DOI: 10.15439/2023F4831
ISSN 2300-5963 ACSIS, Vol. 35

Association Rule Mining for Requirement Elicitation Techniques in IT Projects

Denys Gobov
0000-0001-9964-0339
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, 37, Prosp. Peremohy, Kyiv, Ukraine
Email: d.gobov@kpi.ua

Nikolay Sokolovskiy
0009-0000-1282-5665
Independent researcher
Email: sokolovskynik@gmail.com

***Abstract*—Selecting suitable techniques for requirements elicitation in IT projects is crucial to the business analysis planning process. Typically, the determining factors are the preferences of stakeholders, primarily business analysts, previous experience, and company practices, as well as the availability of sources of information. The influence of other factors is not as evident. One of the possible ways to form recommendations for using techniques is the analysis of industrial experience. This paper is intended to analyze the application of association rules mining to define factors influencing technique selection and predict the usage of a particular elicitation technique depending on the project context and specialist background. The dataset for experiments was formed based on a survey of 328 specialists from Ukrainian IT companies. The associations found to make it possible to speed up the process of choosing elicitation techniques and improve the elicitation process efficiency.**

***Index Terms*—associations rules mining, requirements elicitation, IT project, business analysis.**

I. Introduction

REQUIREMENTS elicitation is the effort expended by the Requirements Engineer to turn implicit desires, demands, wishes, needs, and expectations — which until now were hidden in their sources — into explicit, understandable, recognizable, and verifiable requirements [1]. The outputs of elicitation serve as input for the following tasks from the core business analysis cycle: current state analysis, risk assessment, and requirement specification and modeling [2]. Elicitation activities can be divided into three tasks: preparing, conducting, and result confirming. The effectiveness of elicitation directly depends on the quality of the first – preparation. The requirement engineer/business analyst should define the available source of information, a subset of stakeholders, who should be involved in the following elicitation activities and select appropriate elicitation techniques. Professional guides and standards recommend many techniques practitioners use in IT projects. Due to time and budget constraints, specialists can't use them all and should select a set of techniques best suited to the particular project's conditions. The set of predefined elicitation techniques significantly influences the business analysis, project plan, and the associated costs and resources needed. This study was conducted to analyze the current practices of using elicitation techniques in IT projects and to find associations between project context, specialist's profile, and techniques used for requirement elicitation via Association Rule Mining. The dataset for analysis was gathered via a survey of 328 IT specialists employed by Ukrainian and international companies with branches in Ukraine via a survey [3]. The strong associations identified with Association Rule Mining made it possible to formulate recommendations on using requirements elicitation techniques in IT projects.

II. Problem Statement

The task of selecting best-suited techniques, particularly requirements elicitation techniques, is performed by a business analyst at the start of the project due to defining and estimating a list of business analysis-related activities. But that does not mean it is a one-time task, and a list of used techniques can be updated based on the efficiency monitoring results and project context changes. Considering that the requirements elicitation lays the foundation for further analysis and development activities, the optimal technique selection is an essential business analysis task. The emergence of new techniques and their development in the process of business analysis evolution, as well as the continuously changing business environment, can lead to the complication of this task. A recommendation system that considers the accumulated experience of practicing business analysts and requirements engineers can be applied to solve this problem. An important condition is the explainability of these recommendations, which will allow for checking their applicability in the unique context of each project.

III. The Best Existing Solution

There are many studies regarding solving the choosing appropriate requirement elicitation technique problem using different approaches and models.

Hatim Dafaalla et al. [4] built a model based on an artificial neuronal network (ANN). The model was learned based on the collected dataset with 1684 records about selecting the elicitation technique. By choosing the ROC AUC metric as a score of the model, the authors achieved significant accuracy of the model, which was equal to 82%. Despite good forecasting by modeling, as with any other ANN, this model has a significant weakness. ANN is a net of perceptron (miniature models of neurons). The perceptron is organized in layers, which are connected to each other. The connections might have a different architecture. Each connection of each perceptron has a weight coefficient. The learning process is a process to optimize these coefficients. Unfortunately, a single coefficient and a set of coefficients don't have meaning and can't be explained in business terms. Similarly, connections, layers, and perceptions do not have any sense separately and don't explain how ANN solved a problem. That is the way some decisions of an ANN might be seen as strange, unexplained, and untrusted [5].

Nagy Ramadan Darwish et al. [6] suggested a hybrid approach. The manuscript describes a pipeline of methods. The feature is manually selected based on literature reviews. Then multiple linear regression model was built to select critical attributes influencing technique selection. In the last stage, the ANN was built. The accuracy of the final model was declared as 81%. Despite the remarkable result, the final model has the same limitations as discussed previously. Ihor Bodnarchuk et al. [7] applied goal function for assessment and selection architecture design in the context of "light-weighted" requirements techniques.

Different machine learning approaches were applied not only to technique selection but to related areas as well. Fadhl Hujainah and others [8] suggested using a semi-automated attribute measurement criteria method for requirement prioritization and selection.

Similar method - attributes-based decision making was described by Jinyu Li [9]. Remarkably, semi-automated methods bring a possibility of bias since experts conducted the first assessment.

IV. The Proposed Solution

Associate Rule Mining (ARM) method is a machine-learning technique that combines several remarkable advantages. Firstly it doesn't require data annotation because it is an unsupervised method. Secondly, the method and output are intuitive and could be understood by domain experts and business people, which is a rare property of a machine-learning algorithm.

ARM, also known as basket analysis, was applied first in retail, but now it is widely applied in other areas. For example, Giovanna Castro and colleagues in [10] applied association rules to study the comorbidity of bipolar disorder and premenstrual dysphoric disorder. Chad Creighton [11] used association rules to discover hidden gene expression patterns. Ahmad Mirabadi and Shabnam Sharifian [12] applied the ARM to Iranian Railways data to discover patterns leading to incidents and create management manuals and guidelines. Finally, the method could detect credit card fraud [13]. The Association rules are even included in other algorithms, such as Lamma and other [14] embedded AR, as part of the SLA algorithm.

V.Conditions of the Analysis to Follow

Considered methods are applied to the particular dataset for extraction association rules. It means that if the initial dataset is biased, the found association rules will also have bias.

Moreover, as you will see in the following sections, ARM requires settled initial (apriori) hyperparameters that influence the number of found rules. According to mentioned studies above, there is no standard practice to calculate the metrics, and usually, it comes from the business perspective and domain expert knowledge. During the study, we considered various combinations of rules to find a balance between the number of rules and the reasonability in order to find the most appropriate set of rules.

During the study, we worked with two hyperparameters: support and confidence (see definitions in the next section). We began with a support level of 0.5, increasing by 0.1 while reaching 1.0. We chose 0.5, which means a rule is true for 50% of cases. We obtained an itemset with confidence levels from 0.1 to 1.0 with increments of 0.1 for each new support. Each obtained dataset was estimated among the following questions:

- How many association rules are found out?
- Does an entirely differential rule in the top 100 rules disappear compared with the previous values of hyperparameters?

We stopped the process when we obtained a set with completely differential rules at the top of the list.

VI. Details of the Proposed Solution

A. Association Rule Mining

The problem of discovering association rules was proposed by Agrawal et al. [15]. Let $I = \{i_1, i_2, \dots, i_m\}$ be a set of m items. Let T be a set of transaction $\{t_1, t_2, \dots, t_n\}$, where each t_i is set of items in which $t_i \subseteq I$. Association rules are implication rules:

$$A \Rightarrow B,$$

which is interpreted as "if A, then B". The following statements must be met: $A \subset I, B \subset I$ and $A \cap B = \emptyset$. The A term is an antecedent of the rule. The B term is a consequent of the rule.

The number of rules might be huge, so we need some mechanism for selecting strong rules from weak ones. To do that, let's postulate the following hyperparameters:

- Confidence is a measure that counts how many transactions in T that contain A also contain B. It is the probability of B being true when we already know that A is true:

$$Confidence(A \Rightarrow B) = \frac{Occurence\ of\ A\ and\ B}{Occurences\ of\ A}$$

- Support is a measure of the frequency of the transaction patterns that occur in the T:

$$Support(A \Rightarrow B) = \frac{Occurence\ of\ A\ and\ B}{Total\ transaction\ in\ T}$$

- Lift is a value that gives us information about the increase in the probability of the "then" (consequent) given the "if" (antecedent) part. If the lift equals one, we consider there are no dependencies, but if the lift is more than one, we can consider a dependency. Additionally, the lift can demonstrate the "power" of dependency: the larger the lift, the stronger the rule.

$$Lift(A \Rightarrow B) = \frac{Support(A \Rightarrow B)}{Support(A) * Support(B)}$$

Now we can define the minimal support and confidence values to select strong rules. The rules which have confidence more than the selected minimal value are called strong rules.

B. Apriori Algorithm

The Apriori, proposed by Agrawal et al. in [15], is an algorithm for discovering association rules. The algorithm is based on searching frequent itemsets. It assumes that if rule X has a confidence level of C and $X \subset Y$, so rule Y has a confidence level not less than X. In this way, we can dramatically reduce calculations by excluding many weak rules from consideration based on the frequency of every single i in I.

VII. Analysis

A. Input data

To discover association rules, we used the survey result conducted in 2020 [16]. After data cleaning, the dataset has 324 answers, which will be treated as a transaction. To describe a project context, we asked respondents about the following:

- project size;
- project domain;
- company type (IT-outstaff, IT-outsource, IT product, non-IT);
- company size;
- class of the developed system (business software, embedded software, scientific, etc.);
- belonging to the co-located or distributed team;
- role in the project;
- years of experience;
- passing certification in the chosen role;
- using adaptive, hybrid, or predictive ways of working on the project;
- project category (developing from scratch, reengineering, product or platform customization, etc.);
- involving in different Types of BA activities.

The dataset is available at the link https://data.mendeley.com/datasets/svzv7rs279.

Together the answer's options produced 96 possible items in the itemset.

Before running the apriori algorithm, we discovered the support (frequency) of single items of elicitation techniques. We decided not to consider items (and consequently rules) with a frequency less than 50% (Table 1).

The apriori algorithm was launched across the dataset with the following hyperparameters: minimal support 0.5 and minimal confidence 0.8. After removing autogenerated rules with empty antecedents, there were left 86 association rules.

Table I.
Elicitation techniques with a frequency of more than 50%

Elicitation Technique	Support level %
Interviews	87.3
Document analysis	85.5
Interface analysis	71.3
Brainstorming	69.2
Process analysis/modeling	66.1
Prototyping	66.1
Business rules analysis	54.4

The first look at consequent showed that only two techniques have strong antecedents: Document analysis and Interviews. It means that despite the frequency of other consequents, there is not a strong enough implication between any project context aspects under interest and the consequent itself. Perhaps, the choice of rest elicitation techniques is managed by factors that lay off the considered dataset.

Remarkable that both mentioned methods are often used in pairs. Rule "Document analysis → Interviews" has one of the biggest (0.77) support levels and similar "Interviews → Document analysis". This fact makes sense: a business analyst uses different sources of information due to business analysis information elicitation. Usually, documents and people are the most valuable and accessible sources.

B. Document analysis association rules

First, some rules state implications based on other elicitation methods presented in Table 2.

Table II.
Document analysis association rules

Association Rule	Support level %
Interface analysis → Document analysis	0.65
Process analysis & modeling → Document analysis	0.6
Brainstorming → Document analysis	0.6
Prototyping → Document analysis	0.58

Also, a small subset of rules combines different elicitation methods and another aspect of the project context. For example (here and further, the number in parentheses is a support level): (Business software, Interviews) → Document analysis (0.69), (Interviews, BA Role) → Document analysis (0.69), (Business software, Interviews, BA Role) → Document analysis (0.62), (Interface analysis, Role: BA) → Document analysis (0.58), (Interface analysis, Interviews) → Document analysis (0.58). But these rules have support levels smaller than in rules without other components.

Consider other strongest association rules in this group. Remarkable that BA's role in the project implicates using Document analysis: BA Role → Document analysis (0.76). And the rule includes the class of the system under interest: Business software → Document analysis also has a high (0.74) support level. The situation with mixed rules for role and class system is the same as for mixed rules of elicitation techniques: they have more minor support levels and confidence than the short version. For example, Business software, Role: BA → Document analysis (0.67), Role: BA, Requirements analysis and design definition 0.58

Behind the discovered rules, one more group influences the choice of elicitation techniques. The rule with the strongest support level is (Requirements analysis and design definition, Elicitation & Collaboration) → Document analysis (0.57)

C. *Interviews association rules*

The Interview's association rules are presented in table 3.

TABLE III.
INTERVIEW ASSOCIATION RULES

Association Rule	**Support level %**
Business software → Interviews	0.77
BA Role → Interviews	0.76
Elicitation & Collaboration → Interviews	0.63
Interface analysis → Interviews	0.63
Brainstorming → Interviews	0.62
Process analysis & modeling → Interviews	0.60
Team distributed → Interviews	0.55

As well as for the previous group, there are many more complex rules with three and more antecedents. However, the support level of these rules is less than the listed above, while their confidence level stays the same. Several examples illustrate the thesis: (Business software, BA Role, Document analysis) → Interviews (0.62), (Requirements analysis and design definition, Elicitation & Collaboration) → Interviews (0.57), (BA Role, Requirements analysis and design definition) → Interviews (0.57), (Business software, Requirements analysis and design definition, Elicitation & Collaboration) → Interviews (0.51), (Business software, Document analysis, Process analysis & modeling) → Interviews (0.5)

That could mean that a significant and essential implication in choosing the elicitation technique is laid out in less complex rules. Remarkable that here we can observe rules that postulate implications based on another elicitation technique, such as Interface analysis and Brainstorming.

VIII. CONCLUSION

We analyzed datasets obtained from the survey. The dataset includes 324 transactions containing items from itemset with 96 items. The apriori algorithm was used for discovering association rules. The algorithm's hyperparameters were defined as minimal support equals 0.5 and minimal confidence equals 0.8. We considered only rules with left bigger than 1. The algorithm discovered 86 associated rules.

The most frequently used elicitation techniques are Interviews, Document analysis, Brainstorming, Process analysis and modeling, Prototyping, and Business rules analysis.

The main discovering facts and rules are:

- Among all frequent rules, only two techniques - Document analysis and Interviews- form strong association rules with project context.
- Interviews and Document analysis are used together pretty often.
- Class of developing system (business software) and BA role and BA activity make using Document Analysis elicitation technique.
- Class of developing system (business software) and BA role, distributed team, Process analysis & modeling, and BA activity such as Elicitation and Collaboration and make using Interview technique.
- Some elicitation techniques (Brainstorming, Interface analysis, Process analysis & modeling) implicate using Interview technique.
- The combination class of developing system, role in the project, team distribution, and activity with other aspects of project context have more minor support levels than less complex rules having only one antecedent and could be considered a sub-option.

The following recommendations can be proposed based on found association rules:

- If a person who performs requirements elicitation uses only Document Analysis or only Interview, they might consider Interview or Document Analysis accordingly.
- If a business analyst uses Interface analysis, Process analysis & modeling, Brainstorming, or Prototyping, they might consider Document Analysis as an additional technique;
- If the system under development is business software, then Document analysis and Interview are reasonably chosen;
- If Interface analysis, Process analysis & modeling, or Brainstorming are used, Interview should be considered as an additional technique;
- Interview is a suitable technique in case of a distributed team.

REFERENCES

[1] K. Pohl, "Requirements engineering: fundamentals, principles, and techniques", Springer, New York, USA, 2010, 182 p.

[2] D. Gobov, V. Yanchuk, "Network Analysis Application to Analyze the Activities and Artifacts in the Core Business Analysis Cycle," *2021 2nd International Informatics and Software Engineering Conference (IISEC)*, Ankara, Turkey, 2021, pp. 1-6, doi: 10.1109/IISEC54230.2021.9672373.

[3] D. Gobov, "Practical Study on Software Requirements Specification and Modelling Techniques". International Journal of Computing, 22(1), pp. 78-86, 2023. https://doi.org/10.47839/ijc.22.1.2882.

[4] H. Dafaalla, et al., "Deep Learning Model for Selecting Suitable Requirements Elicitation Techniques, Applied Science, vol. 12 (18), pp. 9060, 2022. https://doi.org/10.3390/app12189060

[5] V. Sharma, S. Rai, A Dev, "A comprehensive study of artificial neural networks." International Journal of Advanced research in computer science and software engineering, vol 2, no. 10, pp. 278-284, 2012

[6] N Darwish, A. Mohamed, A. Abdelghany, "A hybrid machine learning model for selecting suitable requirements elicitation techniques", International Journal of Computer Science and Information Security, vol. 14, no. 6, pp. 1-12, 2016.

[7] I. Bodnarchuk, et al., "Adaptive Method for Assessment and Selection of Software Architecture in Flexible Techniques of Design", IEEE, 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), pp. 292-297, 2018. https://doi.org/10.1109/stc-csit.2018.8526620

[8] F. Hujainah, R. B. A. Bakar, M. A. Abdulgabber, "StakeQP: A semi-automated stakeholder quantification and prioritization technique for requirement selection in software system projects", Decision Support Systems, vol. 121, pp. 94-108, 2019. https://doi.org/10.1016/j.dss.2019.04.009

[9] J. Li, et al., "Attributes-based decision making for selection of requirement elicitation techniques using the analytic network process", Mathematical Problems in Engineering, vol. 2020, pp. 1-13, 2020. https://doi.org/10.1155/2020/2156023

[10] G. Castro, et al., "Applying Association Rules to Study Bipolar Disorder and Premenstrual Dysphoric Disorder Comorbidity," 2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE), Quebec, QC, Canada, 2018, pp. 1-4. https://doi.org/10.1109/ccece.2018.8447747

[11] C. Creighton, S. Hanash, "Mining gene expression databases for association rules", Bioinformatics, vol. 19., no. 1, pp. 79-86, 2003. https://doi.org/10.1093/bioinformatics/19.1.79

[12] A. Mirabad, S. Sharifian, "Application of association rules in Iranian Railways (RAI) accident data analysis", Safety Science, vol. 48, no. 10, pp. 1427-1435, 2010. https://doi.org/10.1016/j.ssci.2010.06.006

[13] D. Sánchez, et al., "Association rules applied to credit card fraud detection", Expert systems with applications, vol. 36, no. 2, pp. 3630-3640, 2009. https://doi.org/10.1016/j.eswa.2008.02.001

[14] E. Lamma, et al., "Improving the SLA algorithm using association rules", Springer Berlin Heidelberg, AI* IA 2003: Advances in Artificial Intelligence: 8th Congress of the Italian Association for Artificial Intelligence, Pisa, Italy, September 2003. Proceedings 8, pp. 165-175, 2003. https://doi.org/10.1007/978-3-540-39853-0_14

[15] R. Agrawal, et al., "Fast algorithms for mining association rules", Proceeding 20th international conference very large data bases, VLDB, vol. 1215., pp. 487-499, 1994.

[16] D. Gobov, I. Huchenko, “Influence of the software development project context on the requirements elicitation techniques selection”, In: Hu, Z., Petoukhov, S., Dychka, I., He, M. (eds) Advances in Computer Science for Engineering and Education IV. ICCSEEA 2021. Lecture Notes on Data Engineering and Communications Technologies, vol 83. Springer, Cham. https://doi.org/10.1007/978-3-030-80472-5_18.