Analytical Technologies for Clients' Preferences Analyzing with Incomplete Data Recovering

Nataliia Kuznietsova[1]

[1] Institute for Applied System Analysis of National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
natalia-kpi@ukr.net

Abstract. The paper is devoted to new analytical information technologies for clients' preferences prediction. The problem of clients' preferences forecasting is now actual for many commercial systems, companies, banks, insurance companies and e-commerce. Various marketing efforts are used to increase demand and attract new customers. The main idea is to understand the customer needs and preferences and to model their behavior with analytical technologies. Such technologies gives the possibility of the clients' data analysis, customer demand evaluation and prediction of the next purchases. The modern approaches to clients' preferences prediction were analyzed and collaborative filtering methods were chosen. The formulation of the task modelling in terms of clients as subjects and purchases as objects was fulfilled. The method for incomplete and missing data recovering, which was proposed by the author, consists of such stages as sample incompleteness evaluation, analysis if the passes are systemic, analyzing of the passes' causes and effects with using Bayesian network, regression modelling for passes recovering. The method of implicit feedback with the combined method for incomplete and missed data processing were built in the existing modern ERP-system and gives the possibility to receive highest accuracy of clients' preferences prediction.

Keywords: Implicit feedback, Missed data, Collaborative filtering, Data Recovering, Clients' preferences.

1 Introduction

The economic growth of any country, the further development of the economy is accompanied by increasing of the incomes and profits for companies and countries' residents at the same time. There is also a significant development of the customer service sphere, increasing demand of the different products. The client-oriented strategies become the priority for customer service companies. The corporations are included in the intensive competition for clients. The key is the understanding of customer needs and preferences [1, 2], modeling of consumer behavior, and therefore, most companies and corporations invest heavily in developing their own solutions or purchasing existing ones [3,4]. Such techniques allow them to analyze customer preferences, their pre-orders and develop models that will include interesting for customers products, which can be offered to them for the next order.

2 Clients' Preferences Analysis

A foreign company specializing in the sale of media products - CDs, DVDs, as well as elite segment products has developed its own Enterprise Resources Planning (ERP) system for collecting statistical information about online store customers, catalogs of goods and knowledge bases on the previous experience. Knowledge are in form of related products recommendations and new acquisitions on the basis of the previous goods purchased in the online store. The company has more than 2 million unique customers and more than 5 million orders statistic. Various marketing efforts such as sending emails are used to increase demand among its users and to attract new customers. The formation of recommendations for related products is carried out by an analyst who selects goods by his own algorithm. Such recommendations are not really precise while people are not able to process such volumes of information. Analyst uses company's ERP system as an automated workplace. The analytics recommendations are stored in the knowledge base of the company, together with information, whether the client used this recommendation.

The company database includes user information, product information, and date of purchase. The database contains missed and lost data. The following simulation tasks are relevant for the company:

- *Forecasting:* sales estimation, server load or server downtime forecasting for providing quick user access to the order directory.
- *Analysis and risk assessment and minimization:* selecting the most promising clients for target e-mailing (the risk of choosing customer who will not buy goods. The losses are calculated as the amount spent on such ineffective mailings).
- *Providing recommendations*: identifying products which can be sold together with the high probability, creating recommendations for preferences.
- *Sequency search*: customer choice analysis while making purchases, forecasting the next possible event.
- *Grouping*: dividing customers or events into clusters of related elements, analyzing and predicting common features.

Modern approaches to analyzing and forecasting the preferences and behaviors of clients within the framework of the advisory systems are ideologically divided into the following types [2,5]:

1. Collaborative filtering methods.
2. Knowledge-based filtering.
3. Methods based on content analysis (Content-based filtering).
4. hybrid methods.

Collaborative filtering is the process of filtering information or samples by sharing multiple technologies, points of review, data sources, and more. Collaborative filtering is usually associated with very large data sets and therefore is appropriate for using in financial systems, Such systems provide financial services, process large amounts of information and combine a large number of financial data sources. In the narrower sense, collaborative filtering is one of the methods for forecast constructing in recommendation systems that uses well-known user group estimates to predict unknown user ratings [3]. The main assumption of collaborative filtration is as follow: those who have equally evaluated any objects in the past tend to give similar assessments of other subjects in the future.

3 Problem Statement

The following problem was solved: to develop the information technology for automation the process of recommendation providing on the accompanying products and next products interesting for the clients.

Let there exist a matrix R of size $u \times o$ with the subjects (clients), objects (goods), and some feedback data (previous orders). It is necessary to find a way of transforming it into one matrix with subjects and their profiles (hidden preferences) $P = (p_{tu})_{|T|\times|U|}$ and one matrix with objects and their profiles (hidden preferences that they satisfying) $Q = (q_{to})_{|T|\times|O|}$. The P and Q matrices contain scales that determine how each subject / object relates to each t. The task is to calculate P, Q in such way that their multiply approximates R as closely as possible: $R \approx P \times Q$.

In the process of iterative assignment of random values in the matrices P and Q, using the method of least squares (LS), we must arrive to the same value of the scales that most closely approximate the matrix R.

In the LS algorithm consistently, at each iteration, the following states of the system alternately change:

- P is fixed, then optimizing Q;
- Q is fixed, then P is optimizing.

This operation is continued until approximation to $R \approx P \times Q$ will be reached.

4 Criteria for Quality Assessing of the Customer Preferences' Prediction

Standard criteria for estimating the quality of forecast such as RMSE or MAE, couldn't be used to assess the accuracy of the solution to the problem of analyzing and predicting customer preferences. It is difficult to estimate if there is a mistake in forecasting model of the clients' preferences or it is the client's decision not to buy this product here and now. Maybe in the future, this customer will buy this product later. It would be advisable to organize the collection of statistical information about the fact of the reference product review, but this is also an indirect characteristic, since the client may not have enough time, but the product is interesting for him, so the recommendation is correct for this client. Such criteria for quality recommendations evaluation [5] were used:

- ✓ Accuracy: $Precision@k = \frac{\xi}{k}$, (1)

where ξ – the number of recommended objects with which the subject has an interaction (that is, the number of correctly predicted preferences); k – number of recommendations. This criterion indicates which rate of recommendations corresponds to the preferences of the subject.

- ✓ Completeness: $Recall@k = \frac{\xi}{N}$, (2)

where ξ – the number of recommended objects with which the subjects had an interaction, and N – the total number of interactions that was performed by the subjects.

Recall@k – evaluates which fraction of interactions performed by clients corresponds to the predicted interactions, i.e. how many of the forecasted picked goods were interesting to customers.

It is possible to evaluate these equation in cash equivalents, by setting the cost of each interaction and fines for lack of interaction.

5 Implicit Feedback and of Customers Preferences' Forecasting

In [6] it was proposed to introduce the following concepts:

$x_{uo} = \begin{cases} 1 & r_{uo} > 0 \\ 0 & r_{uo} = 0 \end{cases}$ – customer preferences:

$\varphi_{uo} = 1 + \alpha r_{uo}$ – level of confidence.

The confidence level is calculated by using the value r_{ui} (feedback, purchases, etc.), which gives more confidence more often, when more often subject interacts with the object. The level of confidence increases due to the linear scaling factor α (which is a "hyperparameter" model). In the confidence level, the constant 1 is always added, indicating $\varphi_{\mathrm{uo}} > 0, \forall \alpha \mathrm{r}_{\mathrm{uo}} \geq 0$ [6].

Then, the mathematical model of the task loss function is formed as:

$$\sum_{u,o} \varphi_{uo}(x_{uo} - p_{tu}^{T} q_{to})^2 + \lambda(\sum_u \|p_{tu}\|^2 + \sum_o \|q_{to}\|^2) \rightarrow min_{p_{tu}^{*} q_{to}^{*}} \quad (3)$$

The component $\lambda(\sum_u \|p_{tu}\|^2 + \sum_o \|q_{to}\|^2)$ is needed to regularize the model in such a way as to prevent retraining. The exact value of the parameter λ depends on the data and is determined by cross-validation.

The loss function contains $u \times o$ values. For typical data sets, this value can be several billions. This enormous amount of values impedes most direct methods of optimization, such as stochastic gradient descent, which is widely used for explicit data collection. Therefore, in [6] were suggested an alternative effective process of optimization. If the entities and their profiles or objects and their profiles are fixed, then the loss function becomes quadratic and can be calculated. This statement leads to the use of the method of alternating least squares [6].

5.1 Predicting Client Preferences

After calculating the preferences profiles of objects and subjects, one can recommend to a particular subject u are K available objects with the highest values of weight $x(u)_i$ – the predicted preferences of the customer u of the product o, namely:

$$\widehat{x(u)_i} = P_i Q^T. \qquad (4)$$

where $\widehat{x_{uo}}$ – symbolizes the predicted preferences of the object u of o.

5.2. The Task of Finding Related Products

Search for related products can be reformulated as a search for such products o that are similar to preferences and which they satisfy to u customers. Denote the numeric expression as sim_score:

$$\widehat{sim\_score} = QQ_i^T, \qquad (7)$$

6 Preliminary Data Preparation

A set of inputs is a statistical information about customers who have purchased certain products. The following characteristics are collected: the unique client identifier in the system (Kunden_Id), the categorical variable for clients’ gender (Geschlecht_Id - contains gaps), the name (Ort) and index (Plz) of the client's city (incomplete data), the client’ birthdate (Geburtsdatum - contains gaps), the unique product identifier (Artikelnummer), the price (Produkt_Preis) and the date of sale (Rechnungsdatum) of the product and the quantity of the product purchased (Anzahl).

That is, there are 4 characteristics with possibly incorrect or missed / lost data. In order to properly handle them, it is proposed to perform a deep analysis of the causes of the gaps’ occurrence and to use the combined method for incomplete and lost data recovering, which is proposed by the author. Method consists following steps.

1 step. Estimation of the data incompleteness of the sample for each characteristic by the criterion for estimating the number of passes.

If $I_{j(missin\,g)} > 20\%$ then, the variable-characteristic is excluded from the simulation and missing values for this characteristic do not make sense to recover.

2 step. Analysis of variables and systematic appearance of missed values

2.1. For a categorical variable assigning missed values to a separate category - filling the spaces with the value:

$V_{категор}$:=”Missing”

2.2. For all numerical variables with gaps we analyze their appearance (S-systematic):

$$S_{j_{num}} = \begin{bmatrix} 1 - \mathit{for\ systemiatic\ passes,\ where\ } I_{j(missin\,g)} \geq 5\% \\ 0 - \mathit{for\ non-systematic\ passes,\ where\ } I_{j(missin\,g)} < 5\% \end{bmatrix} .$$

3 step. Analysis of Causes and Effects.

A Bayesian network is used to establish causal relationships between variables and analyze the consequences of the missing value occurrence. Target (predicted) variable for Bayesian Network – effects.

3.1. To analyze the causes and consequences of the occurrence of passes:

$$C_j = \begin{bmatrix} 1 - \mathit{random} \\ 2 - \mathit{critical} \\ 3 - \mathit{catastrophical} \end{bmatrix}$$

3.2. If for j-th variable $S_{j_{num}} = 0$, $C_j = 1$, then all i-th missed values are replaced as:

$$v_{j_i} = \begin{bmatrix} 0 \\ as \bmod e \end{bmatrix} \text{, where } V_{j_{num}} = \begin{pmatrix} v_{j1} \\ v_{j2} \\ \\ v_{j4} \end{pmatrix} \text{— i-th value is missed.}$$

3.3. Otherwise, a regression equation is used to predict values.

4 step. Regression modeling.

For linear models the representation in the form of first order autoregression:

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k) \ , \ E[\varepsilon(k)] = 0 .$$

Then, the forecast for one step could be calculated:

$$y(k+1) = a_0 + a_1 y(k) + \varepsilon(k+1) ,$$

If the coefficients a_0, a_1 are known, then the forecast as a conditional mathematical expectation is formed as:

$$\hat{y}(k+1,k) = E_k[y(k+1)] = E_k[y(k+1) \,|\, y(k), y(k-1),..., \varepsilon(k), \varepsilon(k-1),...] =$$
$$= a_0 + a_1 E_k[y(k)] = a_0 + a_1 y(k) ,$$

For s-steps the forecast is calculated by the function:

$$\hat{y}(k+s,k) = E_S[y(k+s)] = a_0 \left(\sum_{i=0}^{S-1} a_1{}^i \right) + a_1{}^S y(k) = a_0 \sum_{i=0}^{S-1} a_1{}^i + a_1{}^S y(k) .$$

The sequence of forecasts is a convergent process if the condition $|a_1| < 1$ is fulfilled, that is: $\lim\limits_{s \to \infty} E_k[y(k+s)] = \dfrac{a_0}{1-a_1}$, $|a_1| < 1$

Extension of the forecasting function in the process of autoregression AR(p):

$$\hat{y}(k+s,k) = a_0 + \sum_{i=1}^{p} a_i \, \hat{y}(k+s-i) \ ,$$

where $\hat{y}(k+s-i) = E_k[y(k+s-i)]$.

5 steps. Recovered data application for next simulation.

Applying the proposed method to categorical variable Geschlecht_Id on 2 step, fill it out as "Missing", in this case - 0, and assume that this is "Gender not specified" category. For variables city and zip code (Ort, Plz), fill all gaps as "unknown" and delete all characters except numbers. For the birthdate variable Geburtsdatum, the direct reversals of the gap due to the regression model forecast are not considered appropriate since this characteristic is significant and the gaps can be systematic. While the "age" is usually perceived as the number of full years, it makes sense to do the following. Fill the gaps with the values that correspond to the first day of the first month for the current year. Next, we create a new "Age" variable, which is calculated as the difference between the values of the year in the Geburtsdatum variable and the value of the year for current date. This variable will be an integer and will be greater than or equal to zero. A zero will display a separate case of missed data.

In fig. 1 the visualization of this data set is presented.

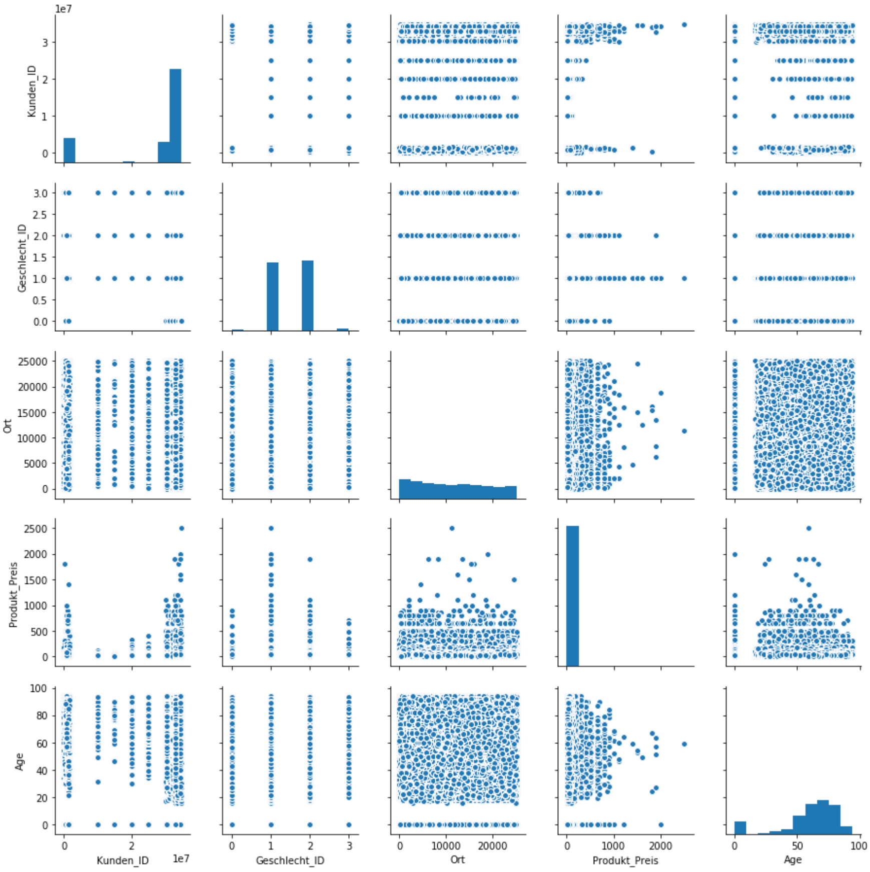

Fig.1 Dependence charts of the characteristics

Charts of dependencies between characteristics show that:

- in the customer database is almost equal quantity of women and men, with a minority of women;

- the distribution of orders among cities is close to uniform;

- the distribution of age values for clients has an average value of 61 years old.

A set of attributes for the implementation of collaborative filtering is the following: Kunden_Id, Artikelnummer, Anzahl. It is important to note that quality testing of recommendations will be completed in 2 stages:

- expert of the company sets different user IDs and subjectively analyzes the issuance of recommendations;

- if the first stage is successful, the next stage is performed, namely, the analysis of the accuracy of the model through the Precision@k and MeanAveragePrecision@k

7 Results of Collaborative Filtration Model Based on Alternating Least Squares (ALS)

Cross-validation, sometimes called cross-check, is a technique of verifying how successfully the statistical analysis by the model is able to work on an independent dataset. Usually, cross-validation is used when the purpose is foresight, and it is important to assess how prognostic model is capable for practice. Cross validation is a way to evaluate the ability of the model to work on a hypothetical test set when it is impossible to obtain such a set explicitly [7].

The model proposed in this paper has hyper parameters:

- Number of iterations - i;
- Number of hidden factors - t;
- The value of the regularization factor is λ;

In all experiments, the quality function will be Precision@k. A "grid" of parameter values is formed as: i - moves from 5 to 50 with a step of 5; t = 50; λ = 0.01.

After performing of 10 iterations, changing the value of the number of iterations, the Precision@k dependency graph is built in dependence from the value of the ALS count iteration indicator and the MeanAveragePrecision@k dependency graph from the ALS count value. By obtaining the first approximation of the optimal iterations number, it is fixed and the optimal value of the regularization coefficient is found. The next "grid" of the parameter values is: i = 10; t = 50; λ - moves from 0.01 to 1 with the step 0.01.

After performing 10 iterations, by changing the value of the regularization factor index, a plot of the dependence of the Precision@k indicator on the value of the indicator λ is built. By changing the value of the indicator, the coefficient of regularization, the dependence of the MAP@k parameter on the value of the indicator λ was constructed. By obtaining the first approximation to the optimal number of iterations and the value of the regularization coefficient, they are fixed and the optimal value of the number of hidden factors is found. The next "net" of the parameter values is formed: i = 10; t - moves from 100 to 1600 in increments of 100; λ = 0.09.

By completing 10 iterations, changing the value of the indicator, the coefficient of the number of hidden factors, we obtained a graph dependence of the Precision@k indicator from the value of t (Fig. 2).

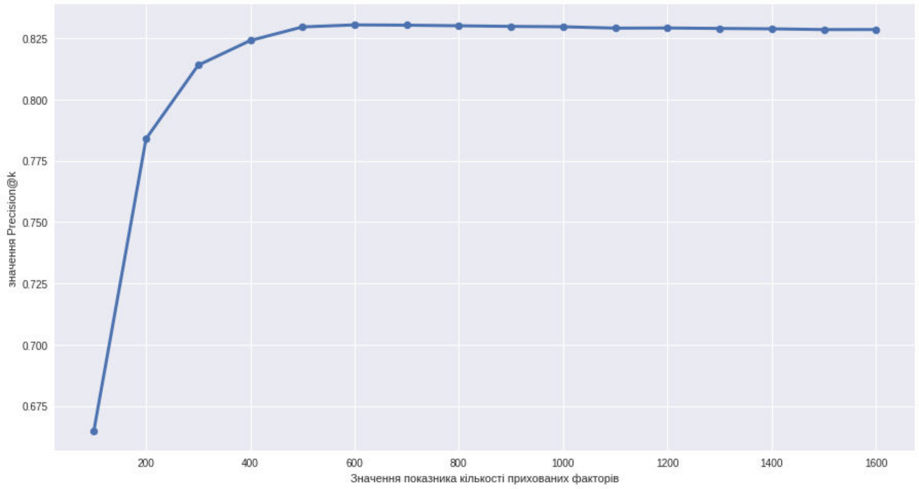

Fig. 2 The dependence of Precision @ k on the amount of latent factors

Similarly, by performing 10 iterations, changing the value of the indicator, the coefficient of the number of hidden factors, the plot of the dependence of the indicator MAP@k on the value of the indicator t (Fig. 3) is constructed.

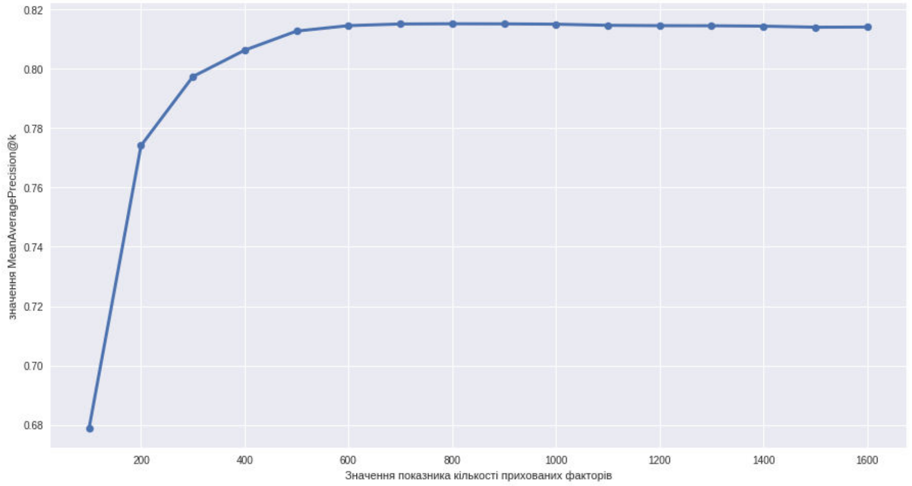

Fig. 3 The dependence of MeanAveragePrecision @ k on the amount of latent factors

8 Analysis of the Results

Both accuracy functions have a declining character, which is expected, since, with the increase in the number of optimization iterations for each of the model components, the model retraining takes place. For Precision@k, the optimal number of iterations is 15, for the statistic MeanAveragePrecision@k the number of iterations in the value of 10 is optimal. Therefore, the least of these values was selected. The optimal value for the regularization factor is 0.09 and the values of the received accuracy indicators correlate with each other.

The optimal value of the number of latent factors is 900 (Fig. 2 and 3). The dependency function is increasing to a certain level and after that the level is almost in the same range. This indicates that an optimal number of hidden factors for the given set of data was determined. The number of latent factors is the most important indicator of this system. This is confirmed by the increase in the quality predictions from 67% to 83%. It should be noted that all values of the hyperparameters of the model are relevant only for the set of data that was investigated during the experiment. For new samples, the process of analyzing data on other sets should start again from the beginning according to the algorithm described above.

9 Conclusions

Ensuring customer loyalty to the company is now a key priority in shaping the relationship between the company and its customers, providing them with high quality products and services which they need. Determining the users' needs and making recommendations when customer is choosing the product it is the main factor in the formation of business models for many companies. Development and using of recommendation systems in the e-commerce market is currently very relevant [9, 10]. The advices of such systems allow companies to use collaborative filtering levers and feature-based recommendations to better serve their customers and increase sales.

Many approaches and methods for recommendation systems constructing have been developed. Most techniques are limited by the fact that they are not able to work on such data as statistics of goods sales, etc. It is necessary to analyze the behavior of users, and for this to determine the key factors of their behavior. Since the entry into force of the "General Data Protection Regulation" Act (GDPR, the European Union), as of May 25, 2016 [8], the personal users' data should not be used without the consent of clients, and therefore such customers should be "forgotten." It turned out to be difficult to work on advisory systems for specific clients. The way of solving this problem is to reclassify clients as 1/0 (old / new client) and extract information from "implicit feedback on binary data", such as, for example, de-personalized statistics on the sale of goods, becoming relevant again. Commercial companies have statistics, therefore, even without the influence of GDPR, the task is relevant.

The methods of searching hidden factors allow solving the problem of analyzing and predicting customer preferences, so are recommended for using in modern business solutions [2, 11].

References

1. Kuznietsova N. V.: Information Technologies for Clients' Database Analysis and Behaviour Forecasting. In: CEUR Workshop. Selected Papers of the XVII International Scientific and Practical Conference on Information Technologies and Security (ITS 2017), 2017, Vol. 2067, pp. 56-62. http://ceur-ws.org/Vol-2067/, last accessed 2018/11/11.
2. Recommendation Systems. Laboratory of Mathematical Logic at PDMI RAS, https://logic.pdmi.ras.ru/~sergey/slides/N16_AIRush.pdf, last accessed 2018/11/11.
3. Data Mining Concepts. Microsoft Documentation Library Homepage, https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/data-mining-concepts?view=sql-server-2017, last accessed 2018/11/11.
4. Data Mining. Base Group Labs Homepage, https://basegroup.ru/community/articles/data-mining, last accessed 2018/11/11.
5. Воронцов К. В. Коллаборативная фильтрация: видеолекции. Школа Анализа Данных Яндекс, https://www.youtube.com/watch?v=kfhqzkcfMqI, last accessed 2018/11/11.
6. Hu Y., Koren Y., Volinsky C.: Collaborative Filtering for Implicit Feedback Datasets. In International Conference on Data Mining 2008, pp. 263-272. Eight IEEE (2008). http://yifanhu.net/PUB/cf.pdf, last accessed 2018/11/11.
7. Cross Validation. LONG/SHORT Blog, http://www.long-short.pro/post/kross-validatsiya-cross-validation-304, last accessed 2018/11/11.
8. EU General Data Protection Regulation Homepage, https://eugdpr.org/, last accessed 2018/11/11.
9. Deshpande M., Karypis G.: Item-based top-N recommendation algorithms, ACM Transactions on Information Systems, vol. 22, pp. 143-177, (2004).
10. Takacs G., Pilaszy I., Nemeth B., Tikk D.: Major Components of the Gravity Recommendation System, SIGKDD Explorations 9 , pp. 80–84, (2007).
11. Вандер П. Дж.: Python для сложных задач: наука о данных и машинное обучение. Спб.: Питер, 576 с. (2018).