# Interactive Data Mining: A Short Background Study on Effective Interaction and Visualization by Association Rules

Valliappan Raman[1], Sundresan Perumal[1], and Putra Sumari[2]

***Abstract*****—** Interactive data mining and visualization is essential to achieve an effective data mining result. This paper is a proposal to develop an interactive visualization approach for association rules. This is mainly due to problems of the existing approaches as stated in this paper. Interactive data mining, association rules and limitation of existing visualization methods were discussed. Specific aims and objectives of this research to solve the problems were presented and a conclusion was drawn at the end.

***Keywords*****—**Data Mining, Association Rules, Visualization

## I. Introduction

Huge amount of data are stored all the time in all types of organizations, such as in the medical sector, financial sector, educational sector e.t.c. Nearly all these organizations have a database with stored data with hidden or undiscovered valuable information that could be processed and extracted to yield useful knowledge which can be of a great advantage to the company or organization. This is where data mining comes in.

Data mining or knowledge discovery is the process of discovering meaningful new correlation, patterns, and trends by digging into large amounts of data stored in warehouse, using statistical, machine learning, artificial intelligence and data visualization techniques [1]. The data mining processes are divided into hypothesis formulation, data collection, data pre-processing, model estimation and interpretation. Interactivity in these processes with humans can encourage learning, improve insights and understandings of the domain, stimulate the exploration of creative possibilities, and help users to solve particular problems [7]. To achieve effective data mining results, the output must be understood by the user in such a way that he can apply it successfully in the problem domain. This is an area of concern because so many patterns are generated as data mining results and without a proper and effective representation methods, these patterns will be too complex to be comprehended by the user. An example of such data mining results is association rules. Some attempts have been made to represent association rules by visualization but these attempts have their limitations either in the number of rules that can be visualize effectively or lack of proper representation of items and measures that define the rule. Our research is going be on integrating interactivity in each stage of mining of association rules and will focus more on developing a highly effective technique for visualization and interaction with the rules in order to overcome limitations of the existing approaches.

This aim of the paper is to make a background study on understanding the interactive data mining by association rules. Data mining review was explained briefly in section 2. Detailed background study was made in section 3. With understanding of the background study, proposed methods with objectives were made in section 4. Finally conclusion was made in section 5.

## II. Review

Interactive data mining is justified with the need for a balance between computer and human control in the data mining processes. Combination of human intelligence and creativity and algorithms implemented in a computer as well as its processing speed will result in an effective data mining. In this form of data mining, all the processes are done interactively and iteratively with a human. Detail research study was made by Wong [15]

- Interactive data preparation - Collect raw data with a specific format
- Interactive data selection and reduction – Selecting data sets that falls into the area of interest and remove those records that do not.
- Interactive data pre-processing and transformation - removing unwanted data and transforming the data set in to a workable one.
- Interactive pattern discovery – Pattern discovery with a human user interactively monitoring the process.
- Interactive Pattern explanation and evaluation – Human user explains and evaluates the discovered pattern. The interpretation is based on the view of the user.
- Interactive pattern representation in a visualized format.
- The interaction done during these phases are in a form of proposition from the user, information acquisition, guidance acquisition and manipulation of objects.

Dr.Valliappan Raman [1]is with the MRG Lab, Universiti Sains Malaysia, Penang, Malaysia

Dr. Sundresan Perumal [1]is with the Universit Sains Islamic Malaysia, Nilai, Malaysia.

Dr. Putra Sumari [2]is with the UniversityiSains Malaysia, Penang, Malaysia.

http://dx.doi.org/10.15242/IIE.E0815001

Chakravarthy [13] cited just to show the limitation of using table as a visualization tool. One of the visualization techniques they used is an interactive rule table to explore mined association rules. But only few rules at a time can be viewed. Buono [16] discussed a visual data mining framework they have developed called DAE (Data Analysis Engine) for data analysis as well as visual exploration of results of clustering algorithms and association rules. They used parallel coordinate and network graph for the visualization. Limitation of parallel coordinate had already been discussed above. Network graph used clearly presented the problem of overlapping of edges and the use of color to visualize antecedent, consequent and support measure didn't aid their idea of reducing cluttering. It still appeared heavily cluttered and a lot of occlusion problems.

In detailed review, similar aspects as above work. The existing works stated the problem of lack of integration of data visualization to show the data associated with a rule. It also states that, there has not been any efficient approach proposed to provide visualization to support interactive rule derivation and evaluation of the derived rules.

## III. Background

### A. Interactive Data mining

Research in interactive data mining has not been fully pursued, although its importance has been clearly stated. Computer systems rely on human users to set goals, to select alternatives if an original approach fails, to participate in unanticipated emergencies and novel situations, and to develop innovations in order to preserve safety, avoid expensive failure, or increase product quality [10][11][12]. For data mining to be effective there is need for human – computer interaction in the mining process. The user with his knowledge and intuition about the application domain should be able to participate in the search for new structures in data and to guide search strategies [1]. By incorporating interactivity, the traditional process can be changed into interactive data preparation, interactive data selection and reduction, interactive data pre-processing, interactive pattern discovery, interactive pattern explanation and evaluation, and interactive pattern presentation [7]. An interactive classification system was proposed [11] to illustrate the concept of interactive data mining process as discussed above. For example, in the data preparation task, the system allows the user to create and explore his/her dataset using tools such as schema editor, data editor and query builder, in data selection, the user can use the data view module to explore and edit both the training and testing data. One of the drawbacks of this approach is that most of the results are represented in a textual format which is not appropriate when dealing with large dataset. This will be further elaborated as the concept of association rule mining. Other software and approaches where presented in [2] [3] [4]] [6].

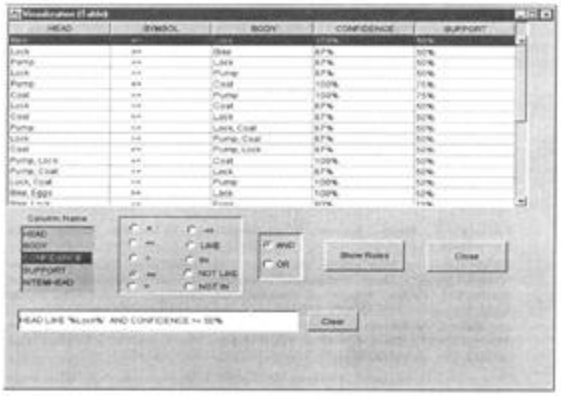
Requirement in interactive data mining development include studying of issues like user modeling, behavior simulation, situation analysis, user interface design, user knowledge management, algorithm/model input setting by users, mining process control and monitor, outcome refinement and tuning but some of these issues cannot be fully addressed by existing data mining approach [5].

### B. Association Rule

Mining of association rule is one of the most popular data mining methods. Let I = {I1, I2,..., Im} be a set of binary attributes called items, and D be a database consisting of subsets of I [12]. Each record, Ti ∈ D(i = 1, 2,..., n) contains a set of items such that Ti ⊆ I. An association rule shows a relationship in the form X→ Y, where X,Y ⊂ I are called the antecedent item and Y the consequent item respectively and both X and Y are mutually exclusive i.e. X ∩ Y = Ɵ. A rule has two attributes, support showing the frequency of the rule and confidence representing the quality or the probability of the occurrence of that rule. A support s of a rule X→ Y is s% of records in D that contain X U Y and the confidence of a rule X→ Y is c if c% of records in D that contain X also contain Y. An item set is frequent if it has a support that is not less than the minimum support specified by the user. Thus, the aim of association rule is to find all frequent itemset and rules that have confidence greater than minimum confidence.

### C. Visualization Techniques

Visualization is needed to represent not only the mined rules but also, the data during the mining process. This can aid the user in understanding the derived rules more. Several visualization methods have been proposed or used in the existing literature. An example is the use of table in [fig.1] to visualize rules [13]. The limitation of this approach is that it applies a type of restriction on the number of rules that can visualize at a time. Another example is the two-dimensional matrix [fig.2] which is a bar diagram with consequents on one axis and antecedents on the other [14]. It has the drawback of representing only one-to-one rules.



| HEAD | SYMBOL | BODY | CONFIDENCE | SUPPORT |
|---|---|---|---|---|
| [illegible] | [illegible] | [illegible] | [illegible] | 50% |
| Lock | <= | Beer | 67% | 50% |
| Pump | <= | Lock | 67% | 50% |
| Lock | <= | Pump | 67% | 50% |
| Pump | <= | Coat | 100% | 75% |
| Coat | <= | Pump | 100% | 75% |
| Lock | <= | Coat | 67% | 50% |
| Coat | <= | Lock | 67% | 50% |
| Pump | <= | Lock, Coat | 67% | 50% |
| Lock | <= | Pump, Coat | 67% | 50% |
| Coat | <= | Pump, Lock | 67% | 50% |
| Pump, Lock | <= | Coat | 100% | 50% |
| Pump, Coat | <= | Lock | 67% | 50% |
| Lock, Coat | <= | Pump | 100% | 50% |
| Beer, Eggs | <= | Lock | 100% | 50% |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |

Column Name: HEAD, BODY, CONFIDENCE, SUPPORT, ITEM-HEAD

○ = ○ <> ○ > ● >= ○ < ○ <= ○ LIKE ○ IN ○ NOT LIKE ○ NOT IN

● AND ○ OR

HEAD LIKE '%Lock%' AND CONFIDENCE >= 50%

Fig 1 Illustrate the Rule table

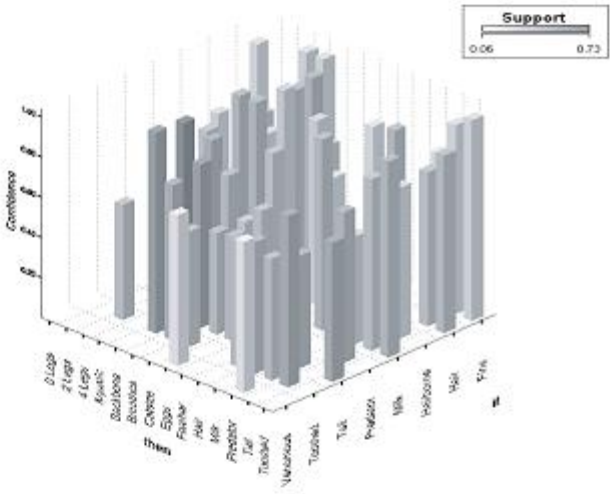


Fig 2 illustrates the Two Dimensional Matrix

A 3-D approach in visualizing many – to – one rule was proposed in [15] to overcome the limitation of the 2-D matrix. It places items in rows and rules in columns. As seen in [fig.3], there is going to be a problem of occlusion when dealing with large number of rules. Also, size of the matrix limits the number of rules that can be displayed. A graph based technique [fig.4] was adopted in [16] in form of a network representing each single side of the rules as a node, the edge of the graph represents the logical implication of a rule. This has the common problem of overlapping of edges when the size of the rules is large. There is also another method called two key Plot [fig.5] which shows the two keys of confidence and support for all discovered association rules [17]. However, this technique has the problem of absence of items in the general view of the rules. Parallel Coordinate [fig.6], another method used in visualizing multidimensional data has a series of parallel vertical axes, each representing a separate variable placed evenly in a horizontal direction [18]. A given data record is represented as polylines between successive vertical axes. As in the case of network graphs, this also has the limitation of overlapping of the polylines.

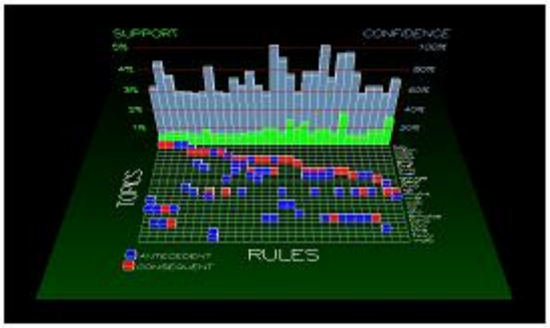


Fig 3 illustrates the Dimensional Matrix

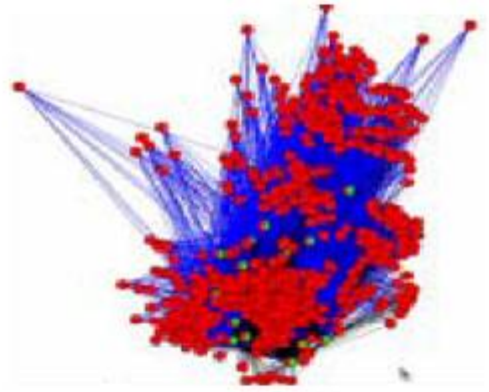


Fig 4 Illustrates the Network Graph

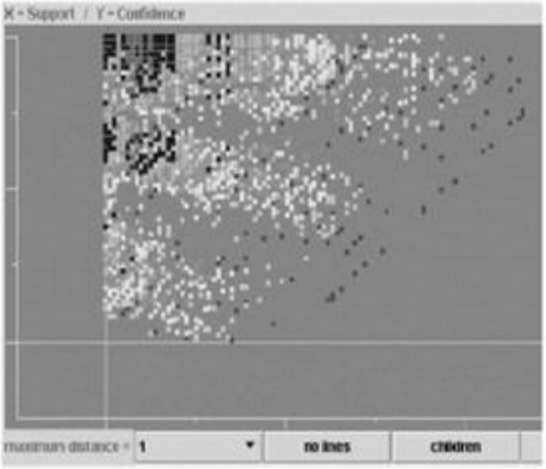


Fig 5 Illustrates the Key Plots

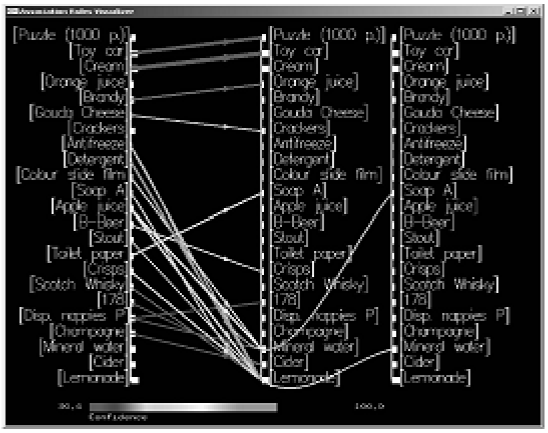


Fig 6 Illustrates the parallel coordinates

Another limitation of visualization of association rules is in terms of the type of rule visualized i.e. one-to-one, many-to-one, many-to-many. There are lots of tools to visualize one-to-one and many-to-one, but a problem arises when visualizing many-to-many rules. An interactive visualization system, VisAR [fig.7] was developed to overcome this problem, also the problem of screen clutter and occlusion [19]. It has a rule-to-item representation. Colour intensity was used to measure the support and confidence of a rule. However, according to McGill and Cleveland [20], shapes and colours are not effective in coding quantitative information. To address this problem, a rule – to – item representation that uses a vertical bar at the top of each column to visualize the support and confidence values of its corresponding rule was proposed in [21]. But as seen in [fig.8], there could be confusion on the implication of the rule. For example, if Black → Red, White → Red, it doesn't necessarily imply Black, White → Red as depicted in the graph.

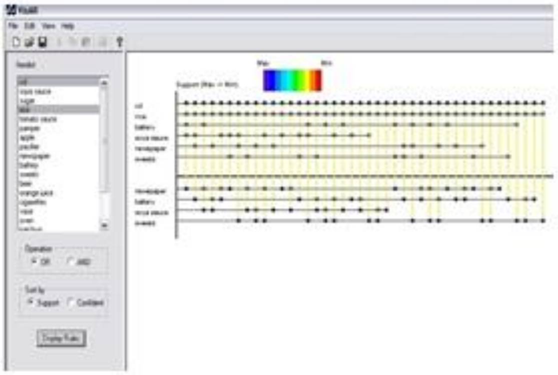
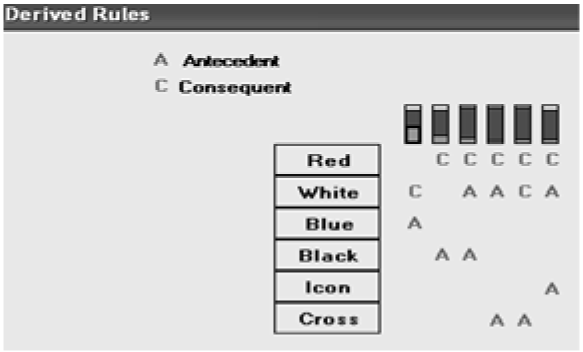


Fig. 7 Illustrates the VisAR Interface



Derived Rules

A Antecedent
C Consequent

| | | | | | | |
|---|---|---|---|---|---|---|
| Red | | C | C | C | C | C |
| White | C | | A | A | C | A |
| Blue | A | | | | | |
| Black | | A | A | | | |
| Icon | | | | | | A |
| Cross | | | | A | A | |

Fig.8 Illustrates the Rule to Item Representation

## IV. PROPOSED OBJECTIVES

Based on the various limitations stated above, the proposed research objectives are:

Integration of interactivity in the association rule mining process from the first stage of the mining process is performed (i.e. data collection to rule presentation and evaluation). An existing association rule algorithm will be used but with the user interacting and guiding the rule generation iteratively.

Due to the trade of between effective visualization of items and support measures in existing visualization approaches. We would research and come up with a visualization method that will be able to overcome this drawback using a hybrid approach of the existing techniques and a novel approach.

A way to aid the users rule exploration process by suggestions. This is to trigger the user's intuition and understanding of the rules. For example, something like a land map in which as a rule in form of a point is clicked, other rules related with that rule can be highlighted as suggestions to the user or to show the user that those rules are related and worth exploring.

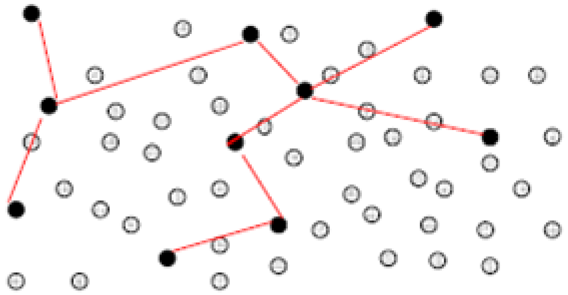


Fig 8 Illustrates the Proposed Framework

This proposed idea is illustrated in the figure above. The dark points and lines appear when one point representing a rule is clicked. The other rules are made less visible may be using a distortion technique. But decision to follow this suggestion will be left to the user.

## V. CONCLUSION

Development of an effective interactive and visual data mining approach will aid immensely on understanding and application of the discovered knowledge on the appropriate domain. This proposal discussed various techniques and limitation of existing literature to show the importance of this research. The objective stated above will be achieved and a highly interactive visualization approach for mining of association rules will be developed as the research outcome.

## REFERENCES

[1] S.N. Sivanandam, S. Sumathi, "Introduction to Data Mining and its applications, Studies in Computational Intelligence", Volume 29, Springer, 2006.

[2] Visumap, 2009, "Visualizing high dimensional complex data", viewed 17th January, 2010. http://www.visumap.net/.

[3] Algorithmic Solutions, 2007, Purple Insight, Mine Set, viewed 17th January, 2010. http://www.algorithmic-solutions.com/leda/projects/mineset.htm

[4] Y. Burcu, G. Mehmet, 2009, Interactive Data Mining for Molecular Graphs, viewed 19th January, 2010. http://downloads.hindawi.com/journals/jammc/2009/502527.pdf

[5] Longbing Cao, "Data Mining and Multi-agent Integration", Springer, 2009. http://dx.doi.org/10.1007/978-1-4419-0522-2

[6] Wu X.D, Zhu X.Q, Chen Q.J, "Ubiquitous Mining with Interactive Data Mining Agents", Journal of Computer Science and Technology, 24(6), Nov. 2009, pp. 1018–1027. http://dx.doi.org/10.1007/s11390-009-9291-7

[7] Yan Z., Yaohua C, Yiyu Y. 2006, "User Centered interactive Data Mining", Proceeding of the Sixth IEEE International Conference on Cognitive Informatics (ICCI'06), p.457-466.

[8] Shneiderman B., 1998, "Designing the user interface: Strategies for Effective Human-Computer Interaction", 3rd edition, Addison-Wesley.

[9] Hancock, P.A. and Scallen S.F, 1996, "The future of function allocation, Ergonomics in Design", 4(4), p.24-29. http://dx.doi.org/10.1177/106480469600400406

[10] Elm, W.C., Cook, M.J., Greitzer, F.L., Hoffman, R.R., Moon, B. & Hutchins, S.G, "Designing support for intelligence analysis", Proceedings of the Human Factors and Ergonomics Society, 2004, p.20-24.

[11] Wang, "On Interactive Data Mining", Encyclopedia of Data warehousing and Mining, 2nd edition, 2008, pp. 1085-1090.

[12] Agrawal R., Imielinski T, Swami A,"Mining association rules between sets of items in large databases", In Proc. of the 1993 ACM SIGMOD international conference on management of data, ACM Press,1993, pp. 207- 216.
http://dx.doi.org/10.1145/170035.170072

[13] Chakravarthy S., Zhang H, "Visualization of Association Rules over Relational DBMS's", Proceedings of the 2003 ACM symposium on Applied Computing, 2003, pp. 922-926.
http://dx.doi.org/10.1145/952532.952714

[14] Bruzzese D., Davino C, "Visual Mining of Association Rules", Visual Data Mining, LNCS 4404, 2008, pp. 103–122.
http://dx.doi.org/10.1007/978-3-540-71080-6_8

[15] Wong P.C, Whitney P., Thomas J, "Visualizing Rules for Text Mining", Proceedings of IEEE Information Visualization, IEEE CS Press, Los Alamitos,1999.

[16] Buono P., Costabile M.F, "Visualizing Association Rules in a Framework for Visual Data Mining", E.J Neuhold Festchrift, LNCS 3379, Springer – Verlag Berlin Heidelberg, 2005, pp. 221-231.

[17] Unwin A., Hofmann H., Bernt K., 2001, "The Two Key Plots for Multiple Association Rules Control. L.De Raedt and A. Siebes (Eds): PKDD, LNAI 2168, Springer – Verlag Berling Heidelberg 2001, pp. 472-483.

[18] Li Yang, "Pruning and Visualizing Generalized Association Rules in Parallel Coordinates", IEEE Transactions on Knowledge and Date Engineering, Vol.17, No.1, January 2005.

[19] Techapichetvanich K., Datta A., 2005, "VisAR: A New Technique for Visualizing mined Association Rules", ADMA, LNAI 3584, Springer – Verlag Berlin Heidelberg, 2005, pp. 88-95.

[20] Cleveland W.S, McGill R, "Graphical Perception: The Visual Decoding of Quantitative Information on Graphical Displays of Data", Journal of the Royal Statistical Society. Series A (General), Vol. 150, No.3, 1987, pp. 192-229.
http://dx.doi.org/10.2307/2981473

[21] Yan Liu, "Design and Evaluation Support to Facilitate Association Rules Modeling", 2006,
http://www.wright.edu/~yan.liu/Pulications/AssociationRules.pdf

**Valliappan Raman**, is a research fellow in school of computer science, Universiti Sains Malaysia. He have worked as team to acquire many external research grants and published papers in impact factor journals. His research interests are in medical imaging, data mining and health informatics.

**Dr.Sundresan Perumal** is a senior lecturer in Universiti Sains Islamic Malaysia. He acquired various external research grants and published papers in impact factor journals. His research interests are in cyber security, medical imaging, data mining and network forensics.

**Assc Prof. Putra Sumari** is Associate Professor at University Sains Malaysia, Penang, Malaysia and leading the research team in Multimedia Research Group Lab, USM. He has supervised more than 100 students in post graduate level. He has acquired many external research grants from the government. He has numerous papers published in proceedings and impact factor journals. His research interests are in, multimedia content and storage, medical image processing and data mining.