Association Rule Mining for Improvement of IT Project Management

Snezhana Sulova [1]

[1] *University of Economics – Varna, Bulgaria*

***Abstract* – In this research we extract knowledge from human resources data, accumulated in IT companies for the right selection of teams to work on software projects. We are looking for interesting and unknown dependencies and connections in the data, based on which managers can form more cohesive and professionally working project teams. The proposed approach to improve the selection of teams working on IT projects is based on association rule mining and can be used by IT managers to select the members of the teams. The approbation of the proposed approach is made using the software product RapidMiner.**

***Keywords* – IT project, Project team, Association Rule Mining, Apriori, FP-Growth.**

1. Introduction

Nowadays information technologies have become a major business function in almost every organization. Most companies have big expectations for the business advantages from investments in IT. Software companies are mainly working on various web design projects and other software applications assigned to them. It should be noted that IT projects have their specificities and are a huge challenge [1]. The result from them is a developed software product which depends on the knowledge and the creative capabilities of the developers, and their teamwork skills. That's why the main factor for the success of an IT project is the performance of the staff [2]. The dynamics in software business and the need to achieve quality results, call for new and innovative approaches based on modern scientific knowledge to be used when managing IT projects and setting up project teams.

2. Formulation of the problem

The projects can be divided into a series of tasks which are the following: have a specific objective to be completed within certain specifications; have defined start and end dates; have funding limits; consume human and nonhuman resources; are multifunctional [3]. When planning IT project activities, the manager is usually the one responsible for the team's staff [4]. The project manager needs to apply knowledge, skills, instruments and technologies to the project activities to achieve the project objectives and requirements.

In modern IT companies, team leaders use a software for project planning, allocation of resources, and project management, these are the so-called project management systems. Although these systems have a wide functionality and support project work in all phases, the project management software does not have enough resources to give advice when putting together a project team. It is well known that computer-aided evaluation tools can give an objective view of employees and help implement the project by shortening the execution deadlines or by improving the quality of the software [4]. In this regard, the main goal of the article is to propose a new innovative approach based on the association rule mining technology to be used by IT project managers to form successful project teams.

3. Literature review

In recent years, more and more researchers, who are looking for ways to gain new knowledge based on the vast amount of accumulated data, use data mining (DM) technologies. DM "is the study of collecting, cleaning, processing, analyzing, and

DOI: 10.18421/TEM74-03
https://dx.doi.org/10.18421/TEM74-03

Corresponding author: Snezhana Sulova,
University of Economics – Varna, Bulgaria
Email: ssulova@ue-varna.bg

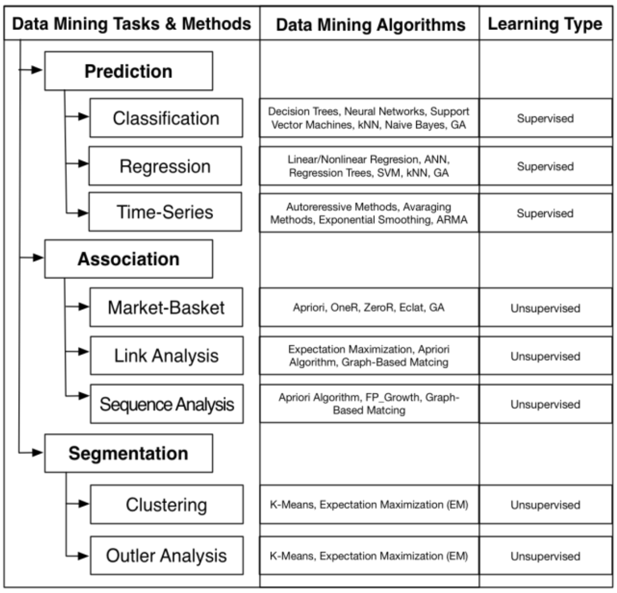
Received: 18 September 2018.
Accepted: 03 November 2018.
Published: 26 November 2018.

© 2018 Snezhana Sulova; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 License.

The article is published with Open Access at www.temjournal.com

gaining useful insights from data" [5]. The term is mainly used to describe the process of exploring and discovering hidden data and knowledge that was previously not known but which is useful to business by using techniques and algorithms in the field of artificial intelligence [6, 7].

The basis of modern Data Mining technology is the concept of templates or models reflecting the multi-dimensional interrelation between the data. These models represent a collection of regularities and a selection of data by given properties, which are appropriately presented in forms easily available to users [8]. Different methods are used to create them. Many researchers are working on various method-related problems, Data Mining algorithms and their application [9, 10]. A good summary of the main theoretical knowledge for data mining tasks and methods is made by Delen [11] and is presented in Figure 1.



Data Mining Tasks & Methods	Data Mining Algorithms	Learning Type
Prediction		
Classification	Decision Trees, Neural Networks, Support Vector Machines, kNN, Naive Bayes, GA	Supervised
Regression	Linear/Nonlinear Regresion, ANN, Regression Trees, SVM, kNN, GA	Supervised
Time-Series	Autoreressive Methods, Avaraging Methods, Exponential Smoothing, ARMA	Supervised
Association		
Market-Basket	Apriori, OneR, ZeroR, Eclat, GA	Unsupervised
Link Analysis	Expectation Maximization, Apriori Algorithm, Graph-Based Matcing	Unsupervised
Sequence Analysis	Apriori Algorithm, FP_Growth, Graph-Based Matcing	Unsupervised
Segmentation		
Clustering	K-Means, Expectation Maximization (EM)	Unsupervised
Outler Analysis	K-Means, Expectation Maximization (EM)	Unsupervised

Figure 1. Data Mining Tasks/Methods [10, p. 104]

It should be noted that the known basic algorithms used to solve various tasks related to classification, clustering, search of association rules, are constantly improving. Researchers offer algorithmic improvements which lead to greater data processing performance and better results.

Associations known as association rule mining are one of the known methods of Data Mining [12, 13]. They are used to obtain regularities from related elements and events and to discover links between elements in large data sets. The dependencies found are presented as rules and can be used for analyzes and forecasts.

For the first time, these rules are used in market-basket analyzes to determine the set of goods that are often purchased together in large data sets [14]. The association rules are defined as follows. For example, if we have a set of elements I = {i1, i2 ... in} and D – multiple transactions D = {t1, t2, ... tn}. Each transaction contains a subset of the elements in I – T⊑I. The association rule is the implication X → Y, where X⊑I, Y⊑I and X∩Y = ∅.

To evaluate the usefulness of the found rules, the following constraints are entered:

- Support – shows what percent of the data contains this rule, i.e. how often a rule occurs in a given set;

$$Support(X \rightarrow Y) = \frac{Number\ or\ records\ with\ X\ and\ Y}{Number\ of\ records\ in\ itemsets}$$

- Confidence – shows the probability of the inclusion of the item X in the transaction also leading to the inclusion of an item Y;

$$Confidence\ (X \rightarrow Y) = \frac{Number\ of\ records\ with\ X\ and\ Y}{Number\ of\ records\ with\ X}$$

- Lift – measure of the performance of association rule;

$$Lift\ (X) = \frac{Confidence(X \rightarrow Y)}{Support(X \rightarrow Y)}$$

The higher the value of the confidence measure is, the stronger the association rule is, and the high value of the lift measure shows that having X and Y together is not accidental but is due to a link between them.

In research, most of the association rules extraction algorithms are ranked according to breadth-first search BFS and depth-first search DFS [15]. The basic known BFS algorithm is Apriori. It generates common subsets of data. The sets expand one by one and the process ends when no further extensions are detected. This algorithm has many variations and enhancements, such as AprioriTID, DIC, in which the number of candidate sets is reduced, and thus productivity is improved.

The most common DFS algorithm is FP-Growth. It extracts association rules by scanning all records one by one and records them in a tree structure (Fp-tree). Another known algorithm is Eclat (Equivalence CLAss Transformation), which searches data in a vertical format [16].

Various applications of association rule mining are mentioned in literature. In addition to the previously mentioned application in market-basket analyzes, the association rules are applied to training systems [17], risk assessment in banking operations [18], medical diagnostics [19], crime prevention [20] and other cases where data links are needed. The goal of this research is to use the data for the employees of the IT companies and find the association rules which can be used in the formation of the project teams.

4. A new approach for Improvement of IT Project Management, based on association rule mining

To analyze the employee data and assist the IT manager's activities, we propose the following approach for selecting a well-functioning project team (Figure 2.):

1. Analysis of the environment and collection of necessary data.
2. Selection of attributes for analysis.
3. Application of different algorithms for association rule mining.
4. Comparing and summarizing the results, analyzing and deciding the staff of the project team.

4.1. Analysis of the environment and collection of necessary data

The first step in the search for association rules and implementing any data mining technique, according to the leading methodology DM CRISP-DM, is business and data understanding [21]. Since our research is looking for dependencies and interesting links between IT staff data, it is advisable at this stage that project team managers work together with database specialists to assess which of the available data can be used and whether it is necessary to supplement the data through a survey to collect additional employee data. The usage of data from different applications as well as additional data from studies suggests that the data are in different formats and need to be transformed, processed in order to form the required data set to be subjected to DM processing. It should be noted that the quality of the results of applying data mining methods directly depends on the quality of data collection and organization.

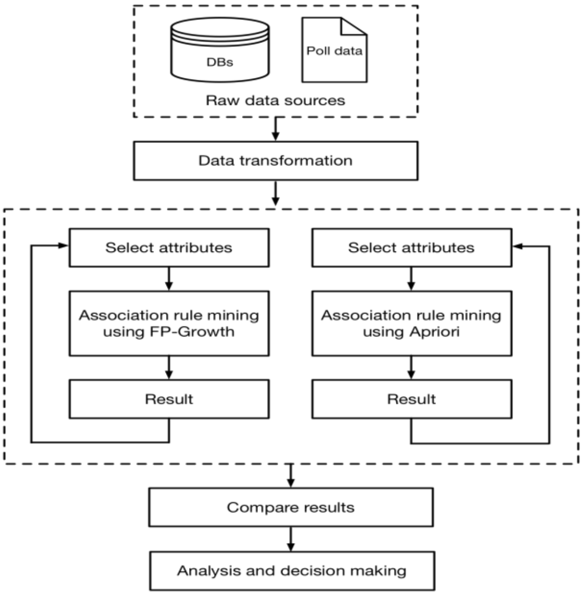


Figure 2. An approach for the search of association rules

4.2. Selection of attributes for analysis

The second main step offered in the approach which assists the employee selection process is the choice of specific attributes for analysis. This stage is largely dependent on what type of project the team is being formed for.

For example, if the team must consist mainly of PHP and MySQL programmers, then one of the attributes will be knowledge about these technologies, and interesting dependencies of these employees can be found, such as whether they can work on any other technologies or belong to some professional community or whether they have a certain quality. If a team is required to quickly complete a specific IT project, then it is appropriate to select attributes, such as professional experience, way of working, etc.

Using the suggested approach, the selection of attributes for analysis can be repeated many times and interesting and useful dependencies between different data samples can be looked for.

4.3. Application of different algorithms for Association rule mining

The third stage – applying different algorithms for Association rule mining involves searching for dependencies with two of the most commonly used algorithms, Apriori and FP-Growth. Essential for this step is to set appropriate values for the support and confidence parameters. As mentioned, the algorithms for the revelation of association rules find rules such as “From X follows Y” (X → Y) with different values of support and confidence. In most cases, it is necessary to limit the number of found rules to the minimum and maximum values of support and confidence. If the rule's support value is too high, obvious and already known rules will be found after the implementation of the algorithm. Too low support value will result in a large number of rules that will not be known and obvious but will be very unreasonable. If the confidence value is low, the rules found may not be valuable for the research.

4.4. Comparison of the results, analysis and deciding the creation of project teams

After applying the two algorithms for finding association rules, we suggest comparing and summarizing the results. The matching dependencies are defined as important, and if there are dependencies of only one algorithm they are identified as less relevant.

In the final analysis and decision-making process, the role of an IT team manager is of great importance. To form a better-working team, the manager can use some of the dependencies found when selecting employees.

5. Approbation and results

We have selected the RapidMiner software product to evaluate and test the proposed approach. Nowadays, this is one of the most widely used software solutions for data mining and predictive analysis worldwide [22].

The test data we have is stored in a two-dimensional table that contains 998 records. Some of the fields in the table are extracted from the employee database, others have been supplemented by an employee of the HR department after a survey. The table contains the following fields:

- ID
- Employment period
- Time in current department
- Gender
- Team manager
- Age
- Family
- Member of professional organizations
- .Net
- SQL Server
- HTML CSS Java Script
- PHP mySQL
- Fast working
- Awards
- Communicative

Since there are missing values in the table for some of them, such as those showing the knowledge, skills and attributes of employees, it is useful to replace the missing values with zero. This means that for example, if we do not know whether an employee is communicative, we will assume that they are not and thus by looking for a communicative person for a project, the connections obtained will be more reliable. We've replaced the selected fields with the RapidMiner operator – Replace Missing Values.

The next step in the approbation, according to the approach presented in point 4, is to compile the association rules extraction models. Figure 3. shows the dependency search model using the FP-Growth algorithm where frequent item sets are defined first and after that association rules are generated by the Create Association rule.

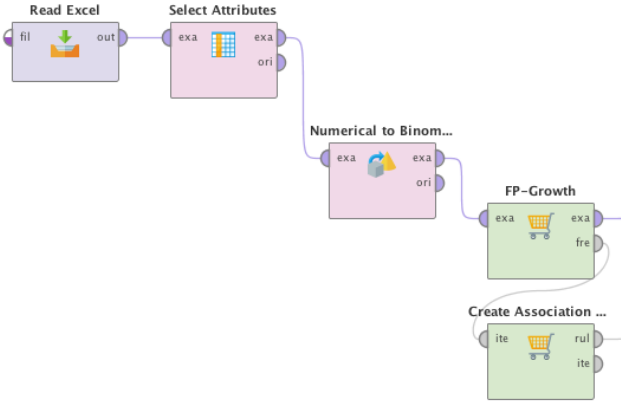


Figure 3. Process of generating association rules using the FP-Growth

Figure 4. shows the Apriori algorithm search pattern, which generates association rules based on frequent sets of data.

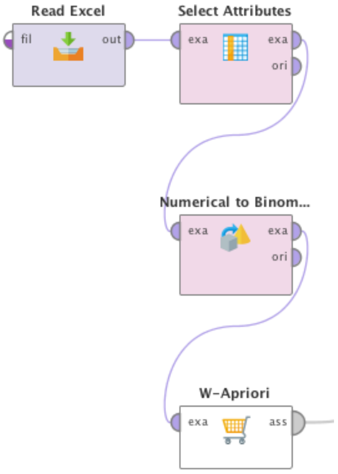


Figure 4. Process of generating association rules using Apriori

In both compiled models there is the Select Attributes operator, which specifies the attributes that will be involved in the association rules search process. For testing purposes, the following attributes are chosen:

- HTML CSS Java Script
- PHP mySQL
- Fast working
- Member of professional organizations
- Awards
- Communicative

Testing the process with the two algorithms shows that using the same values for Minimum support and Minimum confidence, more item sets are detected with the FP-Growth algorithm.

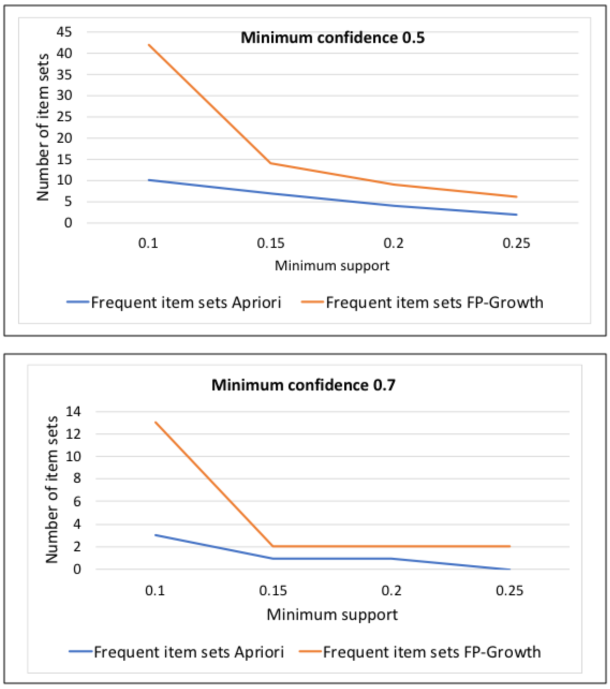


Figure 5. Comparison of the performance of the two processes

To analyze the obtained results, it is also necessary to consider the specific rules found at different values for the Minimum support and Minimum confidence measures. For example, with Minimum support 0.2 and Minimum confidence 0.7, the rules found by the two models are as follows:

- using FP-Growth:
  Member of professional organizations → Communicative
  HTML CSS Java Script, Member of professional organizations → Comunicative
  HPhp mySQL, Awards → HTML CSS Java Script
  Communicative, Awards → Member of professional organizations
  PHP mySQL, Member of professional organizations → Communicative
  Member of professional organizations, Fast working → Comunicative
  Fast working, Awards → HTML CSS Java Script
- using Apriori:
  Awards → HTML CSS Java Script
  PHP mySQL → HTML CSS Java Script
  TML CSS Java Script → PHP mySQL

By looking at and comparing rules, it can be concluded that the rules found with the FP-Growth algorithm are more and show a larger amount of interesting and meaningful connections.

To deepen the analysis or study the dependencies with a precisely defined characteristic of the people, it is possible to bring out all the connections for the people who have a communicative characteristic and to study only those attributes (Figure 6.).

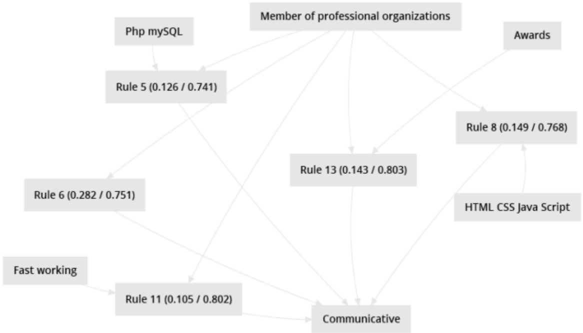


Figure 6. Graph view for the Communicative associations

The analysis process should be repeated with other selected attributes and with different values for support and confidence.

6. Conclusion

In conclusion we should emphasize that the optimal and full allocation of human resources is important for the end results of IT firms. The competitive advantage of software developers would be to work in well-synched teams that work together as a whole to achieve corporate goals. In this regard, we believe that the discovery of unknown links between employees can help create better project teams.

The proposed approach and the experimental results obtained prove that association rule mining can generate useful facts or associations between the data and based on them project managers can make important decisions.

References

[1]. Marinova, O. (2015). New Principles in IT Project Management Through Agile Methodologies. *Izvestia, Journal of the Union of Scientists-Varna, Economic Sciences Series*, (1), 117-124.

[2]. Lientz, B. P., & Larssen, L. (2004). *Manage IT as a business: How to achieve alignment and add value to the company*. Routledge. Routledge.

[3]. Kerzner, H., & Kerzner, H. R. (2017). *Project management: a systems approach to planning, scheduling, and controlling*. John Wiley & Sons.

[4]. Nestorov, K. (2011). Managing modern IT projects in business. Varna: University of Economics, (in Bulgarian).

[5]. Aggarwal, C. (2015). Data mining. The Textbook, New York: Springer.

[6]. Barsegyan, A. et. al. (2009). Analysis of data and processes. 3rd ed., Sankt-Peterburg: BKhV.

[7]. Todoranova, L. (2015). The creation and development of knowledge warehouses. *Izvestia, Journal of the Union of Scientists-Varna, Economic Sciences Series*, (1), 156-161.

[8]. Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, *21*(1), 1-24., doi:10.1017/S0269888906000737

[9]. Larose, D. T. (2006). *Data mining methods & models*. John Wiley & Sons.

[10]. Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.

[11]. Delen, D. (2014). *Real-world data mining: applied business analytics and decision making*. FT Press.

[12]. Larose, D. (2005). Discovering knowledge in data. An Introduction to Data Mining, New York: John Wiley & Sons Inc.

[13]. Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

[14]. Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *Acm sigmod record* (Vol. 22, No. 2, pp. 207-216). ACM. doi:10.1145/170035.170072.

[15]. Hipp, J., Güntzer, U., & Nakhaeizadeh, G. (2000). Algorithms for association rule mining—a general survey and comparison. *ACM sigkdd explorations newsletter*, *2*(1), 58-64.

[16]. Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE transactions on knowledge and data engineering*, *12*(3), 372-390.

[17]. Alzoubaidi, A., Al-Kharouf, R., Salem, A. (2002). Exploring the Usage of Data Mining and Knowledge Discovery Methodology in e-Learning, Latest trends in applied informatics and computing. Proceedings of the 3rd International conference on Applied Informatics and Computing Theory (AICT '12), Barselona, pp. 204-208.

[18]. Xiao, G. (2011). Association rules algorithm in bank risk assessment. In *Advanced Electrical and Electronics Engineering* (pp. 675-681). Springer, Berlin, Heidelberg.

[19]. Salleb, A., Turmeaux, T., Vrain, C., & Nortet, C. (2004). Mining quantitative association rules in a atherosclerosis dataset.. Proceedings of the Sixth European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, Italy, September 20-24, pp.98-103

[20]. Saltos, G., & Cocea, M. (2017). An exploration of crime prediction using data mining on open data. *International Journal of Information Technology & Decision Making*, *16*(05), 1155-1181. doi: 10.1142/S0219622017500250

[21]. Piatetsky, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. *KDD News*.

[22]. Idoine, C. et. al. (2018). Magic Quadrant for Data Science and Machine-Learning Platforms, Retrieved from: https://www.gartner.com/doc/3860063/magic-quadrant-data-science-machinelearning.