ICIC Express Letters
Part B: Applications
Volume **14**, Number **6**, June **2023**

ICIC International ©2023 ISSN 2185-2766
pp. **633–639**

ASSOCIATION RULE MINING OF MARITIME EMPLOYABILITY DEMANDS BASED ON APRIORI ALGORITHM

Fang Peng[1], Yuhui Sun[2], Zigen Chen[1,*] and Jing Gao[2]

[1]School of Maritime Economics and Management
Dalian Maritime University
No. 1, Linghai Road, Dalian 116026, P. R. China
pengfang@dlmu.edu.cn; *Corresponding author: chenzigen2014@dlmu.edu.cn

[2]UniSA STEM
University of South Australia
Mawson Lakes Blvd, Mawson Lakes SA 5095, Australia
Yuhui.sun@mymail.unisa.edu.au; Jing.gao@unisa.edu.au

Received November 2022; accepted January 2023

Abstract. *With the application of emerging advanced technologies in the maritime industrial revolution, the training of high-level and compound maritime talents is facing new challenges and requirements. Existing researches lack a comprehensive investigation outside the traditional professional maritime skillset and lack a detailed analysis of their impacts on employability. This research hopes to study the relationship between the employment status and the maritime graduates' employability to conduct in-depth discussions on the current maritime graduates' employability. This research proposed the improved Apriori algorithm to identify the association rule of the employability indicators and the employment status. Based on the questionnaire data of the maritime graduate, this research conducts the analysis of the employability skillset required for different employment types. The analysis outcome suggests that the improved Apriori algorithm could identify the important association relationship but have certain limitations. This research finds that certain employment types emphasize certain employability skills, such as responsibility and core professional skills.*
Keywords: Employability, Association analysis, Apriori algorithm, Education analysis

1. **Introduction.** The maritime industry involves the core logistics operation for cargo transportation worldwide, and its international influence has increased dramatically in the past decades [1]. With the accelerated application of advanced technologies for the maritime market, the vigorous development of the maritime industry is inseparable from the firm guarantee of adequate maritime talent supply [2]. However, the maritime talents' employability system requires a comprehensive revision to meet the actual industrial requirements, due to the technological revolutions that have taken place in ships, electronic systems, and automatic control systems [3]. Effectively cultivating comprehensive maritime talent is an important research issue. Overcoming such difficulties and challenges from the development of the maritime industry requires an adequate and comprehensive assessment of the current maritime employment status and employability skills. Current studies on maritime employability commonly focus on analyzing the significance of employability skills in different scenarios [4]. Those studies lack the analysis of the correlation between the employability skillset and the general status of the employment market. The association rule mining has been used to assess education; however, the main research focuses are on the teaching methods and the improvement of student performance and attitude [5].

DOI: 10.24507/icicelb.14.06.633

Association rule mining is an essential method for exploring the relationship in large datasets across many research domains, including the education analysis [6]. With the advancement of education digitalization, considerable information and data on the students' performance could be collected for further analysis [7]. Those datasets could contain certain teaching rules and learning patterns, which could be uncovered by the association rule mining. The identified rules could assess the status of the teaching and learning process and guide the improvement of the education practice. Thus, the association rule mining for education has received much attention from scholars. Alangari and Alturki [8] used the data mining technique to assess the performance and its association with student attributes such as course grades. García et al. [9] used the association rule mining to analyze the web-based course data to obtain the association between student usage and the course information, thus establishing the course recommendation structure.

The Apriori algorithm is one of the commonly used algorithms for association rule mining, which has the advantage of easy implementation and intuitive interpretation and has satisfying performance with large datasets [10]. The core principle of the algorithm is that if one itemset is the frequent itemset, then all its subsets are the frequent itemset; if one itemset is not the frequent itemset, then all its subsets are not the frequent itemset. The algorithm identifies the association rule by finding all the frequent itemset in the data and then exploring the association rules from the frequent dataset [11]. However, the algorithm could be slow with a large volume of the dataset, a high dimensional itemset, or a low support value. The large data volume could generate a large initial data set; the low support value could result in a more frequent itemset for each iteration, and the high dimensional dataset could result in significantly more iterations [12]. Also, with a large number of frequent itemsets and many transactions, the algorithm requires significant computing resources such as CPU and memory [13]. Therefore, the Apriori algorithm requires improvement based on the specific task to achieve the desired outcome.

This study aims to identify employability skills by analyzing the employment questionnaire data with the improved Apriori algorithm. The improved Apriori algorithm proposed by this study has simplified the candidate generation method based on the structure of the questionnaire to improve the algorithm efficiency. This study also designed the self-adjust minimum support generation method to improve the accuracy of selecting the association rules. With the improved Apriori algorithm, this study has identified the association rule between employability skills and certain employment types. The identified association rules are further analyzed to explore the employment pattern of the current employability demands. The contribution of the research is two-fold: the first contribution is that the proposed Apriori algorithm could improve the employability analysis efficiency; the second contribution is that the association rules guide the development of the students' employability towards the requirement of the employment market. Following this introduction, the proposed Apriori algorithm for questionnaire analysis is presented in Section 2; Section 3 outlines questionnaire design and data collection as the context of the data analysis; Section 4 presents the detailed analysis process and the identified association rules; Section 5 discusses the employability demands based on the identified rule and the limitation of the algorithm. The last section concludes this study.

2. **Proposed Apriori Algorithm for Questionnaire Analysis.**

2.1. **Apriori algorithm design.**

The essential process of the Apriori algorithm is about calculating the support value and the confidence value. The frequent set is identified by finding all the items meeting the support and confidence threshold. For different items (X and Y), support for item X is the ratio of item X in all itemset.

$$Support(X,Y) = P(X,Y) = \frac{number(X,Y)}{number(All\ sample)} \tag{1}$$

The confidence of the generated association rules from the frequent itemset is the ratio of the number of occurrences of items containing both X and Y to the number of the occurrences of items containing X:

$$Confidence(X,Y) = P\left(\frac{Y}{X}\right) = \frac{P(X,Y)}{P(X)} = \frac{number(X,Y)}{number(X)} \tag{2}$$

The basic process of the algorithm is

- Identifying all the items which fit the parameter $k = 1$, and conducting dataset scanning to determine the support value for each $k = 1$ item,
- Identifying the $k = 1$ itemset with support higher than the minimum support as the frequent itemset to obtain the frequent set $S(k = 1)$,
- Generating new candidate k itemset using the frequent set $S(k - 1)$ from last iteration,
- Rescanning the dataset to determine the support value for each k itemset candidate,
- Identifying the k itemset candidate with a support value higher than the minimum support, and forming the frequent set $S(k)$, and
- Continuing the iteration until the algorithm does not generate a new frequent itemset.

2.2. **Improved candidate generation method.** One of the significant limitations of the Apriori algorithm is that the generation of a large number of itemset leads to a significant computation burden of scanning the database to calculate the support value for each generated itemset. This research proposes a new approach to generating the itemset according to the questionnaire setting and the research objective of this research. This research intends to study the employability skills that are highly demanded for different types of employment enterprises; therefore, the itemset generating has the following rules: From the second iteration, the itemset candidate should only contain one item from the employability indicator itemset "I" and contain one or several items from the background variable itemset "B". For the latter iteration, due to the limited size of the background set "B", each itemset generated in the iteration produces fewer itemset than the original algorithm. The generation processes are

- Generating the itemset candidate with the standard itemset generation method,
- Checking if the itemset's intersection between the background set "B" is not empty and between the indicator set "I" has only one element, and
- Outputting the filtered itemset.

2.3. **The automatic support threshold generation method.** The Apriori algorithm requires the users to set up the minimum support value and the minimum confidence value as the selection criteria. The selection of those criteria thresholds requires experience in algorithm implementation and in-depth knowledge of the study subject. However, for many researchers in the education field, determining the suitable criteria thresholds could pose a challenge. Thus, this research proposes auto-generating the minimum support value based on the attribute of the datasets to effectively and adequately identify the frequent itemsets. The minimum support is calculated by calculating the mean of each item's support plus the standard deviation of the support values. The $Support[i]$ is the support value of item i. N is the total number of items.

$$Mean = \frac{\sum_{i=1}^{N} Support[i]}{N} \tag{3}$$

$$Standard\ deviation = \sqrt{\frac{\sum_{i=1}^{N}(Support[i] - Mean)^2}{N-1}} \tag{4}$$

Both the new itemset candidate generation and the self-generated minimum support values are used for each Apriori algorithm iteration. Since the modifications of the Apriori algorithm just concerns the generations of the itemset candidates and the criteria thresholds, the main workflow of the algorithm remains unchanged.

3. Questionnaire Design and Data Collection.

3.1. The design of the maritime employment questionnaire.

The questionnaire is based on multiple questionnaires for student capability measurement from Dalian Maritime University. The structure of the questionnaire influences the candidate generation method proposed in this paper. The questionnaire contains two parts. The first part is the participants' background information, and the second part requires respondents to assess and score the importance of various employability indicators. There are four catalogues for the background information: the studied major (nautical science, marine engineering, electronic and electrical engineering), geographical origin (coastal provinces, inland provinces), employer type (state-owned enterprise, private enterprise, enterprise dispatching labour abroad), and employer size (large, medium, or small enterprises). There are 14 employability skills which are listed in Table 1. The questionnaire adopts the Likert 5-point scale as the test method to score the corresponding employability indicators.

Table 1. Initial employability indicator system for maritime graduates

Primary indicator	Secondary indicator	Index
Personal qualities	Adaptability	I1
	Learning and self-development capability	I2
	Critical thinking	I3
	Responsibility	I4
Foundational capability	General management capability	I5
	Language skills	I6
	Teamwork and communication capability	I7
	Problem-solving capability	I8
Professional capability	Nautical competence	I9
	Equipment operation and maintenance capability	I10
	ICT skills	I11
	Cargo management capability	I12
	Maritime business skill	I13
	Implementation of international conventions	I14

3.2. Data collection.

This research conducts a large-scale survey to explore the character of employability indicators. This paper selects a wide range of companies hiring maritime graduates recruitment for the large-scale survey, including state-owned companies, private companies, seafarer dispatch agencies, and other types of companies. The questionnaires are distributed through the alumni association, employment, and other related organizations of Dalian Maritime University. Personals are invited to participate in the online survey anonymously. The questionnaire survey started in August 2019 and ended on October 31, 2019; a total of 2,354 questionnaires have been received. After the questionnaires were retrieved, preliminary sorting was carried out to eliminate 742 invalid questionnaires. There were 1,612 qualified questionnaires with an effective questionnaire recovery rate of 68.48%.

4. Data Analysis.

4.1. Data formatting and parameter setting.

The data of the questionnaire adopt the 5-point scale, which is not suitable for the algorithm; thus, the questionnaire data is formatted according to the algorithm design. The core principle of the formatting is about transforming the numerical data into text data with the indication of the question index. There are two sections of the questionnaire: one for background questions and one for employability indicators. For the answers to background questions, the letter "B", the question number and the letter for the answer are used to form the text data (for example, B1_A). For the employability question, the 5-point scale is interpreted with the following methods. The scores of 4 and 5 are regarded as important for the corresponding employment type and marked with the text "I" in the formatting; The score of 3 is regarded as neutral and marked with "N"; the scores of 1 and 2 are regarded as not important and marked as "NI". Therefore, the letter "I", the question number, and the text label are used to form the text data (for example, I1_important). The questionnaire data are exported and formatted with excel. A sample for the data formatting is demonstrated in Figure 1. The only parameter required for the analysis is the minimum confidence to identify the strong association rules.

Original questionnaire data example

Participants number	Background question section				Indicator question section						
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q5	…	Q14
1	C	A	A	B	5	3	4	2	5	…	3
2	A	B	C	A	4	2	1	5	4	…	4

Dataset after formatting example

Participants number	Background item group				Indicator item group						
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q5	…	Q14
1	B1_C	B2_A	B3_A	B4_B	I1_I	I2_N	I3_I	I4_NI	I5_I	…	I14_N
2	B1_A	B2_B	B3_C	B4_A	I1_I	I2_NI	I3_NI	I4_I	I5_I	…	I14_I

FIGURE 1. The example of the formatting of the questionnaire data

4.2. Identified frequent items and rule generation.

After the data formatting, the questionnaire data is processed with the improved Apriori algorithm. The support value threshold is automatically generated with the designed formula and outputted by the algorithm. The minimum support value generated is 0.54025. With the support value threshold of 0.54025, the improved Apriori algorithm outputs a total number of 7 frequent items, which are listed in Table 2.

TABLE 2. The frequent itemset identified from the improved algorithm

No	Frequent items	Support
1	(B1_1, I9_important)	0.55087
2	(B3_1, I4_important)	0.54156
3	(B3_1, I9_important)	0.54218
4	(B3_1, I10_important)	0.54777
5	(B4_1, I4_important)	0.55769
6	(B4_1, I9_important)	0.56638
7	(B4_1, I13_not_important)	0.54404

Then the association rules are generated from the frequent itemset. The association rules with a confidence level of over 0.7 are listed in Table 3. All association rules from the identified frequent itemsets have a high confidence level with a minimum of 0.75. The

TABLE 3. The association rules generated from the frequent itemset

No	Antecedent	Consequent	Attitude	Support	Confidence
1	Nautical science	Nautical competence	important	0.5508685	0.8671875
2	State-owned enterprise	Responsibility	important	0.5415633	0.75649913
3	State-owned enterprise	Nautical competence	important	0.5421836	0.75736568
4	State-owned enterprise	Equipment operation and maintenance capability	important	0.5477667	0.76516464
5	Large enterprises	Responsibility	important	0.5576923	0.78105995
6	Large enterprises	Nautical competence	important	0.5663772	0.79322328
7	Large enterprises	Maritime business skill	not important	0.5440447	0.76194613

high confidence level suggests that the questionnaire data have strong association rules. The identified association rules are mainly related to the employment types of state-owned enterprises and large enterprises.

5. **Result Discussion.**

5.1. **The demands of employability indicators from different employment types.** From Table 3, the state-owned enterprises favour the employability skills of responsibility, nautical competence, and equipment operation and maintenance capability. Most enterprises value responsibility as maritime logistic operations have the characteristics of long operation duration, harsh operating environment, and high reliance on machinery. Therefore, the association rule related to responsibility is expected. The other two indicators are the professional skillset of the maritime major associated with nautical navigation and the equipment operating, which are also the essential aspect of the maritime operation. For the large-size enterprises, the identified association rules are similar to the state-owned enterprise, containing the responsibility and the professional skillset. These rules suggest that the core requirements of different enterprise types are quite similar. However, large enterprises also have one interesting rule: large companies do not value the maritime business skill for graduate employment. As the maritime industry becomes integrated, the general thinking is that business knowledge could be the basic requirement to promote stable and efficient operations. Thus, business skills could be valuable for some enterprises. However, the large enterprise could have significantly more employees, which means that the large enterprise could afford specific personnel to handle the business routine. Therefore, the requirement of graduates in maritime business skills is reduced. For the other two background information types, the Apriori did not produce important association rules. The algorithm detected that the employment type of nautical science emphasizes professional nautical competence. Such an outcome is expected as it is a basic professional skill for that employment type. The location of the maritime enterprise does not appear in the frequent itemsets and the association rules, which means that the enterprise location does not lead to any specific requirement of the employment market.

5.2. **The limitation of applying Apriori algorithm to questionnaire data.** The Apriori algorithm does not generate certain expected outcomes, such as the commonly acknowledged rules that the internationally based enterprise values a high level of language skill. The lack of those expected rules could be explained by the questionnaire participants not being evenly distributed across the background types. For example, most employment in the Chinese maritime market is in the catalogue of state-owned enterprises, which accounts for over 75% of the participants. The participants from enterprise dispatching labour abroad only accounts for about 12%. Thus, the item "B3_3" will not be identified

as frequent, and all association rules regarding the enterprise dispatching labour abroad will not be detected with the current analysis setting. It can be concluded that the Apriori algorithm only identifies the common association relations and ignores the less common and important relationship in the datasets.

6. **Conclusion.** This research proposes an improved Apriori algorithm to study the employability skill demands of different employment types. The results show that the proposed Apriori algorithm could effectively identify the association rule from the educational questionnaire data. However, the algorithm has the issue of cannot identify the association relationship in the minor data category. Therefore, the application of Apriori algorithm has the potential in questionnaire data analysis but requires further improvement. The association relationship finding of this research could help improve the employability cultivation of the maritime graduate to meet the need of different employment types. Further research directions could be optimizing the generating rules of minimum support value for more accurate rule detection and improving pattern matching method to reduce total database scan over the whole calculation.

REFERENCES

[1] I. de la Peña Zarzuelo, M. J. F. Soeane and B. L. Bermúdez, Industry 4.0 in the port and maritime industry: A literature review, *Journal of Industrial Information Integration*, vol.20, 100173, 2020.

[2] D. Lazarus, Promoting self-responsibility: Learning from Australian maritime engineering student and alumni in developing employability competencies, *IOP Conference Series: Earth and Environmental Science*, 2018.

[3] M. M. Nasaruddin and G. R. Emad, Preparing maritime professionals for their future roles in a digitalized era: Bridging the blockchain skills gap in maritime education and training, *Proc. of the International Association of Maritime Universities (IAMU) Conference*, 2019.

[4] P. S.-L. Chen et al., Employability skills of maritime business graduates: Industry perspectives, *WMU Journal of Maritime Affairs*, vol.17, no.2, pp.267-292, 2018.

[5] S. K. Verma and R. Thakur, Fuzzy association rule mining based model to predict students' performance, *International Journal of Electrical & Computer Engineering*, vol.7, no.4, 2017.

[6] T. Li, Application of APRIORI correlation algorithm on music education curriculum association rules, *Journal of Physics: Conference Series*, 2021.

[7] S. Borkar and K. Rajeswari, Predicting students academic performance using education data mining, *International Journal of Computer Science and Mobile Computing*, vol.2, no.7, pp.273-279, 2013.

[8] N. Alangari and R. Alturki, Association rule mining in higher education: A case study of computer science students, in *Smart Infrastructure and Applications*, R. Mehmood, S. See, I. Katib and I. Chlamtac (eds.), Springer, Cham, 2020.

[9] E. García et al., An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering, *User Modeling and User-Adapted Interaction*, vol.19, no.1, pp.99-132, 2009.

[10] T. Pan, An improved Apriori algorithm for association mining between physical fitness indices of college students, *International Journal of Emerging Technologies in Learning (IJET)*, vol.16, no.9, pp.235-246, 2021.

[11] M. Hegland, The Apriori algorithm – A tutorial, *Mathematics and Computation in Imaging Science and Information Processing*, pp.209-262, 2007.

[12] T. M. Hossain, J. Watada, Z. Jian, H. Sakai, S. Rahman and I. A. Aziz, Missing well log data handling in complex lithology prediction: An NIS Apriori algorithm approach, *International Journal of Innovative Computing, Information and Control*, vol.16, no.3, pp.1077-1091, 2020.

[13] L. Liu et al., A task scheduling algorithm based on classification mining in fog computing environment, *Wireless Communications and Mobile Computing*, vol.2018, 2018.