Hybrid method for prediction of users’ information behavior in the Internet based on bioinspired search

V Bova¹, Yu Kravchenko, S Rodzin, E Kuliev

Southern Federal University, 105/42 Bolshaya Sadovaya Str., Rostov-on-Don, 344006, Russia

¹E-mail: vvbova@yandex.ru

Abstract. The paper presents the method of predicting the learner preferences to make recommendations for building the behavior profile for recommender systems (assistants) personalizing the educational activity in the Internet space. To form the personal information offer and to identify the groups of collaborative filtration of the users with similar preferences, the authors propose the hybrid model of collaborative filtration based on the Item-Item CF and User-User CF methods simultaneously. To reduce the size of space during the search process, the author present a novel heuristic bioinspired algorithm based on the fish school behavior in nature. This method can be characterized by the ability to scale and process the sparse data. To estimate the quality of the developed algorithm, experimental research was carried out on the basis of the benchmark dataset MovieLens.

1. Introduction

In terms of the process of moving the learning and self-educational activity into the virtual space, such factors as development of human-internet interaction technologies or ambiguity of the virtual community impact on personality can provide comfortable conditions for self-education and self-development. However, the mentioned factors can produce the risks of deviation of the of informational socialization from the normative behavior [1, 2]. This phenomenon is called victim behavior [3]. The risks of the learner’s victim behavior in the virtual informational and educational environment (IEE) are related with information richness of the Internet space with various content including questionable-quality one, and the absence of pedagogic scenario of the informal behavior of the learners [4].

Victim impact is specifically personal and depends on the ability of a person to choose information suitable for the interests, principles and values independently. In that connection, the main task of this paper is to personalize the learning content providing automated processing of data on the users’ activity and making recommendations to represent the learner’s personal preferences in terms of organization of the effective and psychologically safe information and education activity considering the exponential growth of Internet resources on different topics. To create the personal informational offer and identify groups of the users with similar preferences, we propose the hybrid model of collaborative filtering (CF) [5, 6]. To reduce the size of space of the preference search from the big implicit data on the users’ activity and to provide the best recommendation, the paper presents a Fish School Search bioinspired algorithm, able to scale and process the sparse data.
2. Formal statement of the problem

One of the methods of solving the prediction task is transforming the personal recommendations to the group ones on the basis of CF of the users with the similar preferences. The initial data can be represented in a following way: let \( I = \{i_1, i_2, ..., i_n\} \) be a set of the objects (Internet resources), \( U = \{u_1, u_2, ..., u_n\} \) be a set of the users, \( u_d \) be the current user for which we form the recommendations; \( R = \{(u, i)|u \in U, i \in I, r_{ui} \text{ – is known}\} \subseteq U \times I \) be a set of the pairs including the users and the objects, for which we know the rating of the preferences.

Then \( \{(u_1, i_1, r_1), ..., (u_n, i_n, r_n)\} \) is a matrix of the activity including the information on the users’ behavior. Each record \( ((u_i, t, r_i)) \) means that the user \( u_i \) has evaluated the piece of content \( t_i \) with the value \( r_i \). The task is to build a model \( \hat{r}_{ui} \) for predicting the preference rating for a random pair \( (u, i) \in U \times I \) on the basis of the training set \( \{R, \{r_{ui}\}(u, i) \in R\} \).

Let us consider an iterative process of collective training of the model of preference evaluation predicting. The parameters for model training are specified in a following way: \( \hat{r}_{ui}^{(k)} \) and \( \tilde{r}_{ui}^{(k)} \) denote the evaluation of the object \( i \in I \) by the user \( u \in U \) determined by the individual and collective predictors built at the iteration \( k \); \( \hat{r}^{(k)} \) and \( \tilde{r}^{(k)} \) denote the models of the individual and collective predictors of the object \( r_{ui}, (u,i) \in R \) in re-weighed \( \tilde{r}^{(k)} \) at the iteration \( k-\text{th} \) on the training set. The formula for interrelation between the collective and individual predictors can be demonstrated as follows:

\[
\hat{r}_{ui}^{(k)} = \frac{1}{\sum_{l=1}^{k} \gamma_l} \sum_{l=1}^{k} \gamma_l \tilde{r}_{ui}^{(l)},
\]

where \( T = \{ t_1, t_2, ..., t_k \} \) and \( L = \{ l_1, l_2, ..., l_k \} \) denote the sets of models of the collective and individual predictors respectively, \( \gamma \) denote the weight of the individual predictor in a collective.

3. Hybrid method of the collaborative filtering

To solve the problem of adapting and personalizing the learning content to a certain user, it is necessary to consider uncertainty and spontaneity of the individual’s behavior in terms of a particular Internet resource [6]. Generally, the information on the users’ activity is collected implicitly and has the following properties: large amount, heterogeneity and rapid update of the data over time [6-8]. To find recommendations for a certain user on the basis of behavior profile and identifying groups of objects with the similar characteristic, we propose a model-based method of increasing the information pertinency using the combination of the Item-Item CF and User-User CF methods [6]. CF of a set of preferences on the basis of Item-Item CF and User-User CF methods can be considered as preliminary stage of processing of the data on users’ activity (Figure 1).

To evaluate the degree of the users’ preferences similarity and build the behavior profiles, the paper proposes weighing evaluations of the preferences and calculating the distances between the elements from the user activity matrix \( R = U \times I \) (Figure 2).

---

**Figure 1. Hybrid method of CF**

**Figure 2. Matrix of factorization**
We need to find two matrices, $U \in \mathbb{R}^d$, $I \in \mathbb{R}^{n \times k}$, making the $Y \sim U I$, $Y \in \mathbb{R}^{m \times m}$ matrix factorization-based CF is a factorization model of $Y$, user $i$ is represented by $u_i \in \mathbb{R}^d$ and items $j$ is represented by $i_j \in \mathbb{R}^d$.

In the sparse matrix $R$ each element $r_{ui}$ denotes a vector of hidden (latent) user preferences and Internet resources respectively, $d$ is the dimension of the feature space of the user preferences $n_u$ and object rating $m_i$. Thus, the predicted value is calculated by a following formula $\hat{r}_{ui} = u_i^T I_j$ and $r_{ui} = (u_i, i_j)$. The evaluation of predicting is related with the large amount of the learning model parameters, having $(n_u+m_i) \times d$ of free parameters, which should be restricted in a certain way. Let us introduce the regulating parameter $\lambda$ to minimize the learning error:

$$Y(\langle u, i \rangle) = \min_{u, i} \sum_{i \in R}(\hat{r}_{ui} - r_{ui})^2 + \lambda (\|u_i\|^2 + \|I_j\|^2),$$

where $\hat{r}_{ui}$ and $r_{ui}$ denote the predicted and observed ratings of the preferences of the user $u$ and element $i$ respectively. To improve the effectiveness of the preference prediction, let us apply the weight parameter based on the weight calculating (degree of belief):

$$\omega_{ui,j} = 1 + \eta (r_{ui,j} - \hat{r}_{ui,j}),$$

where $\eta$ denotes the smoothing coefficient. In order to find the optimal $U, I$, we can solve the following optimization problem $f(x) = \min Y(\langle u, i \rangle)$:

$$\min_{u \in \mathbb{R}^{n \times d}, I \in \mathbb{R}^{m \times k}} \sum_{i \in R} \omega_{ui,j}(\hat{r}_{ui,j} - r_{ui,j})^2 + \lambda (\|u_i\|^2 + \|I_j\|^2).$$

The first step is building a model on the basis of the training dataset. The first classifier is trained on all data with equal weight coefficients [9-11]. At each iteration the weights are assigned according to the accuracy of classification. The weights of correctly classified data decrease, the weights of incorrectly classified data increase. At iteration $k$ the dataset weights are updated depending on the errors obtained by the collective predictor at the iteration $(k-1)$. The next step is training the individual predictor on the re-weighed dataset and assigning the weights for newly built individual predictor. To estimate the accuracy of the prediction, we use the Root Mean Square Error:

$$\text{RSME} = \frac{1}{k} \sum_{i \in R} \omega_{ui,j}(\hat{r}_{ui,j} - r_{ui,j})^2 \to \min_{f}.$$

The best model is selected on the basis of the heuristic algorithm Fish School Search (FSS) [12].

### 4. Fish School Search for bioinspired algorithm

The model parameters are optimized with the use of metaheuristic FSS algorithm. In terms of the FSS algorithm each specie is considered as a reactive agent with its own memory to store the information on its weight and the best position (current decision vector) in the search space (feasible region). This feature allows us to avoid the need to find and fix the global best decisions [12-14].

The diagram of the FSS algorithm is demonstrated in Figure 3. The amount of the search agent ($r_{ui}$) $N$ and their positions $x_i, (u, i) \in U \times I$ are set randomly while building the factorization matrix $Y \sim UI$. The initial value of the agent’s weight $\omega_{ui,j}$ is calculated according to formula (1). The fish’s weight is calculated as a difference between the fitness values at current and previous iterations [13].

The next position $n_i$ is calculated considering the distribution probability $\rho$ in the interval [-1,1]: $n_i(t) = x_i(t) + \rho \cdot S_{mov}(t)$.

Individual swimming produces a new position for the further research. The migration step is calculated as: $S_{mov}(t+1) = S_{mov}(t) - (S_i - S_f) / t_{max}$, where $S_{mov}$ denotes movement step, which is a value, evenly distributed in the interval $[S_i=0; S_f = t_{max}]$, the maximum number of the algorithm iteration.

The feeding operator updates the weight $\omega_{ui,j}$ of each fish at each iteration. It is calculated on the basis of the fitness value considering the change of the fitness function of the current and next position $\Delta f(t)$ according to: $f(t) = [f[n_i(t)] - f[x(t)]$. The weight of new position is calculated as follows: $\omega(t+1) = \omega(t) + \Delta f(t) / \max[|\Delta f(t)|]$. In terms of optimization, this stage is used to check if the decision point is included in the feasible region and if the fitness function $f(x)$ is better at this point.

Collective instinctive swimming is implemented after all the agents has finished their individual movements [14]. The movement shift is determined by the following equation:
\[ \gamma(t) = \frac{\sum_{i=1}^{N} \Delta x_i(t) \Delta f_i(t)}{\sum_{i=1}^{N} \Delta f_i(t)}, \]

where \( \Delta x_i(t) = n_i(t) - x_i(t) \) denotes the individual movement of a fish during the iteration [13]. After the migration stage the positions are updated as follows: \( x_i(t+1) = x_i(t) + \gamma(t) \).

**Figure 3.** Diagram of the FSS algorithm

The stage of collective volitive swimming allows us to adjust and shift the fish position in direction of \( B(t) \) (current center of gravity of the population). If the fish school gains weight (the successful search), all agents of the population are shifted towards \( B(t) \) of the whole aggregation [12]. Otherwise, the agents are shifted away from \( B(t) \). \( B(t) \) is calculated as follows:

\[ B(t) = \frac{\sum_{i=1}^{N} n_i(t) x_i(t)}{\sum_{i=1}^{N} n_i(t)}. \]

Collective volitive movement is calculated according to the following formula:

\[ x_i(t+1) = x_i(t) + V_{vol} \text{rand}[0,1] \cdot (x_i(t) - B(t)), \]

where \( V_{vol} \) denotes a random value determining the step size of the agent’s movement.

When all agents are analyzed, the next step is to calculate the RSME, apply the new positions to the agents, and calculate the fitness function. The stop criterion of the algorithm is a free parameter of the iteration number \( t \). The effectiveness of the FSS algorithm is investigated by the task of preferences evaluations optimization. The criteria are the speed and accuracy of the algorithm.
5. Experimental research

To conduct the experiments, we used the real data from the benchmark dataset MovieLens 100K. It consists of 100,000 evaluations of 1,700 Internet resources from 1,000 users. The evaluations are graded from 0.5 to 5 in increments of 0.5. The initial data set is divided into 5 equal parts where we take 5 pairs of the training and testing dataset (in the ration of 80% to 20%). To estimate the effectiveness of the FSS algorithm, we compared it with the Alternating Least Squares (ALS) and Stochastic Gradient Descent (SGD) algorithms [6, 8] in terms of cross-validation on the partitions according to the external parameters demonstrated in Table 1. The internal parameters are estimated on the basis of the training datasets. The accuracy of the final learning model is determined on the basis of validation dataset.

Table 1. External parameters used for estimation of the RSME accuracy

| Parameter | Lambda | Latent factors |
|-----------|--------|----------------|
| Algorithm | 0.01   | 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 10 15 20 25 30 35 40 45 50 |
| SGD       | 1.25   | 1.15 1.08 1.06 1.03 1.02 1.01 0.95 0.93 0.92 0.91 0.9 0.89 0.88 0.91 0.95 0.96 |
| ALS       | 1.33   | 1.15 1.08 1.04 1.02 1.01 0.98 0.94 0.93 0.92 0.91 0.9 0.89 0.88 0.87 0.86 0.85 0.84 0.83 0.82 |
| FSS       | 0.62   | 0.61 0.57 0.49 0.48 0.47 0.46 0.45 0.44 0.43 0.42 0.41 0.4 0.39 0.38 0.37 0.36 0.35 0.34 0.33 0.32 |

Figure 4. Performance according to the parameter Lambda Vs. RMSE

Figure 5. Performance according to the parameter number of latent factors Vs. RMSE
The results are demonstrated in Figures 4, 5. With increase of external parameters Latent factors and Lambda, the FSS algorithm demonstrates lower values of the error than the SGD and ALS algorithms. The increase of Lambda allows us to adjust the model to the hidden data and to improve the ability to generalize.

The diagram in Figure 6 shows that the RSME is mostly invariant to increasing amount of iterations for all algorithms. However, the RMSE value is the best for FSS algorithm and decreases with the increasing amount of iteration. The execution time of the heuristic FSS algorithm is greater than the ALS, but it decreases with the increasing amount of iteration (Figure 7).

6. Conclusion
The paper deals with the personalization of the learning content allowing us to increase the relevance of selecting the learning objects according to personal characteristics, interests and performance of the users of the IEE. The scientific and practical novelty of the proposed method includes the idea of combined use of the methods Item-Item CF and User-User CF, which provides solving the problem of ‘cold start’ and improving the quality of recommendations for the users with low activity and weakly expressed preferences. To increase the accuracy of the preference prediction, we developed the heuristic FSS algorithm for training the individual and collective predictors of building the predicting models. The results of the experiments have confirmed high ability to generalize and accuracy of prediction.
7. Acknowledgments
The reported study was funded by RFBR according to research project № 18-29-22019.

References
[1] Tingting Z, Chen L Y, Liang-Hsien T 2015 Understanding user motivation for evaluating online content: a self-determination theory perspective Behaviour and Information Technology 34 479-491
[2] Lin H F 2009 Examining of cognitive absorption influencing the intention to use a virtual community Behaviour and Information Technology 28 421-431
[3] Deliang W, Lingling X, Chuan C H 2015 Understanding the continuance use of social network sites: a computer self-efficacy perspective Behaviour and Information Technology 34 204-216
[4] Wilson P M, Mack D E, Grattan K P 2008 Understanding motivation for exercise: A self-determination theory perspective Canadian Psychology 49 250–256
[5] Shambour Q, Lu J 2012 A trust-semantic fusion-based recommendation approach for e-business applications Decision Support Systems 54 768-780
[6] Guo G, Zhang J, Thalmann D 2014 Merging trust in collaborative filtering to alleviate data sparsity and cold start Knowledge-Based Systems 57 57-68
[7] Tuzhilin A, Adomavicius G 2005 Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions Knowledge and Data Engineering 17 734
[8] Desrosiers C, Karypis G 2011 A comprehensive survey of neighborhood-based recommendation methods Recommender Systems Handbook ed F Ricci L Rokach (New York: Springer) chapter 4 pp 107–144
[9] Lops P, De Gemmis M, Semeraro G 2011 Content-based recommender systems: state of the art and trends Recommender Systems Handbook ed F Ricci L Rokach (New York: Springer) chapter 3 pp. 73-105
[10] Bova V V, Nuzhnov E V, Kureichik V V 2017 The combined method of semantic similarity estimation of problem oriented knowledge on the basis of evolutionary procedures Advances in Intelligent Systems and Computing 573 74-83
[11] Kravchenko Y, Kursitys I, Bova V 2017 The development of genetic algorithm for semantic similarity estimation in terms of knowledge management problems Advances in Intelligent Systems and Computing 573 84-93
[12] Bastos-Filho C J A, Guimarães A C S 2015 Multi-objective fish school search Swarm Intelligence Research 6 23–40
[13] Filho J B M, Albuquerque I M C, Neto F B L, Ferreira F V S 2017 Improved Search Mechanisms for the Fish School Search Algorithm Advances in Intelligent Systems and Computing 557 362-371
[14] Monteiro R P, Verçosa L F V, Bastos-Filho C J A 2018 Improving the Performance of the Fish School Search Algorithm Swarm Intelligence Research 9 21-46