Analysis of the preferences of public transport passengers in the task of building a personalized recommender system

A A Borodinov[1], **V V Myasnikov**[1,2]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
[2]Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 443001

e-mail: aaborodinov@yandex.ru

Abstract. The paper presents the theoretical and algorithmic aspects for making a personalized recommender system (mobile service) designed for public route transport users. The main focus is on identifying and formalizing the concept of "user preferences", which is the basis of modern personalized recommender systems. Informal (verbal) and formal (mathematical) formulations of the corresponding problems of determining "user preferences" in a specific spatial-temporal context are presented: the preferred stops definition and the preferred "transport correspondence" definition. The first task can be represented as a well-known classification problem. Thus, it can be formulated and solved using well-known pattern recognition and machine learning methods. The second is reduced to the construction of dynamic graphs series. The experiments were conducted on data from the mobile application "Pribyvalka-63". The application is the tosamara.ru service part, currently used to inform Samara residents about the public transport movement.

1. Introduction

The amount of heterogeneous data characterizing the transport situation in the city has increased due to the widespread and active use of modern electronic communications systems, global navigation systems, and various active and passive sensors. Such information is used in navigation and reference systems (services) quite widely [1]. However, along with the development of services and their popularization, the expectations and demands of users and the amount of information that has to be taken into account when planning movements are growing. User demand is individualized from the classic tasks of searching for the “shortest path” [2] or getting a “forecast of arrival at a stop of public transport” [3, 4], shifting expectations from services to Intelligent personal assistants. Although the final decision or choice in such systems remains with the person, the options of the solutions they offer are significantly dependent on the scenario conditions of the request as well as on previous actions and decisions of the user [5, 6]. Accounting for all of the indicated factors is possible in “self-tuning” systems for individual user preferences based on machine learning methods [7]. The imperfection of existing algorithms and the lack of significant experience in the use of machine learning methods in such systems prevents the rapid emergence of such services.

Multimodal routing is determined by the possibility of using several modes of transport in one trip. The analysis of modern literature devoted to recommender systems of multimodal routing [5, 8, 9] allows us to identify some major problems:

– The cold start problem is a well-known and well-researched problem for recommender systems [8, 10]: it is essential to achieve a balance between the accuracy of the recommended routes from system initialization. Thus, the allowable setting time for a personal preference profile should be small.

– The receiving information method from the user is not formalized [11, 12].

– Individual characteristics such as personal income, age, gender, family size, access to public transport influence the choice of the route even for the same purpose of the trip [13].

– User preferences change over time. In addition, context influences user selection [14, 15].

– Typical existing solutions mainly use the Bayesian approach with a sequential parameter recalculation scheme [5, 16].

– It is possible to use transfer learning to improve recommendations [17].

– The problem of determining traffic flow on the vehicle route [18].

This article proposes one of the possible ways of describing and solving the problem of determining individual preferences of users of public route transport and creating a personalized recommender system. The system uses user interaction data with the mobile service as part of the problem of creating a personalized recommender system. The second section of the work formalizes the basic concepts and introduces the basic notation for all objects of interaction. The third section describes the information arising from the public transport user interaction with the mobile application "Pribyvalka-63". Mobile application and service tosamara.ru, are currently used to inform Samara residents about public transport movement and its arrival at the stop. Also in this section are presented the variants of unformalized (verbally described) definitions of "user preferences", suitable for further consideration. The fourth section presents the mathematical formulations of problems, as well as methods and algorithms. Finally, the fifth section presents the results of experimental studies on real data obtained using the mobile application "Pribyvalka-63".

2. Definitions and designations

Let S be the set of public transport stops. Let for each stop $s \in \mathrm{S}$ defined spatial (geographical) coordinates $\mathbf{x}_s \equiv (x_s, y_s, z_s)$ and some unique stop identifier, denoted by *ID*(*s*). Without loss of generality, we can assume that the set S is ordered (for example, by *ID*(*s*)): $S = \{s_1, s_2, \ldots, s_{|S|}\}$.

Let the value *d* determine the calendar date, the value *t* - the time of day, and $w(d) \in W$ - the day of the week, taking values from the set:

$$\begin{aligned} &\mathrm{W}=\mathrm{W}_0 \cup \mathrm{W}_1, \\ &\mathrm{W}_0 \equiv \{MON, TUE, WEN, THU, FRI\}, \quad \mathrm{W}_1 \equiv \{SAT, SUN\}. \end{aligned} \tag{1}$$

Let V determine the set of public transport vehicles, each $v \in \mathrm{V}$ is characterized by the type

$$type(v) \in \{BUS, TRAM, TROL, MARS\}, \tag{2}$$

and has a unique identifier *ID*(*v*) (in practice, the unique identifier may coincide with the vehicle state registration number). For each vehicle at any time, we consider its spatial coordinates to be determined:

$$\mathbf{x}(v,d,t) \equiv \big(x(v,d,t), y(v,d,t), z(v,d,t)\big). \tag{3}$$

Denote the routes set of public transport objects as M. In addition, each route $m \in \mathrm{M}$ is characterized by five arguments:

$$m \equiv \Big(ID(m), N(m), \mathbf{s}(m), N^*(m), \mathbf{x}(m)\Big), \tag{4}$$

where $ID(m)$ - route identifier (in practice, the route number), $N(m)$ - stops number in the route, and $\mathbf{s}(m)$ - stops sequence in an amount $N(m)$ of form:

$$\mathbf{s}(m) = \left(s_1^m, s_2^m, \ldots, s_{N(m)}^m\right), \tag{5}$$

where $s_n^m \in \mathrm{S} \quad \left(m \in \mathrm{M}, n \in \overline{1, N(m)}\right)$. Let $\mathrm{S}(m) \equiv \left\{s_i^m\right\}_{i=\overline{1,N(m)}} \subseteq \mathrm{S}$ be the stops set of the corresponding route, $ind(s,m)$ - stop index s of route m, viz. $ind\left(s_n^m, m\right) = n$. In case $s \notin \mathrm{S}(m)$ the corresponding index is assumed to be "indefinite", and denoted as Δ: $ind(s,m) = \Delta$. Denote the routes set passing through one or a couple stops as follows:

$$\mathrm{M}(s) \equiv \left\{m \in \mathrm{M} : s \in \mathrm{S}(m)\right\}, \ \mathrm{M}(s1, s2) \equiv \left\{m \in \mathrm{M} : s1 \in \mathrm{S}(m) \wedge s2 \in \mathrm{S}(m)\right\}. \tag{6}$$

More detailed information about the route geometry is represented by a pair $N^*(m), \mathbf{x}(m)$, where the first value determines the number of nodes of the polyline describing the route, and the second is the vector defining the coordinates of these nodes:

$$\mathbf{x}(m) \equiv \left(\mathbf{x}_1^m, \mathbf{x}_2^m, \ldots, \mathbf{x}_{N^*(m)}^m\right). \tag{7}$$

For convenience, we will call the pair (m,k), $\left(k = \overline{1, K(d,m)}\right)$ route implementations (RI) on the appropriate day *d*.

Additionally, we denote $t(d,m,k,s)$ - vehicle arrival time assigned to RI (*m*,*k*) on day *d*, to stop $s \in \mathrm{S}(m)$ (in case $s \notin \mathrm{S}(m)$ we consider the time value as uncertain).

Denote the vehicle assigned to RI *(m,k)* on day *d*, as $v(d,m,k) \in \mathrm{V} \quad \left(k = \overline{1, K(d,m)}\right)$.

In addition to the vehicles, pedestrians and passengers, considered in the paper as the users of transport services, are participants in the traffic. Denote by U the set of users, and we will characterize each specific user $u \in \mathrm{U}$ with a unique identifier *ID*(*u*) (mobile device id or hash code) and spatial coordinates at a specific point in time d, t:

$$\mathbf{x}(u,d,t) \equiv \left(x(u,d,t), y(u,d,t), z(u,d,t)\right). \tag{8}$$

If there is no information about user coordinates, we consider that they have "undefined value" Δ (all three at the same time).

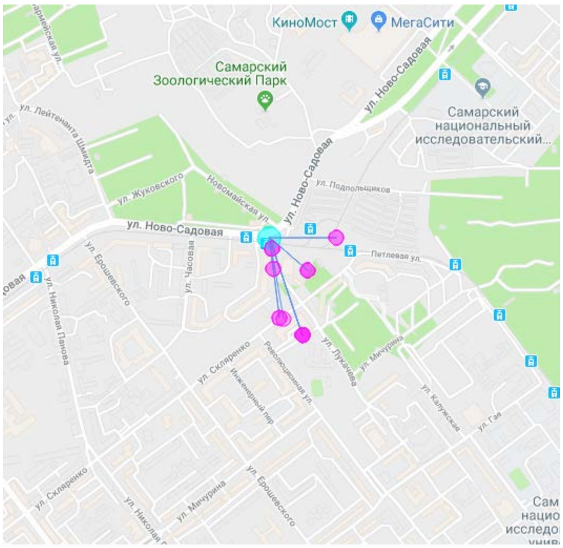

Figure 1. Blue circle – stop location, purple circles – user location points at request time.

3. Mobile application and support service data, «informal preferences» options

For the mobile service "Pribyvalka-63" data for analysis are presented as follows:

- stop data (identifiers and coordinates);
- route information (identifiers and stop identifier list);

- data on the vehicle (identifiers), location coordinates (with a frequency of 2 times per minute), the destination to routes;

- coordinates of users and request parameters are recorded during requests (request results are not saved, since they can be restored from vehicle traffic data) in the form: $ID(s),d,t,\quad ID(u),\mathbf{x}(u,d,t);$

- user response to the request is not saved.

Based on the presented data, the following two options "user preferences" seem appropriate (we consider the user known):

- user-preferred stops at specific space-time coordinates (Figure 1);

- user-preferred "transport correspondence", also considered in the space-time context. "Transport correspondence" refers to the actual movement from one stop to another, the route chosen and the route vehicle type. Information about the "starting" and "end" stops of a particular user is optional (derived) information (Figure 2).

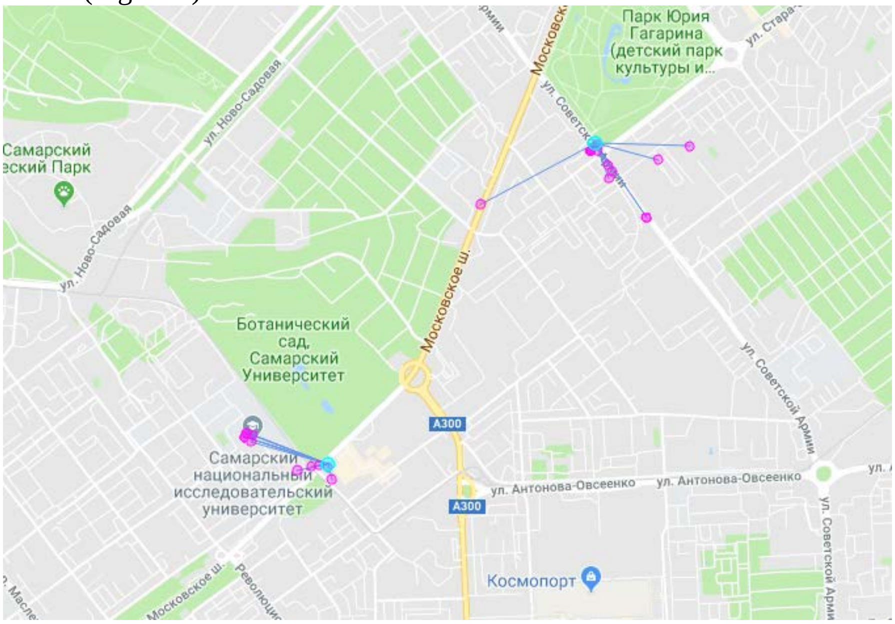

Figure 2. "Start" and "End" stops of a specific user.

4. Methods

4.1 *User-preferred stops*

The task of determining "user-preferred stops" when it is in certain space-time coordinates can be formalized as follows.

Let given precedent set (training examples) for a specific user. Each precedent is a set of space-time context vector-description and the corresponding "answer" (in our case, the stop identifier). It is necessary for the newly received space-time context vector-description of the new situation (which may be absent in the list of precedent examples), to indicate the most "suitable (s) / relevant (s)" answer (s), if necessary, ordering them by degree relevance.

For the case of a single issued response, the described task is a well-known classification theory [19, 20] or machine learning [7, 21], where, according to the object/situation description, the system should indicate the object/situation class as a feature vector. Any recognition algorithm can be represented as two consecutive operators. The first operator translates the description of the object into a numerical value characterizing the "degree of membership" to the class. And the second, according to the indicated value, refers to a specific class [22]. In statistical methods of recognition, the posterior probability [19] is used as a numerical value, in algebraic [23] - estimates, in neural network - the output of the last layer of neurons, etc. Denote the specified numeric value $\Gamma(\mathit{features};\mathit{class})$.

Thus, the formal formulation of the problem for a specific user $u \in \mathrm{U}$ can be represented as follows (where $z \equiv ID(s)$).

Given:

a) precedent set in the form $\left\{\mathbf{x}_i, d_i, t_i; z_i\right\}_{i \in \Im}$. (feature vector; answer)

b) feature vector of the new situation $\mathbf{x}, d, t$.

It is necessary: for the specified vector $\mathbf{x}, d, t$ to determine the permutation of objects $\sigma : \mathbb{N}_{|\mathrm{S}|} \to \mathbb{N}_{|\mathrm{S}|}$ from an ordered set (stops) $\mathrm{S} = \left\{s_1, s_2, \ldots, s_{|\mathrm{S}|}\right\}$ such that

$$\Gamma\left(\mathbf{x}, d, t; ID\left(s_{\sigma(1)}\right)\right) \geq \Gamma\left(\mathbf{x}, d, t; ID\left(s_{\sigma(2)}\right)\right) \geq \ldots \geq \Gamma\left(\mathbf{x}, d, t; ID\left(s_{\sigma(|\mathrm{S}|)}\right)\right). \quad (9)$$

The result for the user is an ordered stops list:

$$s_{\sigma(1)}, s_{\sigma(2)}, \ldots, s_{\sigma(|\mathrm{S}|)}. \quad (10)$$

The formal quality measure of the final decision:

$$\Gamma_\Sigma = \sum_{i=1}^{|\mathrm{S}|} \frac{1}{i} \Gamma\left(\mathbf{x}, d, t; ID\left(s_{\sigma(i)}\right)\right). \quad (11)$$

Solution:

Theory of pattern recognition and machine learning offers a variety of methods for solving the problem. In this paper, we use an approach based on the calculating estimates idea proposed by Yu. I. Zhuravlev [23] and nonparametric estimation of Parzen probability density [19]. Set the value $\Gamma(\,features; class)$, characterizing feature vector belongs to a class

$$\Gamma(\mathbf{x}, d, t; z) = \sum_{i \in \Im} \mu\left(\mathbf{x}, d, t; \mathbf{x}_i, d_i, t_i\right) I\left(z_i = z\right), \quad (12)$$

where

$$\mu\left(\mathbf{x}, d, t; \mathbf{x}_i, d_i, t_i\right) = I\left(\begin{gathered}\left(w(d) \in W_0 \wedge w\left(d_i\right) \in W_0\right) \vee \\ \left(w(d) \in W_1 \wedge w\left(d_i\right) \in W_1\right)\end{gathered}\right) \cdot \exp\left(-\alpha\left|t - t_i\right|\right) \cdot \exp\left(-\beta\left\|\mathbf{x} - \mathbf{x}_i\right\|\right). \quad (13)$$

Event indicator:

$$I(a) = \begin{cases} 1, & a = true; \\ 0, & a = \mathit{false}. \end{cases} \quad (14)$$

The values $\alpha, \beta \in \mathbb{R}_+$ - some coefficients, $\left|t - t_i\right|$ - a numerical value (for example, the number of seconds), which characterizes the difference between t, t_i. As a result, the algorithm for solving the problem of determining "user-preferred stops" will be as follows:

Step 1. For all stops from the set S calculate the values (12):

$$\Gamma\left(\mathbf{x}, d, t; ID\left(s_i\right)\right), \quad i = \overline{1, |\mathrm{S}|}. \quad (15)$$

Step 2. The values set obtained in (15) is ordered in descending order — a permutation is formed $\sigma : \mathbb{N}_{|\mathrm{S}|} \to \mathbb{N}_{|\mathrm{S}|}$ (9).

The resulting permutation is the solution of the problem. An ordered list of stops is provided to the user (10).

«Cold Start» Solution:

To solve the "cold start" problem, the precedent set as follows supplements the initially empty precedent set:

$$\left\{\left(\mathbf{x}_i, d0, t0; ID\left(s_i\right)\right)\right\} \cup \left\{\left(\mathbf{x}_i, d1, t0; ID\left(s_i\right)\right)\right\}, \quad i = \overline{1, |\mathrm{S}|}, \quad (16)$$

where *t*0="0h00min", *d*0 and *d*1 are the dates, respectively, of the weekend and working days preceding the system launch date, and $\mathbf{x}_i \left(i = \overline{1, |\mathrm{S}|}\right)$ - stop coordinates $s_i \left(i = \overline{1, |\mathrm{S}|}\right)$. Analysis of expressions (12) - (13) shows that with such "starting" data, the contribution of the time component $\exp\left(-\alpha\left|t-t_i\right|\right)$ in expression (12) will be the same, and the differences in values $\Gamma\left(\ldots\right)$ will be completely determined by differences in Euclidean distances from the point $\mathbf{x}$ to the stops coordinates $\mathbf{x}_i$ $\left(i=\overline{1,|\mathrm{S}|}\right)$. Thus, the value $\Gamma\left(\mathbf{x},\ldots;ID\left(s\right)\right)$ will be more significant when s closer to $\mathbf{x}$.

4.2 User-preferred "transport correspondence"

The task of determining the user-preferred "transport correspondence" can be presented as the task of estimating the probability characteristics (relative frequency) of correspondences. That is, the movements from the stop *s*1 to the stop *s*2, in the space-time context. The following values are important (all characteristics are related to the behavior of a particular user u):

- $p_u\left(t\left|s1,s2,m,\mathrm{W}_a\right.\right)$ $\left(m\in\mathrm{M}\left(s1,s2\right),\quad a=\overline{0,1}\right)$ - correspondence time distribution density *s*1→*s*2 with the route choice *m* on the weekday W_a; the density function corresponds to the “boarding time” on the route vehicle, and is indicated to stop *s*1;
- $p_u\left(t\left|s1,s2,\mathrm{W}_a\right.\right)$ - correspondence time distribution density *s*1→*s*2 on the weekday W_a;
- $P_u\left(s1,s2\left|\mathrm{W}_a\right.\right)$ - correspondence probability *s*1→*s*2 on the weekday W_a;
- $P_u\left(m\left|s1,s2,\mathrm{W}_a\right.\right)$ - probability of choosing the route *m* for implementing the correspondence *s*1→*s*2 on the weekday W_a;
- $P_u\left(m\left|\mathrm{W}_a\right.\right)$ - probability of choosing the route *m* for implementing the correspondence on the weekday W_a;
- $P_u^*\left(s\mid\mathrm{W}_a\right)$ - the probability that the stop s is the “end/start”.

Additional information about user behavior can be obtained from additional data:

- $p_u\left(\rho\mid\mathrm{W}_a\right)$ - distances distribution that the user is able to overcome without using route vehicles;
- $p_u\left(\tau\left|\rho,\mathrm{W}_a\right.\right)$ - time distribution that the user spends in overcoming the corresponding distance.

All specified values can be calculated on potential user correspondences data collected by the recommender system:

$$\left\{s_i^{start},s_i^{end},m_i^j,k_i^j,t\left(d,m_i^j,k_i^j,s_i^{start}\right),\tau_i^*,\sigma_i^*\right\}_{i\in I_d} \qquad (17)$$

for each day *d* and user. Where s_i^{start},s_i^{end} - correspondence data, information about the route m_i^j,k_i^j $t\left(d,m_i^j,k_i^j,s_i^{start}\right)$ - vehicle arrival time RI at the stop $s_i^{start},\tau_i^*,\sigma_i^*$ - mean and standard deviation of potential boarding on the vehicle.

5. Experiments

The presented method software implementation was written in Python. The results were visualized based on Google Maps. The experiments used "Pribyvalka-63" mobile application and tosamara.ru service data.

The database obtained during the experiments, contains information about requests 57190 users. Each user is represented by a unique identifier *ID*(*u*), which is defined by the device ID hash code and is impersonal. The database contains a total of 4103161 user requests for an arrival forecast at a public transport stop. From 1478 stops of the tosamara.ru service, users made requests to 1417 stops.

For the experiments, we selected common user requests that represent the average user of the service. Maps with different parameters $\alpha,\beta\in\mathbb{R}_+$ and request time were built to visualize the results of the proposed approach. The color of the area on the map corresponds to the first stop from the ordered list (10). An example of determining the preferred stop for the user is shown in Figure 3.

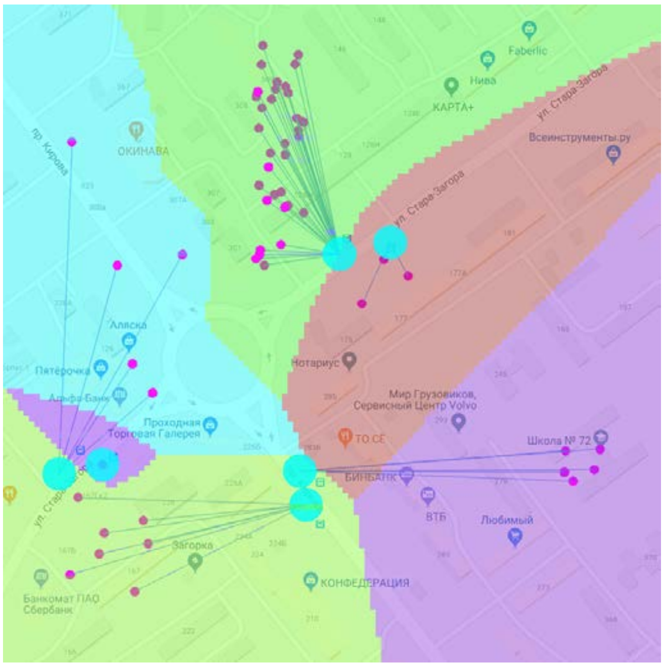

Figure 3. Preferred stops map depending on user location.

Leave-One-Out cross-validation was applied to obtain statistical indicators characterizing the quality of the proposed algorithm. The partitions number $C^1_{|\Im|}$ of the user requests set in this case is equal to $|\Im|$, and the classification accuracy is calculated as follows:

$$Accuracy = \frac{1}{|\Im|}\sum_{i\in\Im} I\left(z_i \equiv ID(s_{\sigma_i(1)})\right)\cdot 100\% \qquad (18)$$

where the indicator $I(a)$ corresponds to (14).

Also, another method was implemented for comparison with the proposed algorithm, in which the user was offered the nearest stop, without taking into account previous requests. The proposed algorithm accuracy was 93%, for the nearest stop algorithm - 65%.

6. Conclusion

In this paper, we presented the informal and mathematical problem formulations of defining user preferences. Users take public transport route in a personalized recommendation system task. We showed the results of an experimental study. An algorithm was developed using pattern recognition and machine learning methods. The determining the user preferred transport correspondence task was formalized and an approach to its solution was specified.

7. References

[1] Chorus C G, Molin E J E and Van Wee B 2006 Use and effects of Advanced Traveller Information Services (ATIS): A review of the literature *Transport Reviews* **26** 127-149

[2] Agafonov A A and Myasnikov V V 2018 Numerical route reservation method in the geoinformatic task of autonomous vehicle routing *Computer Optics* **42(5)** 912-920 DOI: 10.18287/2412-6179-2018-42-5-912-920

[3] Agafonov A A, Yumaganov A S and Myasnikov V V 2018 Big data analysis in a geoinformatic problem of short-term traffic flow forecasting based on a K nearest neighbors method *Computer Optics* **42(6)** 1101-1111 DOI: 10.18287/2412-6179-2018-42-6-1101-1111

[4] Agafonov A and Myasnikov V 2015 Traffic flow forecasting algorithm based on combination of adaptive elementary predictors *Communications in Computer and Information Science* **542** 163-174

[5] Arentze T A 2013 Adaptive personalized travel information systems: A bayesian method to learn users' personal preferences in multimodal transport networks *IEEE Transactions on Intelligent Transportation Systems* **14** 1957-1966

[6] Nuzzolo A, Crisalli U, Comi A and Rosati L 2015 Individual behavioural models for personal transit pre-trip planners *Transportation Research Procedia* 5 30-43

[7] Portugal I, Alencar P and Cowan D 2018 The use of machine learning algorithms in recommender systems: A systematic review *Expert Systems with Applications* **97** 205-2 27

[8] Campigotto P, Rudloff C, Leodolter M and Bauer D 2017 Personalized and Situation-Aware Multimodal Route Recommendations: The FAVOUR Algorithm *IEEE Transactions on Intelligent Transportation Systems* **18** 92-102

[9] Eiter T, Krennwallner T, Prandtstetter M, Rudloff C, Schneider P and Straub M 2016 Semantically Enriched Multi-Modal Routing *International Journal of Intelligent Transportation Systems Research* **14** 20-35

[10] Mikic Fonte F A, López M R, Burguillo J C, Peleteiro A and Barragáns Martínez A B 2013 A Tagging Recommender Service for Mobile Terminals *Information and Communication Technologies in Tourism* (Springer Berlin Heidelberg) 424-435

[11] G. March J 1978 Bounded Rationality, Ambiguity, and the Engineering of Choice *Bell Journal of Economics* **9** 587-608

[12] Campigotto P and Passerini A 2010 Adapting to a realistic decision maker: Experiments towards a reactive multi-objective optimizer *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **6073 LNCS** 338-341

[13] Zhang J and Arentze T 2013 Design and implementation of a daily activity scheduler in the context of a personal travel information system *Lecture Notes in Geoinformation and Cartography* 407-433

[14] Braunhofer M and Ricci F 2017 Selective contextual information acquisition in travel recommender systems *Information Technology and Tourism* **17** 5-29

[15] Braunhofer M, Elahi M and Ricci F 2015 User Personality and the New User Problem in a Context-Aware Point of Interest Recommender System *Information and Communication Technologies in Tourism* (Springer International Publishing) 537-549

[16] Guo S and Sanner S 2010 Real-time multiattribute Bayesian preference elicitation with pairwise comparison queries *Journal of Machine Learning Research* **9** 289-296

[17] Pan S J and Yang Q 2010 A Survey on Transfer Learning *IEEE Transactions on Knowledge and Data Engineering* **22** 1345-1359

[18] Myasnikov V V 2012 Method for detection of vehicles in digital aerial and space remote sensed images *Computer Optics* **36** 429-438

[19] Fukunaga K 1990 *Introduction to statistical pattern recognition* (San Diego: Academic Press)

[20] Vorontsov K V *Machine Learning* URL: http://www.machinelearning.ru/wiki/ (date of the application 10.11.18)

[21] Bishop C M 2006 *Pattern Recognition and Machine Learning* (Springer)

[22] Zhuravlev Yu I and Nikiforov V V 1971 Recognition Algorithms Based on Estimation Calculation *Cybernetics* **3**

[23] Zhuravlev Yu I and Gurevich I B 1989 Pattern recognition and image recognition *Recognition, classification, forecast. Mathematical methods and their application* **2**

Acknowledgments
The work was funded by the Ministry of Science and Higher Education of the Russian Federation (unique project identifier RFMEFI57518X0177).