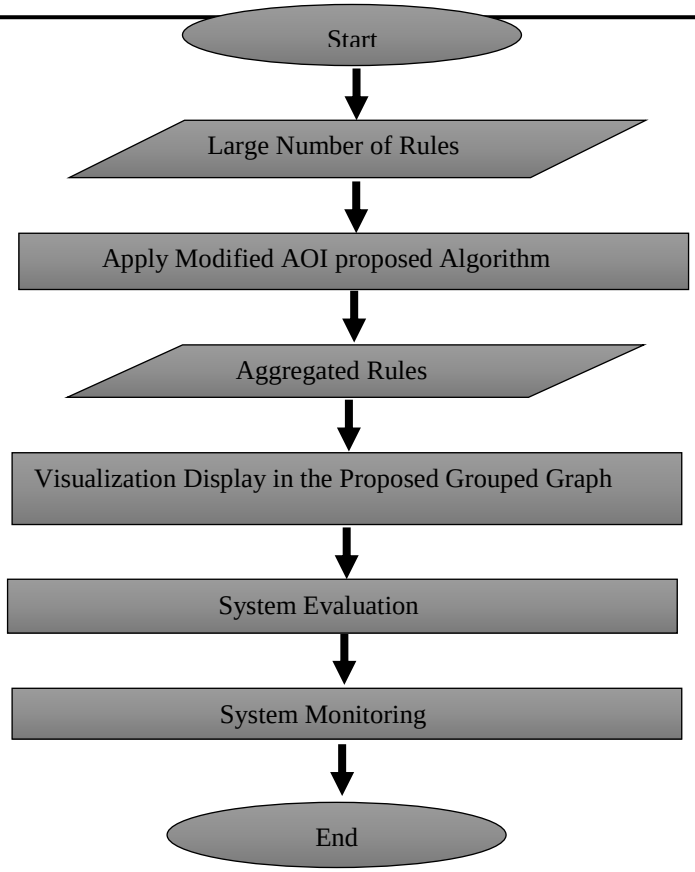

Fig. 2 The general architecture of the proposed system.

Input: hierarchy trees H, set of rules R, Generalization Level (GL).

Output: Aggregated rules.

Method: Modified AOI.

Begin

1. **while** GL!= 0 **do** {
2. **For** each rule R_i (1<=i <=n, where n=# rules) **do**
3. Substitute each $itemset_k$ in the antecedent and consequent of R_i by its corresponding parent in H_k.
4. Merge identical rules
5. **For** each rule R_j (1<= j <n, where n=# rules) **do**
6. Recomput interest measure(s) for generalized rules.
7. GL=GL-1 }

End.

Fig. 3 The Proposed Modified AOI Algorithm.

In step 1 the generalized level is determined and the algorithm is started from it, and in step 7 it is reduced by one in whole the algorithm until it reaches to the root of the concept

hierarchy. From steps 2 and 3 , the rules generalization are made by taking every rule and are replaced each itemsets in corresponding parent in the hierarchy tree, then the result from these steps is a set of redundant rules, and the redundancy of the rules are removed by merging the same rules in step 4 to produce the aggregated rules. In steps 6 and 7, all the existing interesting measure of generalized rules like support, confidence and lift are recomputed that are resulted from the previous steps.

Table 1 is illustrated the difference between traditional AOI algorithm and modified AOI by achieving or not achieving this step.

B. Grouped Graph Visualization

The second main stage is the visualization of the resulting rules. This stage takes the output rules from modified AOI to visualize. The proposed method calls Grouped Graph Visualization because some vertices of the rules in the graph that represents a collection of rules instead of one rule as in the previous graph method.

the LHS (Left Hand Side) and in the RHS (Right Hand Side) from the rules by its

Grouped graph can also visualize every level in the hierarchy tree of the aggregated rules, it can visualize either drill down in the levels to show more detail about the rules or roll up in the levels to show more generalize rules that enable the user to understand large data set and take idea about the nature of the data. Example 1 illustrates this new visualization method.

Example 1: First four transactions are taken randomly from Groceries data set as in Table 2,then apriori algorithm are performed on these transactions to produce 13 rules as in Table 3. The Table 3 visualizes in Fig. 4a.

Secondly, the aggregation performs for rules in Table 3 by the Modified AOI algorithm with levels in Table 4 and show the result in grouped graph visualization. The result from aggregation on Table 3 is 8 rules in level2 as in Table 5 and 6 rules in level1 (more generalize level) as in Table 6. Then the visualization of Table 5 displays in Fig. 4b and Table 6 displays in Fig.49c.

Table 1: Comparison between traditional AOI and modified AOI.

	steps	AOI	Modified AOI
1	generalization on the smallest attribute	✓	✓
2	attribute removal if there is a large set of distinct value but there is no higher-level concept	✓	×
3	concept tree ascension	✓	✓
4	vote value should be accumulative when merging identical tuple in generalization	✓	×
5	threshold control on each attribute that represents maximum number of distinct value of attribute in target class in the final generalized relation	✓	×
6	generalization threshold controls on distinct tuples of generalized relation in target class	✓	×
7	tuple is transformed to conjunction normal form and multiple tuple are transformed to disjunction normal form	✓	×
8	Save the identical tuples before merging its that result from generalization step	×	✓

Table 2: A sample of Groceries transactions

TID	Items
1	{Rolls/Buns, Pastry, Soda}
2	{Whole Milk}
3	{Curd Cheese, Coffee}
4	{Red/Blush Wine, Newspapers}

Table 3: The result 13 rules from apriori Algorithm.

	Rules		Measures		
	Left hand side	Right hand side	Support	Confidence	Lift
1	{curd cheese}	{coffee}	0.25	1	4
2	{coffee}	{curd cheese}	0.25	1	4
3	{red/blush wine}	{newspapers}	0.25	1	4
4	{newspapers}	{red/blush wine}	0.25	1	4
5	{pastry}	{soda}	0.25	1	4
6	{soda}	{pastry}	0.25	1	4
7	{pastry}	{rolls/buns}	0.25	1	4
8	{rolls/buns}	{pastry}	0.25	1	4
9	{soda}	{rolls/buns}	0.25	1	4
10	{rolls/buns}	{soda}	0.25	1	4
11	{pastry, soda}	{rolls/buns}	0.25	1	4
12	{rolls/buns, pastry}	{soda}	0.25	1	4
13	{rolls/buns, soda}	{pastry}	0.25	1	4

Table 4: The levels of Groceries data set.		
Level3	Level2	Level1
rolls/buns	bread and backed goods	fresh products
Pastry	bread and backed goods	fresh products
whole milk	dairy produce	fresh products
curd cheese	cheese	fresh products
Coffee	stimulante drinks	Drinks
red/blush wine	wine	Drinks
Soda	non-alc. drinks	Drinks
newspapers	games/books/hobby	non-food

Table 5: Aggregation 8 rules in level2 from Groceries data set containing 13 rules.					
Rules			Measures		
	Left hand side	Right hand side	Support	Confidence	Lift
1	{cheese}	{stimulant drinks}	0.25	1	4
2	{stimulant drinks}	{cheese}	0.25	1	4
3	{wine}	{games/books/hobby}	0.25	1	4
4	{games/books/hobby}	{wine}	0.25	1	4
5	{bread and backed goods}	{non-alc. drinks}	0.25	1	4
6	{non-alc. drinks}	{bread and backed goods}	0.25	1	4
7	{bread and backed goods}	{bread and backed goods}	0.25	1	4
8	{bread and backed goods, non-alc. drinks}	{bread and backed goods}	0.25	1	4

Table 6: Aggregation 6 rules in level1 from Groceries data set containing 13 rules.

	Rules		Measures		
	Left hand side	Right hand side	Support	Confidence	Lift
1	{fresh products}	{drinks}	0.50	0.6666667	0.8888889
2	{drinks}	{fresh products}	0.50	0.6666667	0.8888889
3	{drinks}	{non-food}	0.25	0.3333333	1.3333333
4	{non-food}	{drinks}	0.25	1.0000000	1.3333333
5	{fresh products}	{fresh products}	0.75	1.0000000	1.3333333
6	{drinks, fresh products}	{fresh products}	0.50	1.0000000	1.3333333

V. EXPERIMENTAL RESULTS

Finally, the general overview is performed about the different data sets in Table 7, the proposed visualization method is compared with different visualization methods to evaluate by the audience, the evaluation of the proposed system is tested by reduction ration measure and monitoring is performed on the performance of the proposed system.

Secondly, The visualization of Association Rules (AR) are often needs four parameters such as sets of LHS items, RHS items, the relation between LHS and RHS, and Interesting Measures (IMs) like support , confidence and lift.

The representation of AR is obtained from The first three parameters , while the evaluation of the AR is obtained from the fourth parameter. The visualization methods are differed in these parameters. Therefore, the comparison between these visualization methods performs according to the above parameters. Now we will discuss four criteria to make difference between the visualization methods as the followings:

Table 7 :Data set description

Data set	Row	Column	Level3	Level2	Level1
Groceries [15]	169	9835	169	55	10
Adult [16]	115	48842	115	112	13
Income[17]	50	6876	50	48	14

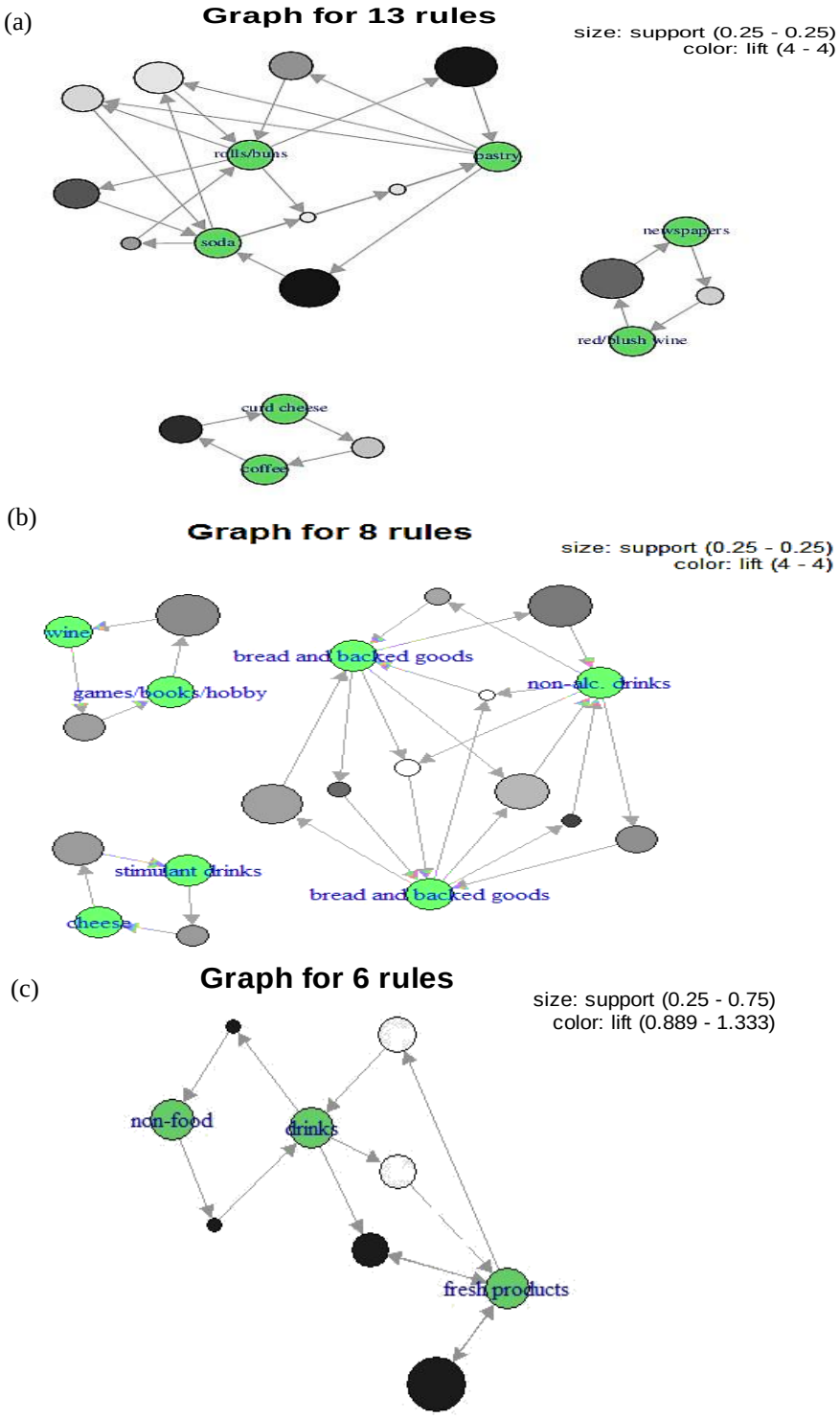

Fig. 4 Graph visualization for original and aggregation rules in level1 and level2.

- Appearance of LHS Items: means Appearance of LHS items of AR in the visualization method. Appearance of LHS items in a clear way helps the analyst to show and evaluate the AR. this criteria ranges [-1, 1]. 1 means Appearance is good, -1 means Appearance is bad.

- Appearance of RHS Item: This criterion is the same as the above criterion except with RHS.

- The clarity of relationship: Easiness knowing the relationship between items in the visualization method means clarity relationship. Therefore, the clarity of relationship gives the analyst easiness to interpret the AR. This criteria ranges [-1, 1]. 1 means Appearance good, -1 means Appearance bad.

- The value of IM: The value of IM instead the shading in the visualization method gives accuracy in evaluation AR. Table 8 is illustrate these criteria.

Thirdly, Reduction Ratio is a ratio of compression of some operation [18] as in Eq. (3) :

$$Reduction\ Ratio\ (RR) = \frac{uncompressed\ number - compressed\ number}{uncompressed\ number} \quad (3)$$

This measure is applied on the result of grouped graph method and to show the ratio of compression from the aggregation by the modified AOI technique. The result from this measure is explained in Table 9 on a number of nodes, edges of the graph and on the number of the rules that resulting from aggregation of the modified AOI technique.

Table 8 Comparison between visualization methods in representation and evaluation of the AR.

Method	Technique	Appearance of RHS Item	Appearance of LHS Item	The clarity of relationship	The value of IM
Scatterplot	Scatterplot	-1	-1	1	×
	Two-Key plot	-1	-1	1	×
Matrix	Matrix-based	-1	-1	1	×
	Matrix-b. (2measures)	-1	-1	1	×
matrix3D	Matrix-b. (3D bar)	-1	-1	-1	×
Grouped	Grouped matrix	1	1	1	×
Paracoord	Parallel coordinates	1	1	1	×
Double-decker	Double decker	1	1	-1	×
Graph	Graph-based	1	1	1	✓
	Grouped graph	1	1	1	✓

Table 9: Reduction ratio of the proposed system on different datasets.

Data Set	Levels	#Nodes	RR %	#Edges	RR %	#Rules	RR %
Groceries	3	5778	-	22220	-	5668	-
	2	1379	76	4838	78	1335	76
	1	211	96	679	97	201	96
Adult	3	23848	-	124343	-	23814	-
	2	12059	49	59338	52	12028	49.5

	1	11751	51	64466	48	11738	51
Income	3	77839	-	415282	-	77781	-
	2	77343	0. 6	413356	0. 5	77288	1
	1	30177	61	174043	58	30163	61

At the last, monitoring of AR performs a monitoring on the changes that take place on the AR before entering to proposed system and after exiting from the proposed system. In this paper, the monitoring performs on the AR before and after the aggregation operation that is performed by the proposed modified AOI algorithm.

The first monitoring is perform on the number of rules before and after the proposed algorithm. In the following Figs., we takes number of the rules and shows the effects of the aggregation on each Levels. The x-axis represents number of the raw AR and before the aggregation. The y-axis represents the AR after the aggregation from modified AOI.

The curves represent the aggregation Levels with different colors. Blue curve represents the Level1 and high Level while red curve represents the Level2 and low Level. The point of intersection the axes only represents y-axis. Figure. 5 displays the AR for Groceries data set.

Know, the second monitoring is perform on memory usage of the number of rules before and after the proposed algorithm. In the following Figs., we takes number of the rules and shows the effects of the aggregation on the memory for each Levels. The x-axis represents number of the raw AR and before the aggregation. The y-axis represents the size of memory that usage by every AR in every levels. The curves represent the usage of memory in different aggregation Levels with different colors. Blue curve represents the Level1 and high Level while red curve represents the Level2 and low Level. The point of intersection the axes only represents y-axis. Figure. 6 displays the AR for Groceries data set.

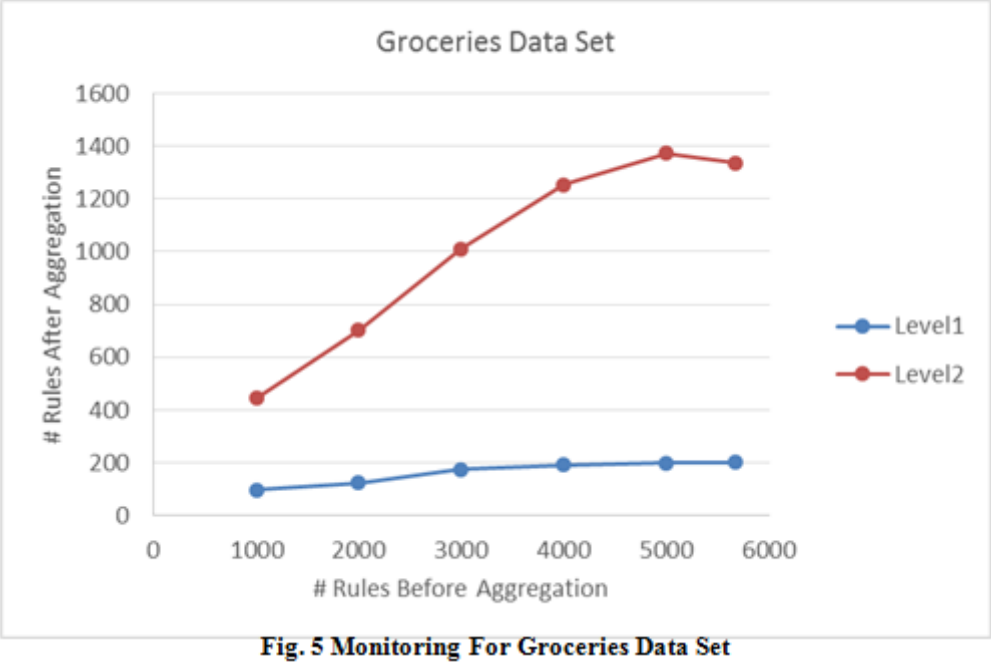

Fig. 5 Monitoring For Groceries Data Set

Fig. 6 Monitoring for Memory Usage for Each Levels of Groceries Data Set

VI. Conclusions

The huge rules results from mining of Association rules that makes difficult for analyzing and understanding these rules. For this purpose, this research combines between the visualization and grouping techniques. The visualization technique solves ease of interpreting of the rules while the grouping technique solves ease of the understanding and reduces the large number of the rules. The proposed technique of this research is graph-based visualization technique and modified AOI algorithm. The result from this system is good in compression and displaying of rules.

References

[1] R. J. B. Jr and R. Agrawa, "Mining the Most Interesting Rules," pp. 145–154, 1999.

[2] H. Hofmann, A. Unwin, and K. Bernt, "The TwoKey Plot for Multiple Association Rules Control," in *Lecture Notes in Computer Science*, no. 1993, 2001, pp. 472–483.

[3] W. A. Hofmann H, Siebes A, "Visualizing Association Rules with Interactive Mosaic," pp. 227–235, 2000.

[4] L. Yang, "Visualizing frequent itemsets, association rules, and sequential patterns in parallel coordinates," *Comput. Sci. Its Appl. …*, pp. 21–30, 2003.

[5] K. Ong, K. Ong, W. Ng, E. Lim, and N. Ave, "CrystalClear: Active Visualization of Association Rules," *ICDM'02 Int. Work. Act. Min. AM2002*, pp. 1–6, 2002.

[6] M. Hahsler and S. Chelluboina, "Visualizing Association Rules in Hierarchical Groups," *42nd Symp. Interface …*, no. Interface, pp. 1–11, 2011.

[7] M. Klemettinen, H. Mannila, R. Pirjo, T. Hannu, and V. A. Inkeri, "Finding Interesting Rules from Large Sets of Discovered Association Rules." pp. 401–407, 1994.

[8] G. Ertek and A. Demiriz, "A Framework for Visualizing Faculty of Engineering and Natural Sciences Department of Industrial Engineering A

Framework for Visualizing Association Mining Results," vol. 4263, pp. 593–602, 2006.

[9] C. P. Rainsford and J. F. Roddick, "Visualisation of temporal interval association rules," *Proc. 2nd Int. Conf. Intell. Data Eng. Autom. Learn. (IDEAL*, pp. 91–96, 2000.

[10] G. K. Gupta, A. Strehl, and J. Ghosh, "Distance Based Clustering of Association Rules," *Proc. ANNIE 1999, St. Louis*, vol. 9, pp. 759–764, 1999.

[11] M. Klemettinen, H. . Toivonen, P. Ronkainen, K. Hätönen, and H. Mannila, "Pruning and Grouping Discovered Association Rules," pp. 1–6, 1995.

[12] B. Lent, a. Swami, and J. Widom, "Clustering association rules," *Proc. 13th Int. Conf. Data Eng.*, pp. 220–231, 1997.

[13] C.-C. H. C.-C. Hsu and S.-H. W. S.-H. Wang, "An integrated framework for visualized and exploratory pattern discovery in mixed data," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 2, pp. 161–173, 2006.

[14] H. J., Y. Cai, and N. Cercone, "Data-Driven Discovery of Quantitative Ruels in Relational Databases," *IEEE Trans. Knowl. Data Eng.*, vol. 5, no. 1, pp. 29–40, 1993.

[15] M. Hahsler, K. Hornik, and T. Reutterer, "Implications of Probabilistic Data Modeling for Mining Association Rules," *Proc. 29th Annu. Conf. Gesellschaft für Klassif. eV Univ. Magdebg. March 911 2005*, vol. 2, no. 14, pp. 598–605, 2005.

[16] a Asuncion and D. J. Newman, "UCI Machine Learning Repository," *University of California Irvine School of Information*, 2007. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html .

[17] T. Hastie, R. Tibshirani, and J. Friedman, "Elements of Statistical Learning: Data Mining, Inference and Prediction," 2001. .

[18] F. Valeur, "Real-Time Intrusion Detection Alert Correlation," *Security*, no. June, pp. 1–199, 2006.