A graph-graph approach to the analysis of the set of associative rules

S V Belim, T B Smirnova, and A N Mironenko
Dostoevsky Omsk State University, 55A, Mira ave., Omsk, 644077, Russia

Abstract. The article proposes a method for processing a set of associative rules, which makes it possible to identify additional relationships between the set of objects under study. The proposed approach consists of three stages. At the first stage, associative rules are revealed from the statistical data. At the second stage, constructed weighted oriented graph of relationships between the objects of the system. The third stage analyses the graph and identifies the community (community), which allows you to determine the groups of the most related objects. As an example, is given an analysis of the activities of public organizations. The result of the work is a method that allows to identify patterns from the analysis of a set of associative rules, and not just from a separate associative rule.

1. Introduction
The traditional use of the method of forming of associative rules from dataset consists of the study of peculiarities of consumers demand for goods in trade networks [1, 2]. Information about the correlation interest to buying of various goods, is used later for their positioning in trading rooms. In recent times, observed an expansion of the use of this method. For example for processing images [3, 4], analysis dissemination of biological species [5], search by photo of the surface points of interest in mining [6] and others.

Quite a large number of works devoted to researching social networks with help of searching association rules. So in article [7] on the basis of association rules are detected mutual influence users of social networks. Then stand out the most influential users, which can be used for effective dissemination of information. In works [8, 9], based on the associative rules formed from Facebook entries, a connection is established between the student's gender and the selected training courses. Similar researching’s [10], based on Facebook entries, and allowed to study the influence of the social network on the educational process of students in Turkey. In [11], the authors on the basis of the analysis of social networks identify factors that affect the hobby of users. The possibilities of using associative rules in criminology are investigated in the work [12]. The authors of this work investigated the correlation between motives of crime and the type of crime. In article [13], with the help of associative rules, constructed a social network is based on the database of terrorist attacks. The paper [14] is devoted to the study of the influence of the nature of people on the addiction to drug addiction through the construction of associative rules based on data from narcological clinics. The application of the method of searching for associative rules to population census data within the framework of the SPIN project is presented in [15].

A set of associative rules are formed automatically based on statistical data. After the associative rules that satisfy certain statistical constraints are obtained, is carried out the search for «interesting»
relationships. To interesting, usually include non-obvious correlations between objects. At this stage, we have to face a number of difficulties. First, the selection of associative rules is carried out manually, which can create significant difficulties with a large database of transactions and a large number of identified associative rules. Secondly, we have to limit ourselves to a relatively small amount of generated transactions with a high degree of support. In this case, the information contained in the remaining associative rules is lost. Thirdly, each associative rule is considered separately, which can lead to the loss of regularities, manifested by transitivity. Thus, the problem of constructing algorithms for automatic processing of a large set of associative rules is actual.

This article proposes a method for processing a set of associative rules formed on the basis of statistical data based on the construction of a graph model, taking into account information contained in associative rules with low support, as well as interrelations between objects, manifested by transitivity.

2. Statement of the problem and methods of solution

Consider the set of associative rules \( A \), which are formed on the basis of the set of transactions \( T \).

Let \( I \) be the set of all answers that can be present in the transaction. Each transaction \( t \) is a collection of elements from \( I \) \( (t \subseteq I) \). \( D \) is the set of all transactions. It is said that transaction \( T \) contains a set of elements of \( X \) if \( X \subseteq t \) and \( X \subseteq I \). An associative rule is the implication \( X \Rightarrow Y \), where \( X \subseteq I \), \( Y \subseteq I \), and \( X \cap Y = \emptyset \).

Every associative rule characterized by a certain set of parameters. The first parameter, called support, shows the frequency detector of this rule in an existing set of transactions. Support rule’s \( X \Rightarrow Y \) is calculated as a percentage of transactions containing lots of \( X \cup Y \):

\[
\text{supp}(X \Rightarrow Y) = \left( \left( N \times (X \cup Y) \right) / |D| \right) \times 100\% 
\]

where \( N \times (X \cup Y) \) is the number of transactions containing the set \( X \cup Y \).

The validity of the rule shows the probability with which \( Y \) follows from \( X \). The reliability of the associative rule \( X \Rightarrow Y \) is calculated as the percentage of transactions containing both \( X \) and \( Y \) in the set of transactions containing \( X \):

\[
\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \Rightarrow Y)}{\text{supp}(X)}
\]

The task of finding associative rules is to find sets of elements whose support is not lower than \( \text{minsupport} \). From the found sets, rules are distinguished with the reliability not less than \( \text{minconfidence} \).

Build a graph \( G \) of correlations based on the identified Association rules. As the set of vertices of the graph, we select the set of objects encountered in the constructed associative rules. The edges will correspond to associative rules, so the graph will be oriented. As the weight of the edges we will use the validity of the corresponding associative rules.

Apply the search community’s algorithm (community) [16] to this graph. To do this, you must first obtain reduce form the weights matrix \( e = E/m \), where

\[
m = \sum_{i,j=1}^{N} E_{ij}
\]

\( N \) is the number of vertices. In the reduced the weights matrix, the element \( e_{ij} \) shows the weight fraction of the given arc in the total weight of the graph. In the future, under the weights matrix, the reduced form will be understood. Easy to see, that

\[
\sum_{i,j=1}^{N} e_{ij} = 1
\]
To identify communities uses modularity function (modularity), showing the optimality of dividing a graph into sub-graphs:

\[ Q(e) = \sum_{i=1}^{N} e_{ii} - \sum_{i=1}^{N} p_{ii} \]  \hspace{1cm} (5)

Where \( p_{ii} \) is expected connectivity. In the canonical model \([16]\), \( p_{ii} \) is defined through the outgoing degree vertices in \( a_i \), and incoming degree vertices \( b_i \):

\[ p_{ii} = a_i b_i \]  \hspace{1cm} (6)

In this case, modularity is written as:

\[ Q(e) = \sum_{i=1}^{N} e_{ii} - \sum_{i=1}^{N} a_i b_i \]  \hspace{1cm} (7)

where

\[ a_i = \sum_{j=1, j \neq i}^{N} e_{ij}, \hspace{1cm} b_i = \sum_{j=1, j \neq i}^{N} e_{ij} \]  \hspace{1cm} (8)

The modularity value shows the difference between the fraction of all edges lying within the communities for a given graph and the expected value computed for a graph in which all vertices have the same degree and the edges are randomly located. The modularity function will have a null value for a graph with a uniform random distribution of edges. Deviation from randomness leads to a positive value for the modularity. Deviation from randomness in the location of edges indicates the presence of regularities in the connection of objects.

To search for communities on graphs algorithm searching couplers is used. Select the sub-graph \( G' \) in the graph \( G \) and replace all its vertices with one vertex, while the vertices of the sub-graph \( G/G' \) remain unchanged. The formed vertex is connected by arcs to those vertices of the graph \( G/G' \) with which the vertices included in the coupler. The sum of the weights of the vertices and arcs that are included in the coupler is equal to the weight of the vertex coupler. The community will call sub-graph source graph, which during the formation of the coupler maximizes the function modularity of the graph \( Q(e) \).

At the output of the algorithm we get a graph consisting of couplers. And the formation of new ties reduces the value of the modularity function. Each scree corresponds to a subgraph in the original graph. Pixels corresponding to the vertices of a single subgraph belong to the same segment. Our task is to identify communities on the graph of associative rules.

3. An example of analysis of a set of associative rules

Consider the results of applying the proposed algorithm to the analysis of the activities of public organizations of Russian Germans. Initial information has been extracted from the questionnaires filled the heads of organizations. As the transaction selected one questionnaire. Each questionnaire questions in different formats. The first type of questions suggested two variants of the answer «Yes» or «No». The second type of question allowed the choice of one of four or five options, and allowed the additional response «other», which was not coded, because it contained uncertainty. Also in the questionnaire there were questions with a nondeterministic answer, but they were not taken into account in the formation of transactions, as all the answers were different and could not lead to the identification of associative rules with any appreciable support. Each variant of the answer was coded by its record. The question was identified by an identifier of one or two Latin letters. For example, the question designated «A» had two response variants, coded «A1» and «A2». In questions containing a choice of two alternatives, one cannot limit oneself to coding only one of them, despite the possibility of an unambiguous recovery of the second. Such a restriction may lead to the loss of associative rules.

A total of 107 were processed questionnaires, each of which was viewed as an independent transaction. After coding the transaction were obtained with different numbers of entries from 24 to 50. On the basis of transaction data, associative rules were searched with the support of not less than 60 %, and accuracy not less than 80 %. For finding Association Rules was used algorithm APriori
The resulting set of associative rules contains the expected interactions. At the same time, it should be noted that in the formation of associative rules, there are only 9 statements from 173 possible. Between the remaining statements, associative rules with sufficiently high support and reliability are absent.

Table 1 lists the notation for claims occurring associative rules. The text of your paper should be formatted as follows:

| Notation     | Description                                                                 |
|--------------|-----------------------------------------------------------------------------|
| $v_1$        | more than 50% of the visitors center are Russian Germans                     |
| $v_2$        | the Internet is used in the work of the centre several times a day           |
| $v_3$        | organization uses language courses for adults                                |
| $v_4$        | German language skills have improved over the past 10 years                 |
| $v_5$        | organizations in Russia are partners in the work of the                     |
| $v_6$        | Cooperation with the MCHK is carried out continuously with high degree of efficiency |
| $v_7$        | In the organization know, that Germany is implementing special programs     |
| $v_8$        | For the study of the German language children's and youth language clubs are used |

Figure 1 show a graph based on association rules with one precondition.
Because the original graph has a small number of vertices, the task can be solved by brute force. Function modularity source graph is equal to $Q=0.1497$. The combining of vertices $v_1$, $v_3$, $v_4$ in a community leads to a value of the function modularity $Q_{1,3,4}=-0.0361$, that is such consolidation is profitable and shows the close communication of these vertices. Combining into one community of vertices $v_1$, $v_3$, $v_4$ and $v_2$ resulted in the value of modularity $Q_{1,3,4,2}=-0.0233$. Other variants of combining do not increase the modularity feature, that is are not profitable.

From this analysis we can conclude about the close association of such aspects of the activity of public organizations of Russian Germans: “more than 50% of the visitors center are Russian Germans”, “the Internet is used in the work of the centre several times a day”, “organizati”. These four areas most closely interrelated and need to be considered together. It should be noted that the relationship identified by the proposed graph-theoretic approach does not follow directly from the formed associative rules. The corresponding set has little support.

4. Conclusion
The article proposes a new method for automatic analysis of a set of associative rules. The representation of connections between objects in the form of an oriented graph allows us to reveal regularities that do not manifest themselves in the traditional «manual» analysis of associative rules. Analysis of the relationship graph with the help of the search for vertex communities makes it possible to determine the most closely related objects of the original set. The increase of informativeness in revealing the regularity in comparison with the traditional approach is based on consideration of more associative rules and taking into account the transitivity relationship. The application of the proposed method to the analysis of the activities of public organizations made it possible to reveal new regularities that do not directly follow from one associative rule.

References

[1] Agrawal R, Imieliński T and Swami A 1993 Proc. ACM SIGMOD International Conference on Management of Data Washington (DC, USA) pp 207–216
[2] Shaheen M, Shahbaz M and Guergachi A 2013 Elsevier Knowledge-Based Systems pp 261–273
[3] Belim S V, Mayorov-Zilbernagel A O 2014 Herald of computer and information technologies vol 12 pp 18–23
[4] Belim S V, Mayorov-Zilbernagel A O 2014 Science and Education of the Bauman MSTU vol 12 pp 716–737
[5] Atepalikhin M S, Kassal B Y and Belim S V 2014 Herald of Omsk university vol 2(72) pp 25–29
[6] Lee I, Cai G and Lee K 2013 46th Hawaii International Conference on System Sciences pp 1580–1588
[7] Erlandsson F, Brodda P, Borg A and Johnson H 2016 Entropy vol 18 p 164
[8] Erlandsson F, Borg A, Johnson H and Bródka P 2016 Advances in Network Science (Springer International Publishing: Cham, Switzerland) pp 126–135.
[9] Nancy P, Geetha Ramani R and Jacob S 2013 Mining of Association Patterns in Social Network Data (Face Book 100 Universities) through Data Mining Techniques and Methods Proceedings of the Second International Conference on Advances in Computing and Information Technology (ACITY) vol 178 (Springer: Berlin/Heidelberg, Germany) pp 107–117
[10] Selman Bozkir A, Güzin Mazman S and Ebru Akçapınar Sezer 2010 Technological Convergence and Social Networks in Information Management IMCW 2010 Communications in Computer and Information Science vol 96 (Springer: Berlin/Heidelberg, Germany) pp 145–153
[11] Yu X, Liu H, Shi J, Hwang J N, Wan W and Lu J 2014 IEEE International Congress on Big Data (Anchorage, USA) pp 310–314
[12] Pereira B L, Brandão W C 2014 IADIS International Conference Applied Computing pp 159–166
[13] Górecki J, Slaninová K 2013 Acta academica karviniensia pp 56–68
[14] Zahedi F, Zare-Mirakabad M R 2014 *Journal of AI and Data Mining* vol 2 pp 135–139
[15] Malerba D, Esposito F and Lisi F A 2001 Proceedings of the Joint Conferences on "New Techniques and Technologies for Statistics" and "Exchange of Technology and Know-how" (ETK-NTTS’01) pp 541–550
[16] Newman M E J 2003 Mixing patterns in networks *Phys. Rev. E* 67 026126
[17] Agrawal R, Srikant R 1994 *Proc. of the 20th International Conference on VLDB* (Santiago, Chile) pp 487–499