On a Novel Machine Learning Based Approach to Recommender Systems

Yuri Zhuravlev, Alexander Dokukin, Oleg Senko
FRC CSC RAS
Moscow, Russia
dalex@ccas.ru, senkoov@mail.ru

Dmitry Stefanovsky
IRIAS
Moscow, Russia
dstefanovskiy@gmail.com

Ivan Saenko
Moscow State University
Moscow, Russia
i.a.saenko@mail.ru

***Abstract*—A new approach for recommender systems design is discussed. The considered system should rely only on the anonymous receipts' data and information about products currently bought by a customer. The preference rating for an arbitrary product is calculated as a classification result of a combined feature description of a product that currently is being bought and products that have been bought previously by the same customer. The corresponding product descriptions are formed by vectors of distances between the products and precalculated product clusters obtained by applying hierarchical clustering technique to binary matrix that descibes relevance between products and receipts . The proposed method was compared with two other techniques in experiments with real retail data. The first one evaluates preference rating simply as a product sales rate. The second technique uses association rules combinations. It was shown by experiments that proposed method outperforms two last ones in terms of areas under ROC curves.**

I. Introduction

Due to the rapid growth of online sales, IT businesses and traditional retailers are looking for various opportunities to improve the quality of customer service and increase revenue. Personalizing an offer is one such opportunity. Personalization is implemented by generating proposals for each client based on previously collected data [1]. These proposals are required to take into account the preferences of each client. The advantages of personalization have two aspects: on the one hand, this leads to an increase in revenue for the retailer by increasing sales as a result of providing customers with a set of products that more closely matches their preferences [1], and on the other hand, it attracts customer attention and contributes to their loyalty and satisfaction [2].

To give an example, today Amazon.com uses collaborative filtering to provide recommendations to its users [1], [3]. Also, there are many startups that invest in highly personalized online shopping technologies, including Stitch Fix, a women's clothing retailer who periodically sends boxes of five items of clothing tailored to each client's taste (for example, size, favorite brand and color, budget). Trunk Club is another similar company serving male customers. Ropazi is a text messaging service that specializes in personalized clothing offerings for children.

Online stores collect a lot of customer data (e.g. demographic, transactional, etc.). However, given the wide range of customer profiles, collecting enough transaction data for each customer profile may not be possible. This, in turn, limits the seller's ability to accurately assess preferences and offer personalized assortments.

The purpose of this article is to study the effective use of data to personalize offers to retail customers, mainly with online purchases. In general, the model of the relationship between the retailer and the buyer is as follows: the appearance of the client is considered consistent, so the seller offers each buyer a range of products, then the client decides whether to make a purchase. The retailer may encounter restrictions in terms of display or capacity, which limit the number of products in the offer. The retailer's goal is to maximize the expected total revenue for the sales season. Usually, customers are divided into different profiles according to information about past transactions (for example, purchase history, payment method). This information is exogenous and is available to the retailer after the purchase. In the current case, the client identifier: a bank card or a client card may not be present. Therefore, all identification is based on a set of goods being bought.

It is worth noting the fact that with large assortments of the same type of product, the customer's choice has a large number of degrees of freedom. However, we assume that from the point of view of the retailer, customers with a common profile are homogeneous in relation to the preferences of their products. In practice, the definition of profiles reflects the degree of customer information available to the retailer and the level of detail that allows the retailer to implement other marketing solutions.

The models are designed to understand customers based on ranking derived from noisy data about user preferences, similar, for example, to a web search, where the user's choices allow enriching the query keywords with the user information [4], [5]. It can be assumed that users are not alike, but have somewhat similar behavior with a certain preferences inherent only to them. Therefore, it is possible to build a population model that explains the similarities and diversity. Segmentation or clustering is the key to effective personalization and identifying preferences. It should be noted that the machine learning approach for personifying offers has long attracted the attention of researchers and it is worth noting the Mallows models [6] and the Plackett-Luce model [7].

Our approach is motivated by the interest of retailers in identifying segments of customers with similar preferences and in offering personalized assortments for their customers. Unlike most of the existing works on offline settings (using

historical data), we offer an algorithm for personalizing the offer, which can be implemented in real time. In addition, the proposed clustering policy is quite general and flexible, since the number of clusters (segments) is endogenous and does not require a preliminary determination.

We illustrate the practical value of a clustering policy in real conditions using a dataset from a major Russian retailer. The data set consists of roughly ten thousand cosmetic and related goods purchased in different combinations in about one hundred thousand transactions over two months period. We compare the effectiveness of the proposed policy with a data-intensive policy that ignores any potential similarity of preferences in different profiles and, thus, evaluates the product preferences for each profile individually.

Let's describe the problem and the methods involved more formally. Recommender systems are a popular tool aimed to give a customer an advice which good in the best way corresponds to his/her demands [8]. Many techniques can be used to implement a recommender system. Context based systems use some supplementary information about customers or goods. However such information is often hard to achieve. Another approach employs information about customer's preferences expressed by them in one way or another, for example, in form of product rating at the shop's site. In the latter case some very efficient mathematical tools involving matrix decomposition can be used. But getting client's preference data is associated with additional costs in offline shopping. Finally, recommendations can be based on digital traces left by anonymous customers, i.e. the set of customers' receipts registered up to a certain point. That is the system considered in the present research.

Each receipt Z can be described as a binary vector of length N corresponding to a total number of products sold by a certain retail organization. The vector's elements corresponding to products sold in that particular transaction are marked as ones, and the rest are marked zeros. New customer's recommendations are made on the basis of the previously collected receipt data and the currently performed transaction Z_0 which can be described in the same binary form.

New customer's preference ratings can be achieved in many ways. The simplest method could rely solely on the products frequencies over all known transactions. Obviously such a recommendation would not employ the Z_0 information. There are other popular methods which lack that drawback, for example association rules [9], [10], clustering [11] based techniques or collaborative filtering [12], [13].

Association rule is a way of measuring consequence like relationships between objects. In this case the relation between a product X being bought, i.e. $Z_0(X) = 1$, and a product Y to be recommended. The relation is characterized by two criteria, namely support and confidence. Let's denote the set of all known receipts or digital traces by S, and by $S(D)$ its subset containing products from and arbitrary set D. Thus, $S(\{X\})$ and $S(\{Y\})$ will stand for subsets of S containing product X or Y correspondingly, whereas $S(\{X, Y\})$ will denote a subset containing both goods. The support of X is then defined as

$$Sup(X) = \frac{|S(\{X\})|}{|S|},$$

whereas confidence of the X, Y pair is

$$Conf(X, Y) = \frac{|S(\{X, Y\})|}{|S(\{X\})|}.$$

Pairs of objects with large enough values of both criteria form association rules which can be used to estimate probability of buying product Y subject to product X purchase. The conclusion may be made based on a single best association rule or by their ensemble.

When using clustering techniques the set of binary receipt vectors is divided into several groups or clusters in which the digital traces are considered close to each other in terms of a selected metric. Then the Y product's preference rating can be calculated by frequencies present in the cluster containing the Z_0 trace.

Finally, collaborative filtering is based on selecting a subset of k vectors $Z_1, \ldots, Z_k$ closest to Z_0. The preference rating is then formed by frequency of product Y in the group $Z_1, \ldots, Z_k$ obtained the described way.

The methods mentioned above are quite simple heuristics which might be outperformed by the modern machine learning techniques [17] which do nicely in various applied tasks. Those are neural networks, gradient boosting decision tree ensembles [18], support vector machines, etc. Weak use of the mentioned methods in recommender systems design can be attributes to the lack of a direct reduction of the recommender problem to a standard classification one in the described vector space. A novel object description is proposed in the following section.

II. Clustering Based Feature Descriptions

Clustering methods are used in recommendation systems to select groups of customers with similar preference profiles [14]. Here we suggest another technique where clustering is used to select groups of complementary products. The derived set of clusters is further used to generate multidimentional feature description of products. Such descriptions allow effective application of machine learning tools.Agglomerative hierarchical cluster analysis is preferable due to the fact that the number of resulting clusters can be regulated by choosing a threshold for the similarity between two clusters that are being merged. A preliminary number of clusters is not required. Moreover, the size distribution of clusters is effectively controlled by setting the metric between the clusters.

The authors of the present research have already shown that agglomerative hierarchical grouping method applied to the described binary data produces well interpreted set of product clusters [15]. The description of an arbitrary product X can be formed by its similarity estimates to the products of different clusters by the nature of interest of various customers in them. The similarity measure can be expressed in form of binary metrics using the fact that each product X can be described in binary form as $G(X) = [Z_1(X), \ldots, Z_M(X)]$, where M is the power of the set of receipts available for training S_T. Evidently, the use of the best known metrics such as Hamming or Euclidian distances is inappropriate for the description of shopping activity similarity. Those metrics basically count the number of matches between components and hence the absence of a product in two receipts is equally encouraged as their presence. Thus, higher similarity will be estimated

for products that are present in receipts with small number of goods. The binary metrics that are widely used in biology, ecology and other domains are preferable. The Jaccard index and chi-squared metrics were tried in our previous research. The chi-squared metric [16] has been proved advantageous [15]. So, this metrics is used in current research. Let's describe it formally.

Let's consider two arbitrary binary vectors x, y of the same length. Let's denote by a the number of positions in which both vectors are ones, i.e. xy^T. Let's define b, c, d in the similar fashion, i.e. $b = x(\bar{1}-y)^T$, $c = (\bar{1}-x)y^T$, $d = (\bar{1}-x)(\bar{1}-y)^T$, where $\bar{1}$ is a vector of ones of the same length as both x and y. The chi-squared metric ρ is then defined as

$$\rho(x,y) = \frac{(ab-cd)^2 sign(ab-cd)}{\sqrt{(a+b)(b+c)(c+d)(d+a)}}.$$

Similarity between two binary vectors x and y is proportional to so defined $\rho(x,y)$. Let we have L non-intersecting clusters $C_1,\dots,C_L$ in the S_T set. The distance of product Y to the i-th cluster is calculated as

$$P(Y,C_i) = \frac{1}{|C_i|}\sum_{X\in C_i}\rho(G(Y),G(X)).$$

Vector $P(Y) = [P(Y,C_1),\dots,P(Y,C_L)]$ can serve as a good feature description of the product Y since it is continuous and it reflects customer's interest to the product in terms of his interest to different clusters of products.

Let we now want to assess product preferences for customers who bought a particular set of goods, i.e. who received a particular receipt, and describe those preferences by vectors of similar length. Let's consider an arbitrary receipt Z with a single limitation: it has to consist of more than r products. Let $X_1,\dots,X_r$ be its top r goods in order of their frequency decrease in S_T. We deem that preference information of another arbitrary product Y is contained to a large degree in the combined vector $[P(X_1),\dots,P(X_r),P(Y)]$ and that it can be extracted with machine learning techniques.

In the present research the underlying clustering was performed with hierarchical grouping method. Starting with the initial set of clusters equal to separate goods two closest clusters were joined then on each step until a given distance threshold was reached. The following function was used as a similarity metric between clusters C_i and C_j. Let

$$\rho_{UA}(C_i,C_j) = \frac{1}{|C_i||C_j|}\sum_{X\in C_i}\sum_{Y\in C_j}\rho(G(X),G(Y))$$

$$\rho_{CL}(C_i,C_j) = \min_{X\in C_i,\,Y\in C_j}\rho(G(X),G(Y)).$$

Using $\rho_{UA}(C_i,C_j)$ implements unweighted average linkage clustering, while using $\rho_{UA}(C_i,C_j)$ corresponds to complete linkage clustering that prevent merging big clusters. Thus combining of two metrics allows to control distribution of clusters by size. It was found that good results may be achieved when simple sum of two similarity metrics is used:

$$\rho(C_i,C_j)\rho_{UA}(C_i,C_j) + \rho_{CL}(C_i,C_j).$$

Clusterization results. Agglomerative hierarchical clusterization technique using $\rho(C_i,C_j)$ proximity metrics was applied to dataset including 98500 receipts and about 6000 goods. As a result about 270 clusters with size of at least 5 products were outlined. In most cases clusters includes products of similar assignment. At that many products inside each cluster belong to the same brand. Such clusterization results were evaluated as reasonable by experts. However clusterization validity may be really correctly assessed when it is used inside recommender system topredict customers preferences.

To validate algorithms' performance the available set of receipts S was divided into training and control parts S_T and S_C, respectively. The training set was then used to calculate product frequencies, pairwise distances and clustering. Further it was used to generate predictive algorithm with the help of machine learning techniques. The control part was used solely to estimate algorithms' performance as described later. In the final experiments the division was made randomly in proportion 7 training receipts to 3 control ones.

III. Recommender Algorithms

A. Predicting models

Our goal was to compare efficiency of three different techniques aimed to evaluate products preference ratings. As we initially don't have any information about buyers' preferences except the same receipt data, we base our evaluation on those receipts. For that the receipts from the control set were reduced to top r goods from each of them, i.e. the r products with the highest sales rate or highest support in other words. Then the ratings of all other products were estimated including both goods which were actually bought in the real transaction and goods which were not. The prediction made with the calculated preference ratings was then compared to the real receipt to compare methods' performances. The described approach has an obvious drawback as we can't be sure that the considered receipts contain all the relevant goods and only them. But given the amount of receipts and the lack of alternatives we assume that the described approach reflects method's performance well enough.

The considered techniques are: machine learning algorithm A_{ML}, associative rules algorithm A_{AR} and frequency based algorithm A_F. Of those the first two really base their preference rating estimation for an arbitrary product Y on the top r products from the receipt. A_{ML} uses previously achieved clustering of goods to generate their informative multivariate descriptions. Then the combined description of the top r goods and Y is processed by the trained recognition algorithm implementing some machine learning method. The possibility of application of machine learning methods in this task comes from their two stage nature. The so called margins or probability estimates generated at the first stage of recognition can be treated as preference ratings in this particular task. The second technique A_{AR} is based on associative rules relating evaluated good Y to top r products. A_F, on the contrary, makes constant prediction independent of the receipt considered. Preference rating of Y is estimated simply as Y's sales rate.

Let's consider the algorithms in more details starting with A_{ML}. After performing all the preliminary steps including product clustering the training set S_T is filtered to leave only receipts containing at least $r+k_1$ goods. k_1 here reflects the

number of positive examples which can be derived from each receipt. This subset of S_T will be further referred to as $S_T^{r+k_1}$. Then each receipt from $S_T^{r+k_1}$ was used to generate $k_1 + k_2$ training objects containing feature descriptions of the top r goods from the receipt and the description of one product other than top r ones.

More formally, let Z_j be an arbitrary receipt from $S_T^{r+k_1}$. Let $X_1^j, \ldots, X_r^j$ be the top r goods from the receipt in the order of decrease of their frequency in the S_T. Products $Y_1^j, \ldots, Y_{k_1}^j, Y_{k_1+1}^j, \ldots, Y_{k_1+k_2}^j$ are chosen randomly from the remaining $N-r$ ones $\{X_1, \ldots, X_N\} \setminus \{X_1^j, \ldots, X_r^j\}$ such that $Y_1^j, \ldots, Y_{k_1}^j$ are present in the considered receipt and the rest are not. Combined feature descriptions of the top r objects and an additional one form the training objects derived from the receipt Z_j: $[P(X_1^j), \ldots, P(X_r^j), P(Y_k^j)]$, $k = 1, \ldots, k_1 + k_2$. The target value of the said objects is determined by the presence of the corresponding Y product in the receipt, i.e. it is equal 1 if $Y_k^j \in Z_j$ and 0 otherwise.

The combined training sample was then used to train different two-class machine learning methods [17] A_{ML} including logistic regression, support vector machines, decision forests with gradient boosting [18]. Their performance was estimated with the multifold cross-validation with the class 1 gap as preference rating for a product. The best model, which turned out to be gradient boosting decision forest, was selected to use in the final algorithm.

B. Steps of proposed algorithms

Thus training procedure for A_{ML} algorithm may be summarized as follows.

- Step 1. Hierarchical grouping method is used to receive optimal clustering of products $\{C_1, \ldots, C_l\}$ by full initial training set S_T.
- Step 2. Set $S_T^{r+k_1}$ is selected from initial training set S_T. Each receipt from S^{r+k_1} includes more than $r + k_1$ products. Let $X_1, ..., X_r$ be the products with the greatest sales rate.
- Step 3. Combined training set is generated from $S_T^{r+k_1}$.
  - Step 3 (a). Set $\{X_1 \ldots, X_r\}$ and one product Y that does not belong to set $\{X_1 \ldots, X_r\}$. At that Y may be in receipt Z or may be not in Z. Label 1 is assigned to product in case when product Y was really bought and label 0 is assigned otherwise.
  - Step 3 (b).For each object from training set S_T descriptions of products $X_1, \ldots, X_r, Y$ are calculated as vectors of distances to clusters $C_1, \ldots, C_l$.
  - Step 3 (c). Vector description of each objects is calculated as concatenation of products $X_1, \ldots, X_r, Y$ vector descriptions.
- Step 4. Machine learning method is used to generate A_{ML} from training set S_T version that was received at steps 2-5.

To evaluate preference rating for some good Y from new receipt Z it is necessary to take top goods from Z and to calculate combined description according steps 3(a) and 3(b). Then preference rating is calculated by combined description with the help of algorithm A_{ML}.

Let consider in more details heuristics A_{AR}. Support and confidence are calculated for all products from the training set S_T and their pairs respectively. The associative rules preference ratings for the owner of Z_j receipt can then be calculated as

$$A_{AR}(Y, Z_j) = \frac{1}{|\{X_1^j, \ldots, X_r^j \mid Sup(X_i^j) > 0\}|} \times \times \sum_{X \in \{X_1^j, \ldots, X_r^j \mid Sup(X_i^j) > 0\}} Conf(X_i^j, Y)\,.$$

To implement calculation of preference ratings for good Y by top r products with the help of algorithm A_{AR} it is necessary

- to find all possible associative rules at training stage,
- to select all associative rules that are associated with Y and one of the top products.

The frequency based reference rating calculation involves only the support value, i.e.

$$A_F(Y, Z_j) = Sup(Y)\,.$$

IV. Experiments

A. Performance validation

Performance validation of algorithms A_{ML}, A_{AR}, A_F was done with the same set of objects. So, the sample structure was mainly defined by the requirements of A_{ML}. That is the control set of objects S_C was filtered to leave only receipts containing $r + 1$ goods or more. Each receipt was then used to generate $k_3 = 1000$ testing objects containing feature descriptions of its top r goods in combinations with k_3 other objects chosen randomly independent of their presence in the receipt. The correspondent target value was again determined by the presence of the respective additional product in the receipt. Each testing object's preference rate was estimated by the three algorithms and the resulting values were used to compare methods with ROC analysis [19].

The initial set of about one hundred thousand receipts generated in some time interval was provided by a distribution network. 68955 of them were used for calculating frequencies and association rules and the rest 29553 were used for control. After that two tests were run with r, i.e. the number of products required to be in the receipts, being 3 and 4. After the filtering the control set reduced to 10010 and 7152 receipts respectively, generating 1000 times as much testing objects. The respective ROC curves are shown in the figures 1 and 2. In the legend "boostings", "sum" and "const" stay for A_{ML}, A_{AR} and A_F respectively. Both pictures show advantage of the proposed method, slightly increasing with r.

It is seen from plots (1) and (2) that a curve breaks exist for FNR about 15.5% and at interval $(0.155, 1)$ ROC curve for algorithm AR reduces to linear dependence. Such effect is

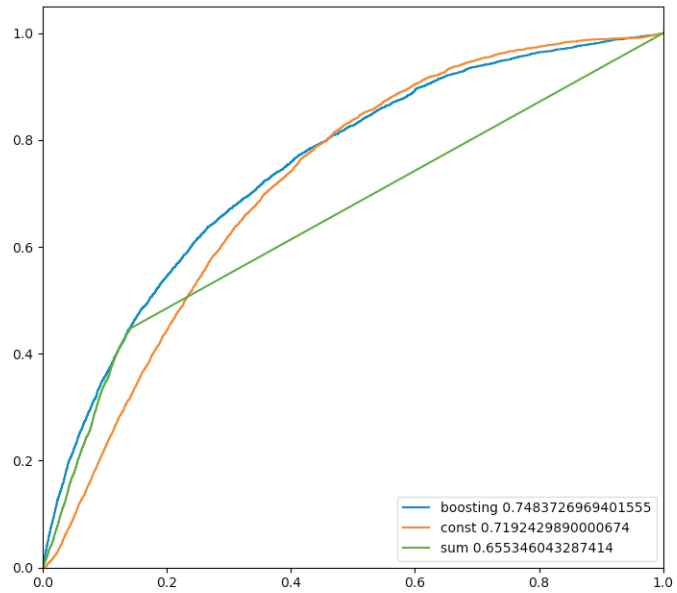

Fig. 1. ROC curve comparison for $r = 3$

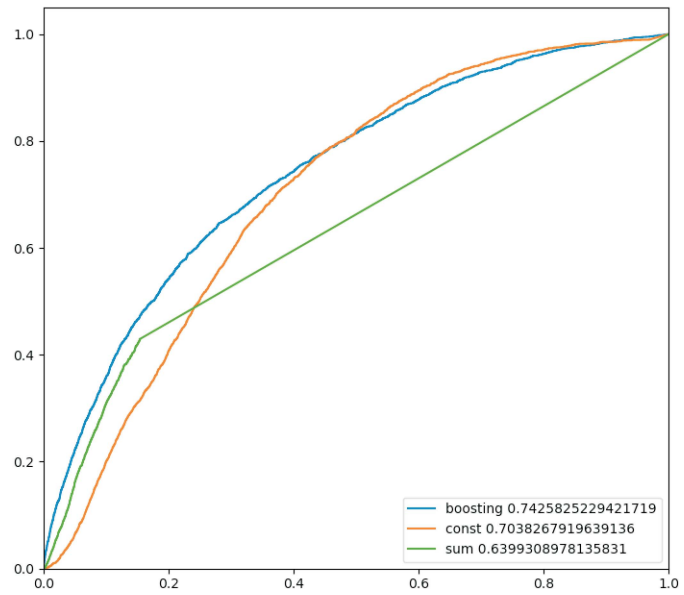

Fig. 2. ROC curve comparison for $r = 4$

related to absence of associative rules with $Conf(X_j, Y) > 0$ for 82% of evaluated products.

B. Mutual relationships between ratings

It is interesting to know if preference ratings calculated by different techniques are correlated. Existence of a strong correlation may indicate that it is sufficient to use a technique with the best performance and it is not necessary to employ another technique or their combinations. On the contrary a weak correlation or absence of correlation makes it useful to use combined solutions.

Relationship between preference ratings calculated with the help of different methods are given at figures 3-5. Scatter plot (3) represent relationship between preference ratings calculated by algorithm A_F and algorithm A_{AR}. Existence of strong linear correlation is seen for majority of observations. However for relatively small group of cases preference ratings calculated by A_{AR} are significantly higher than preference ratings calculated by algorithm A_F.

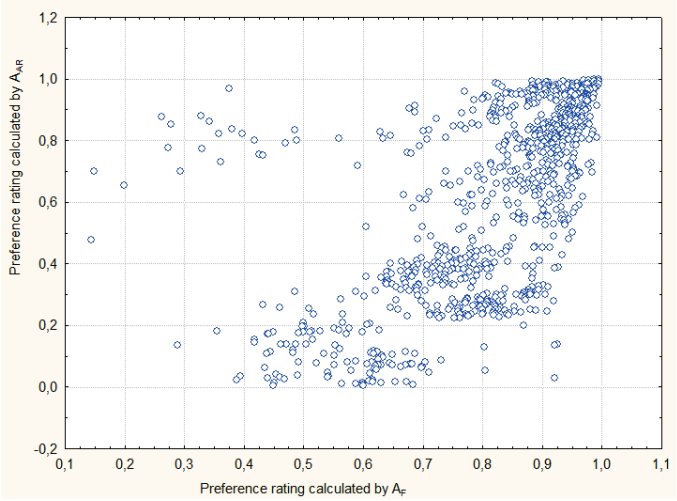

Fig. 3. Relationship between A_{AR} and A_F

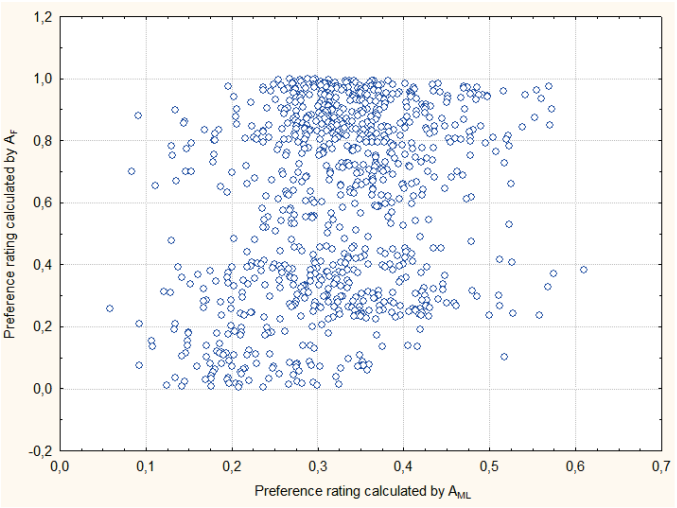

Fig. 4. Relationship between A_{ML} and A_{AR}

It is from scatter plot 4 that practically there is no correlation between ratings calculated by A_{AR} and by A_{ML}. It is seen from plot (5) that correlation beween A_{ML} and A_F is weak. Pearson correlation coefficient between preference ratings calculated by A_F and A_{AR} is equal 0.603,correlation coefficient between preference ratings calculated by A_{ML} and A_{AR} is equal 0.27, correlation coefficient between preference ratings calculated by A_{ML} and A_{AR} is equal 0.495.

V. Conclusion

A new method has been developed for estimating customer preferences by the anonymous cash receipts data. The method involves several steps. At the first stage, a system of clusters

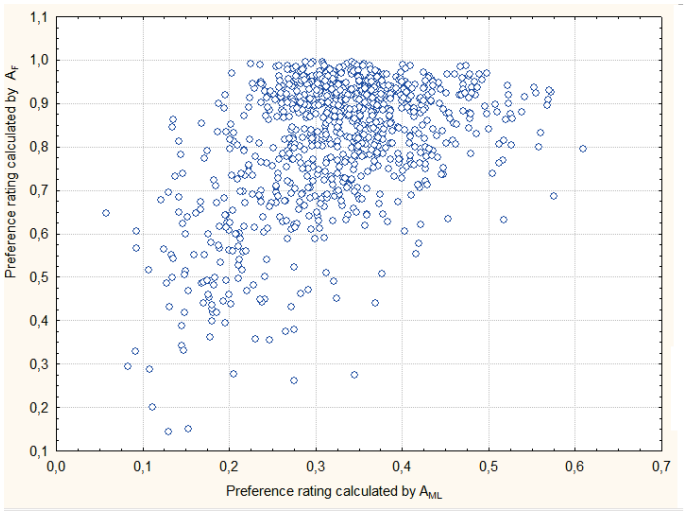

Fig. 5. Relationship between A_{ML} and A_F

of related products is constructed using hierarchical clustering methods. At the second stage, multidimensional feature descriptions of goods are formed, which are defined as a set of distances of products to clusters. To rank the products a customer is not currently buying the description of each of them combined with the description of several products he is buying is processed by a previously trained machine learning algorithm. The achieved estimates are used to sort the products and recommend the most promising ones. Training is carried out using a sample that includes the same descriptions. The sample includes both cases of purchased and non-purchased goods. The real data experiments were conducted with different numbers of products form each receipt being selected as the description base. In all cases the most frequently purchased products were selected as such. The experiments indicate the prospects of the proposed approach.

Firstly, the experiments showed that in this setting the effectiveness of the method based on the use of the machine learning method turned out to be slightly higher than the effectiveness of methods based on assessing preferences only by the frequency of goods' purchase, as well as the method based on ensembles of associative rules. Evaluation was performed by ROC AUC.

Secondly, It is important to mention that though the ROC AUC values of the proposed method the frequency based algorithm are quite close the experiments showed significant differences between their recommendations in terms of goods. The machine learning algorithm suggests rarer products which may be advantageous for the shop owner. Also, the correlation value indecate that method ensembles might provide some improvement in future research.

In general, this article discusses a large amount of data on the sales of the Russian retailer, with a large assortment of products, as well as the algorithm that for several purchased products, customers assign them some profiles. Customer preferences are unknown to the seller and should be studied over time. To do this, it is proposed to distinguish the core of the clusters that describe a particular type of consumption, and as soon as the client purchases 3-4 products from one core, the rest of the core products can be offered to him, and then products in the same cluster but not from the core. Such products can be called peripheral products. Peripheral products can show the nearest core, and thus you can diagnose and show a possible change in the type of consumption. This will allow you to adaptively adjust the composition of customer segments based on purchase information.

Thus, it is possible to form a recommender which, for several products, makes a conclusion about the type of client and possible changes. It is traditionally believed that as the number of customer profiles increases, marginal profitability from the information itself and its consolidation from different points of sale decreases, that is, the data does not allow revealing structures. Our approach related to the cluster core and its periphery does not work with clusters as a whole, but with a structure consisting of a core and a periphery, which in turn can be associated with another core. Thus, we can work together and separately with the nuclei and peripherals, which is another area for future research.

ACKNOWLEDGMENT

This work was supported in part by the Russian Foundation for Basic Research, projects no. 18-29-03151, 18-01-00557.

REFERENCES

[1] Ghose A., Arora N., et al. "Putting one-to-one marketing to work: Personalization, customization, and choice." *Marketing Letters*, vol.19, no.3, Feb.2008, pp. 305-321.

[2] Ansari A., Mela C.F. "E-customization." *Journal of marketing research*, vol.40, no.2, 2003, pp. 131-145.

[3] Linden G., Smith B., York J. "Amazon.com recommendations: Item-to-item collaborative filtering." *IEEE Internet computing*, vol.7, no.1, 2003, pp. 76-80.

[4] Burges C.J.C. "From RankNet to LambdaRank to LambdaMART: An Overview." *Microsoft Research Technical Report MSR-TR-2010-82*, Jun.2010.

[5] Young P. "Optimal voting rules." *Journal of Economic Perspectives*, vol.9, no.1, 1995, pp. 51-64.

[6] Mallows C.L. "Non-null ranking models. I." *Biometrika*, vol.44, no.1/2, Jun.1957, pp. 114-130.

[7] Plackett R.L. "The analysis of permutations." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol.24, no.2, 1975, pp. 193-202.

[8] H. Sohlberg, "Recommending new items to customers - a comparison between Collaborative Filtering and Association Rule Mining. Master's Thesis.". Stockholm: KTH Royal institute of technology.school of computer science and communication (CSC), 2015.

[9] W. Li, J. Han, J. Pei, "CMAR: accurate and efficient classification based on multiple class-association rules", *Proceedings of the 1st IEEE International Conference on Data Mining (ICDM '01)*, Dec.2001, pp. 369–376.

[10] X. Sun, F. Kong, H. Chen, "Using Quantitative Association Rules in Collaborative Filtering". *In: Fan W., Wu Z., Yang J. (eds) Advances in Web-Age Information Management. WAIM 2005. Lecture Notes in Computer Science*, vol.3739, 2005.

[11] G. Dunn and B. Everitt, *An Introduction to Mathematical Taxonomy*. Cambridge: Cambridge University Press, 1982.

[12] F. Cacheda, V. Carneiro, D. Fernandez, V. Formoso, "Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems", *ACM Transactions on the Web (TWEB)*, vol.5, no.1, Feb.2011, 2.

[13] Jin Hyun Jooa, Sang Won Bangb, Geun Duk Parka, "Implementation of a Recommendation System using Association Rules and Collaborative Filtering", *Procedia Computer Science*, vol.91, 2016, pp. 944-952.

[14] J.D. West, I. Wesley-Smith, C.T. Bergstrom, “A recommendation system based on hierarchical clustering of an article-level citation network”, *IEEE Transactions on Big Data*, vol.2, no.2, 2020, pp. 113-123.

[15] Yu. Zhuravlev, A. Dokukin A., O. Senko, D. Stefanovskiy, “Use of Clasterization Technique to Highlight Groups of Related Goods by Digital Traces in Retail Trade”, *Proceedings of 9th International Conference on Advanced Computer Information Technologies ACIT-2019*, Jun.2019, pp. 84-88.

[16] S.-S. Choi, S.-H. Cha, C. C. Tappert, “A survey of binary similarity and distance measures”, *Journal of Systemics, Cybernetics and Informatics*, vol.8, no.1, 2010, pp. 43-48.

[17] T. Hastie, R. Tibshirani, J. Friedman, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction.” Springer Science & Business Media, 2013.

[18] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine”, *The Annals of Statistics*, vol.29, no.5, 2001.

[19] T. Fawcett, “An introduction to ROC analysis”, *Pattern Recognition Letters*, vol.27, 2006, pp. 861–874.