Big Data Visualization for Occupational Health and Security problem in Oil and Gas Industry

Daniela Gorski Trevisan[2], Nayat Sanchez-Pi[1], Luis Marti[3], and Ana Cristina Bicharra Garcia[2]

[1] Instituto de Logica, Filosofia e Teoria da Ciência (ILTC), Niterói (RJ) Brazil.
nayat@iltc.br
[2] Fluminense Federal University, Computer Science Institute, Niteroi (RJ) Brazil.
daniela,bicharra@ic.uff.br
[3] Dept. of Electrical Engineering, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro (RJ) Brazil.
lmarti@ele.puc-rio.br

Abstract. Association rule learning is a popular and well-researched set of methods for discovering interesting relations between entities in large databases in real-world problems. In this regard, an intelligent offshore oil industry environment is a very complex scenario and Occupational Health and Security (OHS) is a priority issue as it is an important factor to reduce the number of accidents and incidents records. In the oil industry, there exist standards to identify and record workplace accidents and incidents in order to provide guiding means on prevention efforts, indicating specific failures or reference, means of correction of conditions or circumstances that culminated in accident. OHS's employees are in charge of analyzing the mined rules to extract knowledge. In most of cases these users has two main challenges during this process: i) to explore the measures of interestingness (confidence, lift, support, etc.) and ii) to understand and analyze the large number of association rules. In this sense, an intuitive visualization of mined rules becomes a key component in a decision-making process. In this paper, we propose a novel visualization of spatio-temporal rules that provides the big picture about risk analysis in a real world environment. Our main contribution lies in an interactive visualization of accident interpretations by means of well-defined spatio-temporal constraints, in the oil industry domain.

Keywords: data visualization, big data applications, decision support systems, oil and gas industry

1 Introduction

Occupational health and security (OHS) issues are priority matter for the offshore oil and gas industry. This industry is frequently in the news. Much of the time, it is because of changes in prices of oil and gas. Other |less frequent but perhaps more important| subject of media attention is when disasters strike, as is the case of offshore oil drilling

platform explosions, spills or fires. These incidents have a high impact on lives, environment and public opinion regarding this sector. That is why a correct handling of OHS is a determining factor in this industry long-term success.

There is an important effort of oil and gas industry to reduce the number of accidents and incidents. There are standards to identify and record workplace accidents and incidents to provide guiding means on prevention efforts, indicating specific failures or reference, means of correction of conditions or circumstances that culminated in accident. Besides, oil and gas industry is increasingly concerned with achieving and demonstrating good performance of occupational health and safety (OHS), through the control of its risks, consistent with its policy and objectives. Today, with the advances of new technologies, accidents, incidents and occupational health records are stored in heterogeneous repositories.

Similarly, the amount of information of OHS that is daily generated has become increasingly large. Furthermore, most of this information is stored as unstructured or poorly structured data. This poses a challenge, which is a top priority, for industries that are looking for ways to search, sort, analyze and extract knowledge from masses of data. Data mining can be applied to any domain where large databases are saved. Some applications are failure prediction [2], biomedical applications [7], process and quality control [6]. Association rule learning is a popular and well-researched set of methods for discovering interesting relations between entities in a large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. Many algorithms for generating association rules were presented over time.

Some well-known algorithms are Apriori [1], Eclat [14] and FP-Growth [9], but they only do half the job, since they are algorithms for mining frequent item sets. Another step needs to be done after to generate rules from frequent item sets found in a database.

In most of cases users has two main challenges during this process: i) to explore the measures of interestingness (confidence, lift, support, etc.) and ii) to understand and analyze the large number of association rules. In this sense, an intuitive visualization of mined rules becomes a key component in a decision-making process. In this paper we propose a novel visualization of spatio-temporal rules that provides the big picture about risk analysis in a real world environment.

Our main contribution lies in an interactive visualization of accident interpretations by means of well-defined spatio-temporal constraints, in the oil industry domain. The paper is organized as follows. After introducing the OHS problem, Section 2 briefly describes the state of the art on visualization techniques for association rules. After that, a case study involving rules visualization is presented in Section 3. Finally, Section 4 presents some conclusive remarks and outlines the current and future work been carried out in this area.

2 Foundations

Association rule mining algorithms typically generate a large number of association rules, which poses a major problem for understanding and analyzing rules. In this sense,

several visualization techniques are proposed in order to facilitate this reasoning process. Parallel coordinates, introduced by Inselberg in 1981 [12], represent a very useful graphical tool to visualize high dimensional data-sets in a two-dimensional space. They appear as a set of vertical axes where each axis describes a dimension of the domain and each case is represented by a line joining its values on the parallel axes. The Mosaic plots [10] and its variant of Double decker plots can be used to visualize the contingency table. They were introduced to visualize each element of a multivariate contingency table as a tile (or bin) in the plot and they have been adapted to visualize all the attributes involved in a rule by drawing a bar chart for the consequence item and using linking highlighting for the antecedent items.

The main drawback of Double Decker plot lies in the possibility to represent one rule at a time or at least all the rules generated from the different combinations of the items belonging to a given rule. In order to have the possibility to represent simultaneously many rules, Hofmann and Wilhelm [11] proposed the matrix of Association Rules with and without additional highlighting but in this case, only one-to-one rules are taken into consideration. In [8] the authors present and compare a set of visualization techniques implemented in arulesViz tool, which can be used to explore and present sets of association rules. The comparison criteria is based on the size of the rule set which can be analyzed, the number of interest measures which are shown simultaneously, if the technique offers interaction and reordering and how intuitive each visualization technique is. They found that Scatterplot (including two-key plots) and grouped matrix plot are capable to analyze large rule sets.

These techniques are interactive to allow the analyst to zoom and select interesting rules. Matrix-based can accommodate rule sets of medium size. Reordering can be used to improve the presentation. To analyze small rule sets the matrix-based method with 3D bars, graph-based methods and parallel coordinates plots [3-5] are suitable. Graphs for large rule sets can be analyzed using external tools and at last, double decker plots only can visualize a single rule. The techniques discussed in [8] can also be categorized based on the number of interest measures simultaneously visualized. Most methods can represent two measures and scatter plots are even able to visualize three measures for each rule in one plot. Scatter plot and graph based techniques are the most intuitive while matrix-based visualization with two interest measures, parallel coordinates and double decker require time consuming to learn how to interpret them correctly. It is important to note that most of these categories are only evaluated qualitatively, and the results presented are only meant to guide the user towards the most suitable techniques for a given application.

The only work that we have found reporting the user evaluation of the applied visualization techniques is described in [13]. They are using a matrix view to provide an overview of all of the association rules, allowing the user to filter and select rules that are potentially useful. The second visualization technique is a graph view and it shows the subset of rules selected from the matrix view, illustrating the relationships between the LHS (left-hand side) items and the RHS (right-hand side) items. At any time, the features of specific rules can be accessed within the detail view by highlighting the rules within the matrix or graph views. The matrix view, graph view, and detail view

of the association rules are visualized separately since during association rule exploration, users often seek rules based on their interestingness measures first. In this study they performed the user testing evaluation with 12 participants, which were, characterize as knowledgeable users, i.e. with knowledge on databases or data mining techniques.

These users should find the ten significant association rules in different dataset sizes. As results, they found that even though the number of rules increased by factors of 5 and 10 over the smallest group of rules, the time to find the required set of rules, the error rates, the perceived confidence, and the perceived ease did not change. In particular, the ability to filter the association rules using the matrix view, and then examines the rules using the graph view and detailed view made the task equally as easy even as the number of rules available for examination increased. The main drawback of such approach is the fact the users are not interpreting the rule itself they are mainly using a data mining visualization tool to filter and to find rules. We cannot conclude if experts users in the application domain but not in data mining techniques could be able to reach the same performance.

3 Visualization Approach

The approach proposed in this paper includes a visualization technique for mined association rules. It is applied in the post-processing stage in a straightforward way, and visualization results are used together with the mined rules. As already explained, the task of providing context-based information calls for the processing and extraction of information in the form of rules. One of the possible ways of obtaining those rules is to apply an association rule algorithm. In this work we employ Apriori [1] and FP-Growth [9] algorithms in parallel in order to mutually validate the results from each other. Association rules are not always simple to understand because results set are often very large and/or because the rule itself demands explanation. For the former problem, many techniques have been proposed to filter the most relevant set of rules. However, the later problem has received less attention.

Many techniques are quite difficult to understand and to correlate items making hard for the user to take a decision. In this direction, our work proposes two kinds of rules visualization. The first one is focused on visualization of one-to-one rules and the second one is focused on the n-to-n rules. In fact, for both visualizations we have two filters options, one filter that can be applied before mining, it is useful to filtering attributes to be mined, and the other filter can be applied after the mining process to filtering rules. For instance, the rule filter allows the user to choose to see all rules that have the ACIDENT anomaly in the right hand (consequent) of the rule.

3.1 Visualization of Intra Association Rules

For the Intra association rules visualization we have chosen to show in two circular viewers all attributes involved in the mining process. The attributes are shown in the exterior part of each circle while their values are show in the interior area of each circle.

There is a slider to navigate between rules and when the next rule is selected its correspondent attributes and values are highlighted in the circles indicating which are the antecedent and consequent items involved in that rule. As a result of the attributes filter we can see in Fig.1 the visualization of 3 attributes involved in that rules while in the visualization illustrated in Fig.2 shows only 4 attributes involved in that set of rules. Also it is possible to filter rules based on the number of occurrences (i.e. the lift criteria filter) and its confidence level.

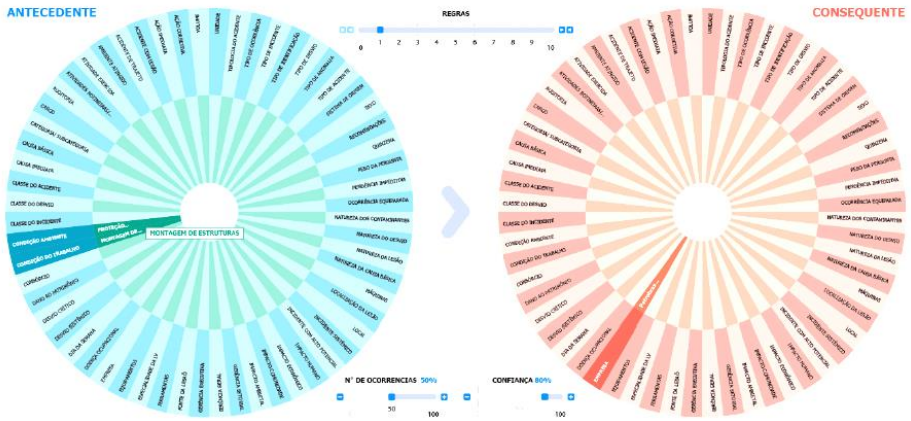

Fig. 1. Visualization of an Intra-anomaly rule with 3 attributes involved in that rule. The circle visualization is dealing with 60 attributes mined.

The minimum values for these filters are defined before mining. Therefore, after mining the user can filter rules from this minimum predefined value until the maximum value while there are rules that can´t these values well. An advantage of this visualization is to compare rules, it is easy to see which attributes has been changed along rules and which ones did not change.

Fig. 2. Visualization of an Intra-anomaly rule with 4 attributes involved in that rule. The circle visualization is dealing with 30 attributes mined.

3.2 Visualization of Inter Association Rules

The visualization of spatiotemporal association rules calls for more complex interactive mechanisms in the attempt to explain the rules and its relationships once it can be any combination of *n-to-n* rules. An example of inter-anomaly mined rule interpretation is presented in Fig.3 and its visualization is presented in Fig.4 and Fig.5. We start by defining three different colors to identify each kind of anomaly. Fig.6 is showing the three anomalies representation where blue color is related to the Deviation anomaly, the green color is related to the Incident anomaly and the red color is related to the Accident anomaly. For each rule, we visualize the anomalies (items) involved in the antecedent and consequent sides of rule. For instance, in Fig. 4 (see area 1) we have a visualization of *n-to-1* rule configuration and in Fig. 6 (see area 1) we can see visualization for *n-to-n* rule configuration. Area 1 is always showing the antecedent (left hand) and consequent (right hand) sides of the rule. Interactions with the consequent side of the rule (in area 1) will generate changes in the visualization of occurrences in area 2. Interactions with the antecedent side of the rule (in area 1) will generate changes in the visualization of occurrences in area 3.

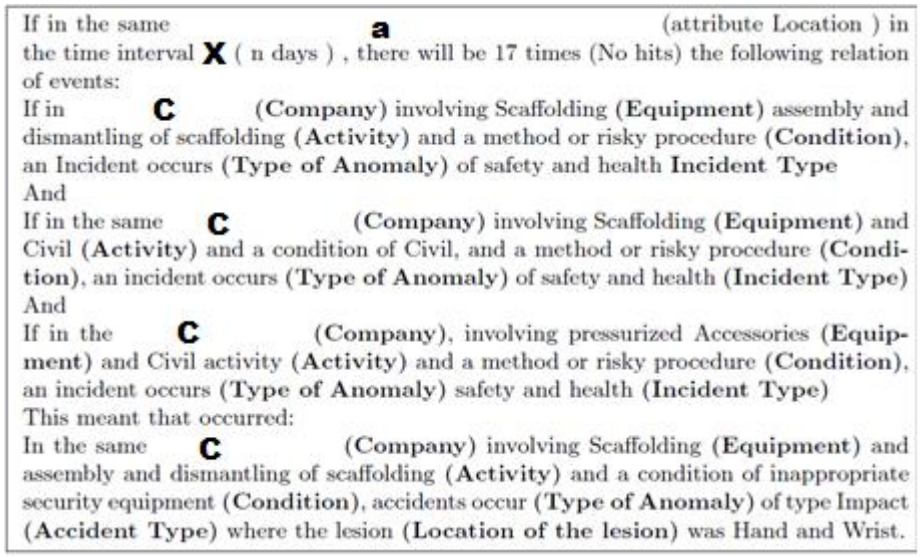

Fig. 3. Example of Intra-anomaly mined rule involving Incidents and Accidents anomalies

By dragging and dropping one occurrence circle (for instance from the area 2 in Fig. 4) to the timeline area (area 3 in Fig. 4 and in Fig.5) the user can see all events involved in that occurrence according to the a given amount of precedence time. For instance, the user could have defined before mining that he would like to analyze events occurred 30 days before of this event happened. Also by selecting a circle anomaly in the consequent side (area 1 in Fig. 6) the user can see all correlated occurrences of that anomaly (see area 2 in Fig. 6). When the user selects a circle anomaly at the antecedent side (area

1 in Fig. 5 and in Fig. 6) all occurrences in the timeline area which are shared with him are highlighted (see the highlighted circles in area 3 of Fig. 5 and Fig. 6).

Fig. 4. Inter-anomaly interactive visualization of *n-to-1* rules

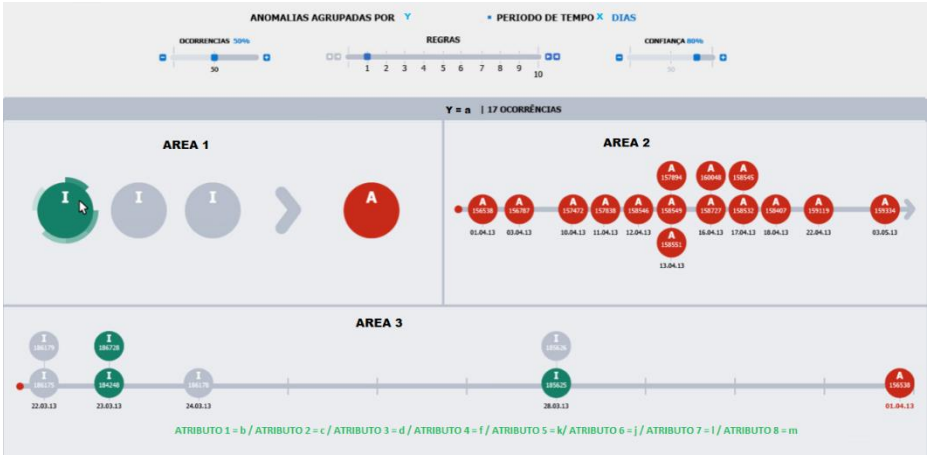

Fig. 5. Inter-anomaly interactive visualization of correlated events

Besides that, it is possible to investigate anomalies shared attributes. By clicking in one anomaly circle in the antecedent side (see Fig.7) the user can see which are the attributes involved in that anomaly as well which of these attributes also appear in other anomalies. Finally, visualization resources for zoom in and zoom out in the timeline area are also available.

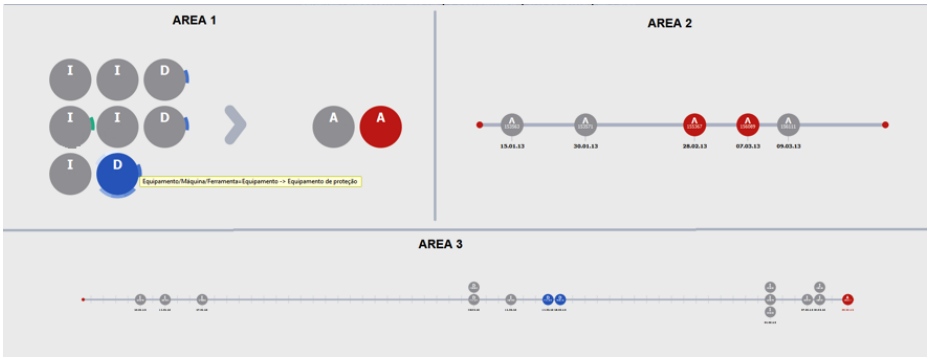

Fig. 6. Inter-anomaly interactive visualization of *n-to n* rules

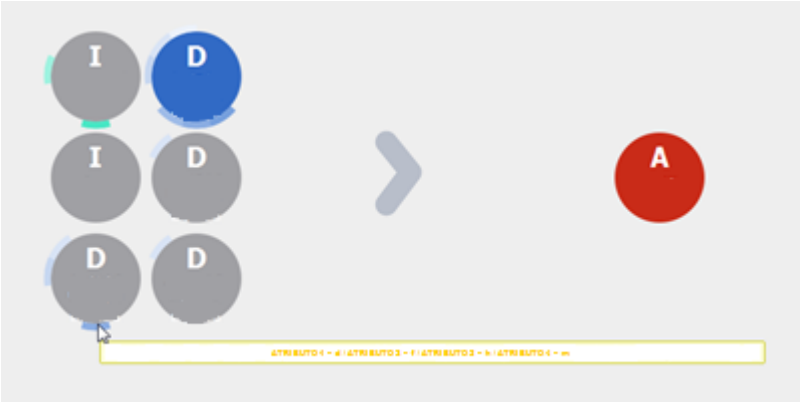

Fig. 7. Visualization of anomalies shared attributes

4 Conclusions

In this work, we introduced a novel interactive visualization of association-mined rules specially devised to deal with massive problems. We have applied this interactive visualization to mined rules as part of Occupational Health and Security (OHS) in offshore oil extraction and processing plant. Although we have introduced and presented the interactive visualization of association rules problem itself, it must be pointed out that, this approach is currently deployed as part of a larger system that rely of the mining and classification modules. In addition, a set of usability tests will complement this study. The global system is currently in use by a major petroleum industry conglomerate of Brazil and is to be presented as a whole in a forthcoming paper. Readers must be warned that the results presented here had to be transformed in order to preserve the sensitive details of the data. Further work in this direction is called for and is currently being carried out. An important direction is the formal understanding of the user-centric evaluation of the proposal.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases. VLDB '94, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1994) 487-499.
2. Borrajo, M.L., Baruque, B., Corchado, E., Bajo, J., Corchado, J.M.: Hybrid neural intelligent system to predict business failure in small-to-medium-size enterprises. International Journal of Neural Systems 21(04) (2011) 277-296.
3. Bruzzese, D., Buono, P.: Combining visual techniques for association rules exploration. In: Proceedings of the working conference on Advanced visual interfaces, ACM (2004) 381-384.
4. Bruzzese, D., Davino, C.: Visual post-analysis of association rules. Journal of Visual Languages & Computing 14(6) (2003) 621-635
5. Bruzzese, D., Davino, C.: Visual mining of association rules. In Simo, S., Bhlen, M., Mazeika, A., eds.: Visual Data Mining. Volume 4404 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2008) 103-122
6. Conti, M., Pietro, R.D.,Mancini, L.V., Mei, A.: Distributed data source verification in wireless sensor networks. Information Fusion 10(4) (2009) 342-353.
7. De Paz, J.F., Bajo, J., Lopez, V.F., Corchado, J.M.: Biomedic organizations: An intelligent dynamic architecture for KDD. Information Sciences 224 (2013) 49-61.
8. Hahsler, M., Chelluboina, S.: Visualizing association rules: Introduction to the r-extension package arulesviz. R project module (2011)
9. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: ACM SIGMOD Record. Volume 29., ACM (2000) 1-12
10. Hofmann, H., Siebes, A. P. J. M and Wilhelm, A. F. X. Visualizing association rules with interactive mosaic plots. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '00, pages 227-235. New York, NY, USA, 2000. ACM.
11. Hofmann, H., Wilhelm, A.: Visual comparison of association rules. Computational Statistics 16(3) (2001) 399-415
12. Inselberg, A.: N-dimensional graphics, Part I - lines and hyperplanes. Harwood g. kolsky papers edn. International Business Machines Corporation (IBM). Los Angeles Scientific Center (1981)
13. Sekhavat, Y.A., Hoeber, O.: Visualizing association rules using linked matrix, graph, and detail views. International Journal of Intelligence Science 3 (2013) 34
14. Zaki, M.J.: Scalable algorithms for association mining. IEEE Transactions on Knowledge and Data Engineering 12(3) (2000) 372-390.