Formation of Life Quality Indicators System through Search Algorithm of Association Rules

Lyudmila P. Bilgaeva, Dashidondok Sh. Shirapov, and Grigoriy V. Badmaev

East Siberia State University of Technology and Management, Russia
http://www.esstu.ru

Abstract. The paper is devoted to the search of association rules for the formation of the indicators system that affects the quality of life. The search of association rules is carried out in the transactional database based on the method of AprioriTid algorithm to calculate such metrics as support, confidence and lift. It results in the extraction of useful association rules showing the relationship of life quality indicators, which can be used later to solve the problems of analysis and forecasting.

Keywords: extraction algorithm of frequent sets of database, the property of monotony, the associative search of life quality indicators, truncation of candidates

1 Introduction

At present, issues of life quality are relevant, as the current economic crisis has primarily affected the population. In general, the standard of living depends on a competent social policy pursued by the state. Solving social problems requires the adoption of management decisions based on real information. This requires research aimed at identifying the main factors affecting the life quality.

In this paper we propose to use methods of searching association rules to identify the most important indicators of life quality that will enable the authorities to plan and implement certain measures to improve the population living standards.

To search association rules is one of the tasks of Data Mining, the modern technology of intellectual data analysis, which includes finding regularities between some related events, the identification of related objects and their location in the space of states. To find associations such a database is typically used in which all objects are connected to each other, provided that the database is consistent and integrative.

2 Basic theoretical principles of association rules search

There are many techniques, which allow solving the problem of finding association rules. They have the same mathematical approach, but the ways of the method implementation are different. Let us consider the basic theoretical principles of these methods.

The association rule of context K is an expression of the form

$$A \to B,$$

where $A, B \subseteq M$.

The context K is a tuple (G, M, I), where G is a set of objects, M is a set of features, but $I \subseteq G \times M$.

When association rules are searched, special metrics are used: Support, Confidence, Lift.

Association rule $A \to B$ Support is a quantity defined by the formula:

$$\text{Support}(A \to B) = \frac{|(A \cup B)'|}{|G|} \tag{1}$$

The Support value indicates which part of the G objects contains $A \cup B$. The Confidence of the association rules is defined by the formula:

$$\text{Confidence}(A \to B) = \frac{|(A \cup B)'|}{|A'|} \tag{2}$$

The Confidence value shows, which part of the objects that contain A, also contains $A \cup B$.

The following quantity is called the association rule utility (Lift):

$$\text{Lift}(A \to B) = \frac{|(A \cup B)'|}{|A'| \cdot |B'|} \tag{3}$$

In other words, the utility is the ratio of Confidence$(A \to B)$ to the Support(B). The Lift value indicates the usefulness of the rule. If the found utility value is more than 1, then the rule is considered to be useful.

The task of mining Association rules is to find all Association rules of the context for which the values support and confidence exceed certain set values `min_support` and `min_confidence`, respectively.

Searching the frequent sets of data is limited to the minimum support value (`min_support`), which is set by the user [1–3]. Search of association rules is made within the frequent sets of data and is limited to the minimum confidence (`min_confidence`) and utility value. The minimum confidence is generally set by the user.

AprioriTid method, as well as the Apriori method, is based on the anti-monotony property, the key property when finding multielement frequent sets of data [4, 11]. It is formulated as follows:

$$\forall A, B \subseteq M, \quad A \subseteq B \Rightarrow \text{Support}(B) \le \text{Support}(A) \tag{4}$$

It means that:

- with an increase of the set size its support either decreases or does not change;
- for any set of characteristics support does not exceed the minimum support of any of its subsets;
- the set of n size characteristics will be frequent only if all its $n-1$-element subsets are frequent.

3 Valid method choice

To select a search method of the association rules the authors developed definite criteria and comparatively analyzed the certain amount of methods. The results are given in Table 1.

Table 1. Comparative analysis of methods of association rules search

No.	Methods	Criteria				
		Implementation simplicity	Small number of candidates	Speed	Application of ID transactions	Possibility of candidates' truncation
1	AIS	+	−	−	−	−
2	SETM	+	−	−	+	−
3	Apriori	−	+	+	−	+
4	AprioriTID	+	+	+	+	+
5	AprioriSome	−	+	+	−	+
6	FPG	−	+	+	−	+

The most appropriate method to solve the task of the associative search of indicators affecting the population life quality, is the AprioriTid method proposed by the group of authors [2].

Simplicity of implementation is associated with such a data structure as a table for storing intermediate results.

Other methods, e. g. Apriori or FPG, use trees as data structures. These data structures are more complicated to implement [6, 7, 9]. There are methods of searching for association rules based on the Boolean matrix [5, 8].

It is convenient to extract data from a database applying database records identifiers, i. e. TID. TID also enables you to identify whether the generated rules belong to a particular database record.

The possibility to truncate candidates allows cutting useless and unreliable rules at their generation stage in order to optimize the memory used.

4 Software module of associative search of population life quality indicators

To solve the problem of the formation of a system of indicators that affect life quality of the population, we developed a system the architecture of which is presented in Figure 1.

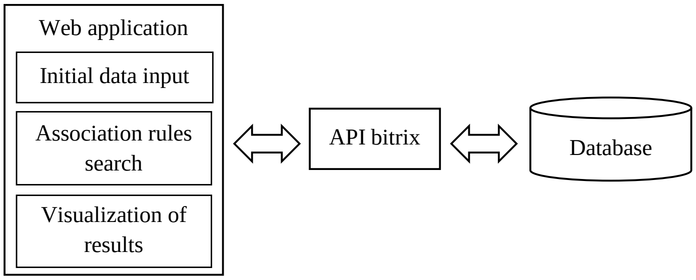


Fig. 1. Architecture for Association rules mining system

The system consists of a web application and a database interacting with each other through the API bitrix component. The database was created in a DBMS MySQL. As a web server a freely distributable program OpenServer is used. It is a portable server platform, which is a medium for web development. The Web application is composed of five pages: the main page, parameters setup, transactions and attributes management, generation of rules, and visualization of results.

The system starts with setting up the parameters, such as the minimum support (`minsup`), the minimum confidence (`minconf`) and a serial number of the experiment. The transaction content, i. e. each record in a database table, is a set of possible attributes which are coded indicators of life quality. For example, in a database entry {1, 5, 7}, 1 is an indicator of “Actually available income of the population, %”, 5 – “Life expectancy at birth in years”, 7 – “The Gini coefficient (income concentration factor)”.

Minimum support and minimum confidence are specified by the user.

While conducting experiments one can consider various transaction and attributes sets, therefore such a parameter as a “serial number of the experiment” is used.

The function of rules generation is based on the AprioriTid method, the block diagram of which is shown in Figure 2.

It starts with generating single-element data sets that are candidates for rules. Support, i. e, the number of repetitions in all database transactions involved in the experiment, is counted for each of them.

Then two-element sets, three-element sets, ..., i-element sets, where $2 \leq i \leq k$, are generated in the iteration.

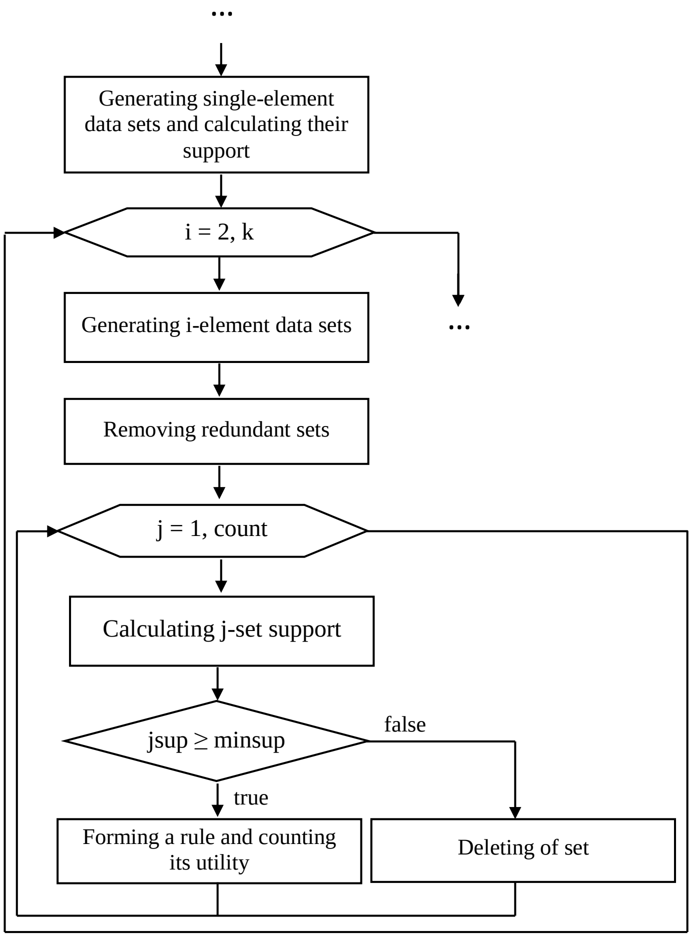

...
Generating single-element data sets and calculating their support
i = 2, k
...
Generating i-element data sets
Removing redundant sets
j = 1, count
Calculating j-set support
jsup ≥ minsup
false
true
Forming a rule and counting its utility
Deleting of set

Fig. 2. Block diagram of association rules generation

The same sets that are redundant are removed from the resulted sets.

After that support is calculated for each of the remain database sets, then the current set support value `jsup` is compared with the minimal support `minsup`, set by the user.

If the condition `jsup` $\leq$ `minsup` is met, then the association rule formation begins, otherwise the current set is removed.

Confidence and utility (Lift) are calculated for the generated rule.

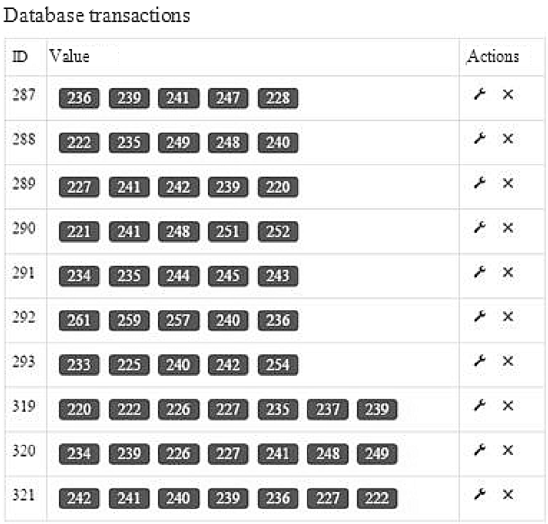
If the Confidence value is greater than or equal to the minimum confidence value and the lift value is greater than or equal to 1, then the rule is considered to be credible and useful, otherwise it is deleted.

Visualization of the results allows us displaying the initial transactions, frequent sets of data and their support, the generated association rules and the values of the confidence and utility parameters for each of them.

5 The results of the experiments

We made many experiments with the AprioriTID method of association rule to search for a system of indicators that affect life quality. The subsystem of the indicators proposed by the authors in [10] was taken as input data. This subsystem provides eight main indicators of the population life quality and the factors that influence on each of them.

Database transactions were formed from the original data, which contained a various number of attributes representing the coded life quality indicators and factors distinguished according to the experts' opinion. Overall, there were formed 25 transactions with the various number of attributes from five to seventeen. When using the transactions with five attributes and more, fourteen ones included, there were no results of the experiments. The generation of association rules begins with using 15 attributes in a transaction. Figure 3 shows a fragment of the original database transaction with five and seven attributes.

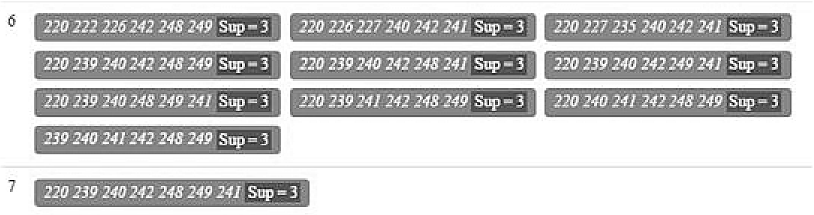


Database transactions

ID	Value	Actions
287	236 239 241 247 228	🔧 ✕
288	222 235 249 248 240	🔧 ✕
289	227 241 242 239 220	🔧 ✕
290	221 241 248 251 252	🔧 ✕
291	234 235 244 245 243	🔧 ✕
292	261 259 257 240 236	🔧 ✕
293	233 225 240 242 254	🔧 ✕
319	220 222 226 227 235 237 239	🔧 ✕
320	234 239 226 227 241 248 249	🔧 ✕
321	242 241 240 239 236 227 222	🔧 ✕

Fig. 3. Original transactions with five and seven attributes

In Figure 4 you can see a fragment of frequent item sets containing six or seven attributes, the support value of which is equal to three. Four valid useful rules presented in Table 2 were generated based on the frequent item sets above.



```
6  220 222 226 242 248 249 Sup = 3   220 226 227 240 242 241 Sup = 3   220 227 235 240 242 241 Sup = 3
   220 239 240 242 248 249 Sup = 3   220 239 240 242 248 241 Sup = 3   220 239 240 242 249 241 Sup = 3
   220 239 240 248 249 241 Sup = 3   220 239 241 242 248 249 Sup = 3   220 240 241 242 248 249 Sup = 3
   239 240 241 242 248 249 Sup = 3
7  220 239 240 242 248 249 241 Sup = 3
```

Fig. 4. Fragment of the frequent item sets with six or seven attributes with their support values

Table 2. Valid useful rules

Rules	Confidence	Lift
248 → 249	0.857142857143	1
251 → 252	1	1
257 → 259	1	1
234, 243 → 244	1	1

The experiment resulted in the generation of fourteen valid and useful association rules. Since any association rule is an operation of implication, it is possible to combine them through a conjunction operation provided that the conjunction is true. After converting a logical expression five association rules were obtained. They are represented in Table 3.

Table 3. Results of the experiments

No.	Number of database transactions	Association rules
1	15	$251 \to 252$
2	16	$(248 \to 249) \wedge (251 \to 252)$
3	18	$(248 \to 249) \wedge (251 \to 252) \wedge (257 \to 259)$
4	20	$(257 \to 259) \wedge (234 \wedge 243 \to 244) \wedge (248 \to 249)$
5	23	$(248 \to 249) \wedge (257 \to 259) \wedge (234 \wedge 243 \to 244) \wedge (230 \wedge 235 \to 238 \wedge 239 \wedge 241)$

Here it is seen that to generate the association rule 251 → 252 15 database transactions were used. This rule means that the “Mortality” indicator (252) is affected by the “Birth rate” indicator (251).

Or, for example, Rule $230 \wedge 235 \to 238 \wedge 239 \wedge 241$ means that “Life quality index” (230) and “Purchasing power” (235) indicators are influenced on with such indicators as “Paid services volume per capita” (238), “Growth rate of the minimum subsistence level" (239) and "Employment rate of the population" (241).

During the experiments the graphs were plotted. Figure 5 shows the graph of relation between the number of rules and the number of transactions, a trend line was made.

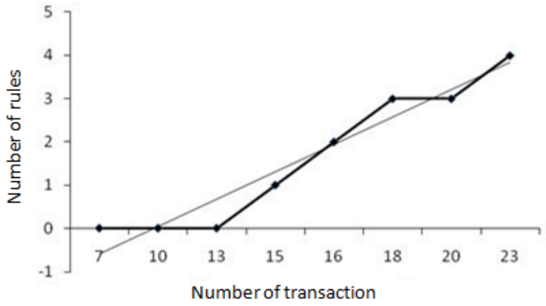


Fig. 5. Graph of relation between the number of rules and the number of transactions

Figure 5 demonstrates that the number of rules depends on the number of database transactions. The greater the number of transactions is, the more association rules are generated, as evidenced by the trend line.

In another chart shown in Figure 6, you can see the dependence of the number of rules on the number of features in the transaction and the trend line. It should be noted that the more elements in the transaction are, the more association rules are generated. For example, if you have 12 features in the transaction the maximum number of rules generated is equal to 4. You can see that the value 4 corresponds to 23 transactions, each one including 12 features, as shown in Figure 6.

Therefore, we can conclude that the number of rules depends on the number of database transactions and the number of features in these transactions.

6 Conclusion

Computational experiments with the developed software were carried out. They enabled us to obtain valid and useful association rules for the population life quality indicators, the number of which depends on the input data.

The experiments outcome shows that the indicators and factors in each association rule are interrelated. In addition, the results obtained demonstrate that it is possible to generate valid and useful association rules based on a transactional database. Having performed logical transformations over them, one can create a system of life quality indicators, which then can be used to solve problems of analyzing and forecasting the population life quality.

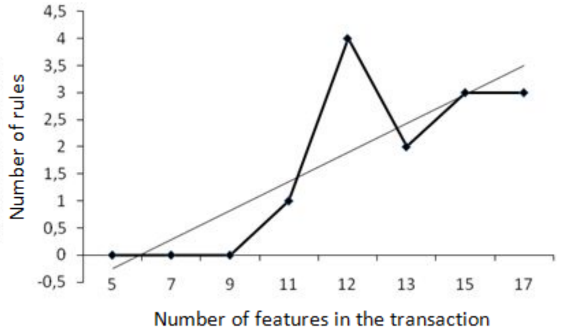


Fig. 6. Dependence of the number of rules on the number of features in the transaction

This approach will enable the state authorities to correct and reasonably develop strategic social and economic programs to improve the population life quality.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the ACM SICMOD conference on management of data. pp. 207–216. Washington, D.C. (1993)
2. Agrawal, R., Mannila, H., Stricant, R., Toivonen, H., Verkamo, A.I.: Advances in knowledge discovery and data mining, chap. Fast Discovery of Association Rules, pp. 307–328. American Association for Artificial Intelligence Menlo Park, CA, USA (1996)
3. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20^{th} International Conference on Very Large Databases. pp. 487–499. Santiago, Chili (1994)
4. Billig, V.A., Tsaregorodcev, N.A., Ivanova, O.V.: Building association rules in medical diagnosis. International Journal of Software & Systems 2, 146–157 (2016)
5. Liu, H., Wang, B.: An association rule mining algorithm based on a boolean matrix. Data Science Journal 6, Supplement, 559–565 (2007)
6. Olson, D.L., Delen, D.: Advanced Data Mining Techniques. Springer Publishing Company, Incorporated (2008)
7. Oreshkov, V.: Fpg – an alternative search algorithm for association rules (2014), uRL: https://basegroup.ru/community/articles/fpg
8. Rao, C.S., Babu, D.R., Shankar, R.S., Kumar, V.P., Rajanikanth, J., Sekhar, C.C.: Mining association rules based on boolean algorithm – a study in large databases. International Journal of Machine Learning and Computing 3(4), 347–350 (2013)
9. Sahaaya Arul Mary, S.A., Malarvizhi, M.: A new improved weighted association rule mining with dynamic programming approach for predicting a user's next access. In: Proceedings of the ICAITA conference. vol. 2, pp. 105–122. Dubai, UAE (2012)

10. Saktoev, V.E., Sadykova, E.T.: Sustainable Development of Regional Economic Systems with Environmental Regulations. ZAO “Economy”, Moscow, Russia (2011)
11. Zayko, T.A., Oleinik, A.A., Subbotin, S.A.: Association rules in data mining. Bulletin of NTU “KhPI” 39(1012), 82–95 (2013)