UDC 631.15:33

FEATURES OF SEARCHING FOR ASSOCIATIONS IN LARGE VOLUMES OF MARKETING DATA

D. R. Rahel
PhD in Economics, Associate Professor of the Department of Economics of the Belarusian State University of Informatics and Radioelectronics
ragel@bsuir.by

D.R. Rahel

In 2000 he graduated from the Faculty of Economics of the Belarusian State University of Informatics and Radioelectronics with a degree in Economic Informatics. In 2016 he graduated from the postgraduate course of the Academy of Management under the President of the Republic of Belarus. In 2018 he defended his PhD thesis. Research interests: data mining in marketing, process modeling, data analysis, statistical forecasting, macroeconomics.

Annotation. The article describes an approach to searching for associations in baskets of goods and searching for patterns in transactions. In addition, approaches are given to the analysis of big data on the behavior of the customer base, which can be obtained in the course of market research. The stages of the analysis of consumer behavior and the search for patterns in it are described.

Keywords: data analysis, marketing data, big data, data array, marketing research, associations, analysis of the actions, buyers activity, classification, clustering, frequency pattern.

As part of marketing research, one of the main tasks is often to find links and dependencies between elements obtained from different data sources. In fact, we are talking about an associative analysis of disparate and disordered information.

If we consider customer transaction data, which collects data on customer purchases, that is, the goods purchased on a specific date or within a fixed period of time are specified, then the main goal of further marketing forecasting of customer activity is the association between groups of goods or individual goods in the considered datasets. In this case, by association we mean patterns that can be obtained analytically, which can later lead to the development of rules that help optimize the company's commercial activities and classify product categories and customer groups. Many goods are divided into "baskets" with patterns that are inherent in the goods that make them up. For example, if one of the baskets contains “bread, flour, butter”, then it is likely to contain milk as well. And if there is a “hammer, glue” in the basket, then milk is unlikely to be present in it, but rather something from the categories close to these goods, such as, for example, “nails”.

The search for association rules based on the received and systematized data is an activity to maximize income based on the formation of the company's product basket. Taking into account the processing of data and the formation of "baskets", further decisions are made on discounts for certain categories of goods, on the termination of the commercial operation of certain categories of goods, or the launch of new product categories with a higher commercial potential on the market.

Collecting data on transactions is now a normal practice in almost all companies. The formation of association rules for categories should be accompanied by probabilistic estimates, since the implementation of associations in practice is probabilistic. Thus, in the formation of associative rules, the condition and the consequence from it play a role. For example, the condition is “hammer, glue”, and

"nails" is already a consequence that has some kind of probabilistic assessment depending on the practical implementation.

At the stage of practical implementation at the level of a business entity, the following rules are formed:

• at the level of the condition, that is, what was bought with the subsequent generalization of the most typical consequences; (1)

$$\{\text{hammer, glue}\} \rightarrow \{\text{nails}\} \quad (1)$$

• at the level of the investigation, with the formulation of the rules that eventually led to it. (2)

$$\{\text{nails}\} \rightarrow \{\text{hammer, glue, nails}\} \quad (2)$$

If we touch upon the problem statement and apply associations to the database with records of the company's commercial transactions, then for such a database (DB) with a set of data on transactions Tr_1, Tr_2, Tr_3, ..., Tr_n, it is required to find a set of templates that can express any then the minimum of transactions for the considered period of time. The number of transactions that are described within the templates can be denoted as N. The parameter N, in turn, can be represented as the number of transactions selected from the general set that fall under this template. At the same time, each transaction in the data set is expressed as a binary vector, one of the coordinates of which is an attribute that describes the presence of an order for a specific product in a specific outlet. For example, let's take 2 some sets from outlets M1 and M2. Thus, M1 => M2 is an association with a minimum support of H under the conditions:

1. M1 U M2 - is a frequency pattern;

2. The ratio M1 U M2 to M1 has a minimum certainty and is less than 1 if the sets are not equal to each other.

From a computational point of view, a pattern is a vector in the same space as the set of transactions in question. In this case, the template is the result of applying the logical operation "and" to the set of transactions under consideration.

The most typical frequency pattern search algorithm is a pattern-merge based algorithm by sequentially combining patterns of length k to sequentially generate frequency patterns of length k+1. The main goal is to reduce the search space for frequency patterns based on the observation that no pattern can contain low frequency patterns.

1. Input:

• Database.

• Minimum C support.

2. Output:

• Lots of SN frequency patterns.

3. Algorithm:

• Generation of two sets of frequency patterns M1 and M2 respectively with duration D1 and D2.

• k=2

• As long as the template set is not empty:

i. Generation of all possible templates of length k+1 due to pairwise union of available templates;

ii. Exclusion from the set of all patterns of length k+1 that have subpatterns without sufficient support;

iii. Generation of a finite set by copying all the remaining templates of dimension k + 1 into it and calculating their support.

iv. Return a final set.

The search for associative links based on patterns in marketing activities has a number of useful applications for practical use:

• Analysis of the activity of buyers in the framework of the operating activities of trading companies. Transaction information in this case is a collection of information about sets of jointly purchased goods. In this case, the development of templates is the generation of recommendations to potential buyers based on previous experience. If the current basket intersects with the template available in the database, then the buyer receives a recommendation regarding the rest of the goods from the template.

• Carrying out classification or clustering of customers or customer base based on the company's existing patterns of purchases, when all similar transactions are combined in the database based on compliance with existing pattern types of clusters;

• Analysis of the actions of service users. In this case, the database may consist of sets of elementary user actions online, templates in turn for generating recommendations.

• Analysis of application errors. The application generates standardized sequences of logs, while paying attention to the analysis of low-frequency patterns for their fallacy.

References

[1] Мыльников Л.А. Статистические методы интеллектуального анализа данных. – БХВ-Петербург, 2021. – 119 с.

[2] Data Science and Big Data Analytics. A Step by Step Guide to learn Data Science from Scratch with Python Machine Learning and Big Data / Andrew Park. – Published by Andrew Park, 2021. – 124 p.

[3] Nicholson W.L. Exploring Data Analysis. – Nobel Press, 2012. – 421 p.

ПОИСК АССОЦИАЦИЙ В БОЛЬШИХ ОБЪЕМАХ МАРКЕТИНГОВЫХ ДАННЫХ

Д. М. Рагель
к.э.н., доцент кафедры экономики Белорусского государственного университета информатики и радиоэлектроники

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
e-mail: ragel@bsuir.by

Аннотация. В статье изложен подход к поиску ассоциаций в корзинах товаров и поиск закономерностей в сделках. Кроме этого приводятся подходы к анализу больших данных о поведении клиентской базы, которые могут быть получены в ходе исследований рынка. В завершении описываются этапы анализа потребительского поведения и поиска в нем закономерностей.

Ключевые слова: анализ данных, маркетинговые данные, большие данные, массив данных, маркетинговые исследования, ассоциации, анализ действий, покупательская активность, классификация, кластеризация, частотная характеристика.