VILNIUS UNIVERSITY

AURIMAS RAPEČKA

INCREASE OF THE EFFICIENCY OF RECOMMENDER SYSTEMS IN SOCIAL NETWORKS

Summary of Doctoral Dissertation

Physical Sciences, Informatics (09 P)

Vilnius, 2015

The doctoral dissertation was prepared at the Institute of Mathematics and Informatics of Vilnius University in 2010-2014.

Scientific Supervisor

Prof. Habil. Dr. Gintautas Dzemyda (Vilnius University, Physical Sciences, Informatics – 09 P).

The dissertation will be defended at the Council of the Scientific Field of Informatics of Vilnius University:

Chairman

Prof. Habil. Dr. Antanas Žilinskas (Vilnius University, Physical Sciences, Informatics – 09 P).

Members:

Prof. Dr. Romas Baronas (Vilnius University, Physical Sciences, Informatics – 09 P),

Prof. Habil. Dr. Kazys Kazlauskas (Vilnius University, Physical Sciences, Informatics – 09 P),

Assoc. Prof. Dr. Raimundas Matulevičius (University of Tartu, Physical Sciences, Informatics – 09 P),

Prof. Dr. Dalius Navakauskas (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07 T).

The dissertation will be defended at the public meeting of the Council of the Scientific Field of Informatics Sciences of Vilnius University in the auditorium number 203 at the Institute of Mathematics and Informatics of Vilnius University at 10 a. m. on the 29th of September 2015.

Address: Akademijos st. 4, LT-08663 Vilnius, Lithuania.

The summary of the doctoral dissertation was sent out on 28th of August 2015.

A copy of the doctoral dissertation is available for review at the Library of Vilnius University or on this website: *www.vu.lt/lt/naujienos/ivykiu-kalendorius*

VILNIAUS UNIVERSITETAS

AURIMAS RAPEČKA

REKOMENDACINIŲ SISTEMŲ SOCIALINIUOSE TINKLUOSE EFEKTYVUMO DIDINIMAS

Daktaro disertacija,

Fiziniai mokslai, informatika (09 P)

Vilnius, 2015

Disertacija rengta 2010–2014 metais Vilniaus universiteto Matematikos ir informatikos institute.

Mokslinis vadovas

prof. habil. dr. Gintautas Dzemyda (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P).

Disertacija ginama Vilniaus universiteto Informatikos mokslo krypties taryboje:

Pirmininkas

prof. habil. dr. Antanas Žilinskas (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P).

Nariai:

prof. dr. Romas Baronas (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P),

prof. habil. dr. Kazys Kazlauskas (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P),

doc. dr. Raimundas Matulevičius (Tartu universitetas, fiziniai mokslai, informatika – 09 P),

prof. dr. Dalius Navakauskas (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Disertacija bus ginama viešame Vilniaus universiteto Informatikos mokslo krypties tarybos posėdyje 2015 m. rugsėjo 29 d. 10 val. Vilniaus universiteto Matematikos ir informatikos instituto 203 auditorijoje.

Adresas: Akademijos g. 4, LT-08663 Vilnius, Lietuva.

Disertacijos santrauka išsiuntinėta 2015 m. rugpjūčio 28 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir VU interneto svetainėje adresu: *www.vu.lt/lt/naujienos/ivykiu-kalendorius*

INCREASE OF THE EFFICIENCY OF RECOMMENDER SYSTEMS IN SOCIAL NETWORKS

1. Introduction

Research Area

With a rapid development of technologies and the increasing number of internet users, more and more products and services are transferred into the virtual space. Various offers to buy something by means of internet or to use a certain service without leaving the home should save the client's time. However, new problems occur here.

First of all, how to choose a product when the majority of the offered products are very similar and the client lacks experience? Secondly, how to find the necessary product among others, more often unnecessary products? Recommendation systems are widely used to solve these problems.

For a great number of algorithms, used for creating user's recommendations, the data sets of users, products, and product evaluation by users are required. These sets are widely accumulated in the internet shops and social websites where people have an opportunity to converse, share their opinions and, in that way, directly or indirectly evaluate products and services. Usually, social networks store huge amounts of information, and that may negatively influence users' social actions and reduce a possibility to find useful information quickly, so recommender systems (RS) are necessary here. Thus, it is considered that internet shops and social networks are supposed to be the most useful medium for using RS.

Relevance of the Problem

There are a lot of methods to recommend products or services. Each method has its advantages and disadvantages. For example, widely distributed universal methods (e.g. k-nearest neighbors) demonstrate good results in most data sets, but, because of basic operating principles (calculating the correlation coefficient between the target and every other user in a dataset), these methods require large computing resources. With limited computation resources, application of these methods is not successful. This problem is very important for online recommender systems.

In practice, there are two basic types of recommender systems: recommending by the digital ratings and recommending by "used – not used" (binary ratings 0 and 1). The latter type is more popular, because such data does not need to be collected from a direct customer feedback (it is not necessary to evaluate the movie watched, product consumed, etc.). World's most popular method now is the k-nearest neighbors method. This type of methods demonstrates good results with the main part of datasets. Methods, that are faster and suitable for big data, are not universal and demonstrate best results only on specific types of datasets, depending on their filling or size. This fact has also been confirmed by the experimental results in Chapter 3 of the dissertation.

In order to reduce calculations of the k-nearest neighbors method, it is possible to group users by clustering methods. However, here it is important to determine the optimal number of clusters. This number can vary significantly in different datasets. The second problem is to assign a proper cluster for new user.

The method, based on the user clustering, is introduced in Chapter 4 of the dissertation. This method is suitable for high density datasets and determines the specifics of user groups when generating recommendations.

The Aim and Tasks of the Research

The aim of this work is to propose a new recommendation method, that determines the specifics of user groups when the user-item matrix is of high density.

To realize the aim of research, it is necessary to solve the following tasks:

1. To perform an analytical review of the basic principles of recommender systems.
2. To systematize knowledge about the recommendation methods and their efficiency.
3. To perform experimental study of efficiency of the popular recommendation methods.
4. To create a new recommendation method suitable for high density datasets for determinity the specifics of user groups when generating a recommendation.

Scientific Novelty

In this work, a new recommendation method is proposed and experimentally examined. This method determines the specifics of user groups when the user-item matrix is of high density.

The new method is suitable for datasets with a large number of users and relatively a small number of products. This type of datasets is very popular in specialized online stores and in the specific web directories.

Statements to be Defended

1. The proposed method determines the specifics of user groups.
2. User clustering speeds up the generation of recommendations.
3. There are no universal recommendation methods – the results depend on the specifics of the data set.
4. The effectiveness of the proposed method depends on the density of the user-item matrix.

Approbation and Publications of the Research

The main results of the dissertation were published in 8 research papers: five papers are published in periodicals, reviewed scientific journals; three papers are published in conference proceedings. The main results have been presented and discussed at 9 national and international conferences. The main results of the work has been applied in three projects: "Development of the Short Term Prognosis Model for the New Book Demand", "Development of the Methodology for a Recommender System in the Bookstore Manoknyga.lt" and "Theoretical and Engineering Aspects of e-Service Technology Development and Application in High-Performance Computing Platforms". The research "The Possibilities of User Clustering in Recommender Systems" has been awarded by Association "INFOBALT" at 2013.

Outline of the Dissertation

The dissertation consists of 5 chapters, references, and appendix. The chapters of the dissertation are as follows: Introduction, A Review of Recommender Systems and

Recommendation Methods, Experimental Evaluation of the Popular Recommendation Methods, A New Recommendation Method and Area of its Applications, Summary of the Results and Conclusions. The dissertation also includes the list of notation and abbreviations. The scope of the work is 113 pages that include 31 figures and 10 tables. The list of references consists of 74 sources.

2. A Review of Recommender Systems and Recommendation Methods

Social networks on the internet are a very new phenomenon in our lives. On the other hand, this phenomenon is now experiencing a rise. To better evaluate the need for users in social networks and to recommend suitable products, these users are analyzed in full.

One of the users of social networking analysis and suitable products selection techniques is recommender systems. The main fields of application of these systems are electronic commerce and social networks. In social networks, there are ideal conditions for the application of recommender systems: there are many users that evaluate and comment various products. These evaluations are used for generation of recommendations.

In order to better identify needs of the users and to provide more precise recommendations, recommender systems use various methods of data mining (clustering, factorization, neural networks etc.).

This section presents a review of recommender systems, recommendation methods, datasets that are suitable for experimental research, performance measurements of recommendation methods, and recommender system software.

Each RS has two subjects: the user and the product. The subject, is using this system and gaining new product recommendations about various products, is called an RS target user. RS operates with the user-item matrix that filled with user ratings (evaluations) of products. Usually, a part of the matrix is not filled. The density of the user-item matrix is the ratio between the known evaluations and maximal possible evaluations in this matrix. The density is not large, as usual.

Various methods are used for the creation of recommendations. They are divided into two major groups: content-based and collaborative filtering-based methods. The

collaborative filtering-based methods are divided into two subcategories: memory-based and model-based. To solve specific problems, hybrid RSs are used. These RSs join the content-based and collaborative filtering-based methods. In the case of user's data privacy, RSs are divided into two major groups: personalized and not personalized recommender systems. In the case of the source for generating recommendations, RSs are divided into two large groups: user correlation-based and items correlation-based RSs.

Not personalized recommender systems recommend products only by the average opinion of all other users to the products. Recommendations do not depend on a user, so all users receive the same recommendations. These RSs are very common in small electronic shops, because they need a small amount of computing resources. Personalized and not personalized systems differ by one viewpoint only – logging history of a target user. In this case, the user is identified at the moment of return to the system and recommendations are provided not only by the averages of other users' evaluations of products, but also by the target user's activity history.

Users' correlation-based RSs recommend products by similarity between the target user and all other users' behavior. That is often called as collaborative filtering and is widely used in recommender systems. Items correlation-based recommender systems recommend compatible products taking into account the sets of products that are purchased by other users. Items content-based RSs recommend a product by its specific characteristics.

RS can give two types of outputs: prediction or recommendation. Prediction is expressed by some number, which means a predictable evaluation of the product by the target user, and the recommendation is expressed as a set of products that should be most relevant to the target user.

The precision of recommendations depends on a wider context of the sales of the product, therefore RSs are closely related to global product demand forecasting methods. A principled demand forecasting and recommender systems interaction scheme is presented in Fig. 1. There are many global product forecasting methods. For example, time series analysis-based demand forecasting methods investigate historical data of demand. Once the model of demand change is developed, it is possible to use this model

for demand forecasting in the future. Various demand forecasting methods can increase the precision of recommendations. Therefore, some of the global demand forecasting methods are reviewed in the dissertation. The forecast is accurate when the prediction error is relatively small. The following measures of the forecasting error are used most commonly: Mean Deviation (*MD*), Absolute Mean Deviation (*AMD*), Mean Squared Error (*MSE*), Root Mean Squared Error (*RMSE*), Mean Absolute Error (*MAE*) and Normalized Mean Absolute Error (*NMAE*). These measures can be used to evaluate the precision of recommendations too.

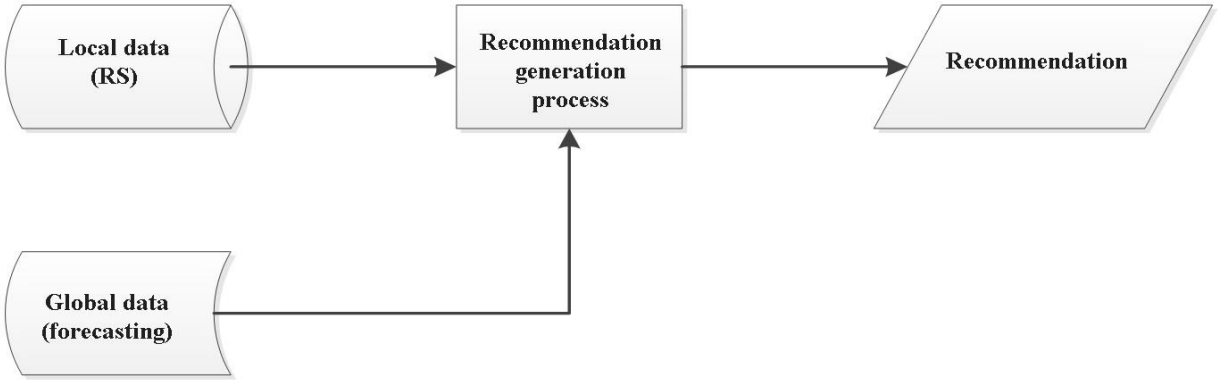

Figure 1. *Demand forecasting and recommender systems interaction scheme.*

The most popular groups of recommendation methods are as follows:

- Random offer,
- Recommendation of the most popular products,
- Recommendation of the most popular products by attributes,
- ***k*** nearest neighbors,
- Bayesian (content-based),
- Clustering based,
- Other methods.

In order to determine how accurately a recommender system predicts if a target user has purchased a product, several efficiency measures are used. The most popular measures are as follows: Precision, Specificity, Mean Average Precision (*MAP*), *recall@k*, *prec@k*, Normalized Discounted Cumulative Gain (*NDCG*), Mean Reciprocal Rank (*MRR*), and some others.

Table 1. *Comparision of free and open source Recommender Systems.*

Recommender system	Last update	Programming Language	Count of RS methods	Possibility to realize own method	Rating prediction	Product recommendations
MyMediaLite	2013 02	*C#*	>20	Yes	Yes	Yes
Apache Mahout	2012 06	*Java*	3	No	Yes	Yes
GraphLab	2012 05	*C++*	15	Yes	Yes	Yes
LensKit	2012	*Java*	?	Yes	Yes	Yes
Waffles	2013 04	*C++*	>5	No	Yes	No
easyrec	2012 02	*Online*	1	No	Yes	Yes
RecLab	2011 02	*Online*	1	No	Yes	No
Crab	2011	*Python*	>5	Yes	Yes	Yes
recommenderlab	2011 11	*C++*	4	Yes	Yes	Yes
Jellyfish	2012 12	*Python*	1	No	Yes	No
wooflix	2009 06	*Python*	1	No	Yes	No
OpenSlopeOne	2010 06	*PHP*	1	Yes	Yes	No
AppRecommender	2011	*Python*	>10	No	Yes	Yes

The data of product estimates by users are important for decision making. That is the reason why big social websites and internet shops do not disclose them. However, some material of this type is available for analysis. Each dataset has its own specifics, it depends not only on the structure of a dataset, but also on users, who have created these datasets. In scientific research, *MovieLens*, *Jester*, *Manoknyga.lt* and *Sapnai.net* datasets were used.

In the dissertation, an analytical comparison of the most popular free and open source software is presented. The results are generalized in Table 1. There are many recommendation methods and recommender systems, and this variety show that all these methods are not universal. We note that the *MyMediaLite* recommender software is better than others. Thus, effectiveness of the most popular recommendation methods was tested using *MyMediaLite* software.

The area of recommender system usage is very sensitive from the viewpoint of ethics. In the dissertation, this issue attracted attention because RSs use personal data of the users. The main problem is not the usage of personal data for generating recommendations, but a possibility to collect and sell these data to the third party.

3. A Experimental Evaluation of the Popular Recommendation Methods

In this section, the experimental evaluation of most popular recommendation methods is performed. The aim is to examine the efficiency of recommendation methods with several different datasets.

The cross validation model was used in this research (Fig. 2). Data of the users, who rated less products than K, are removed from the set. Then the data of the remaining users were divided into K equal parts $V_1, \dots V_K$. Afterwards, one part from $V_1, \dots V_K$ is considered as a validation set, and all the other $K-1$ are used in forming the training set. The removed data of users, who rated less than K products, now are included into the training set, too. Experiments were carried out consentively choosing one of $V_1, \dots V_K$ as the validation set.

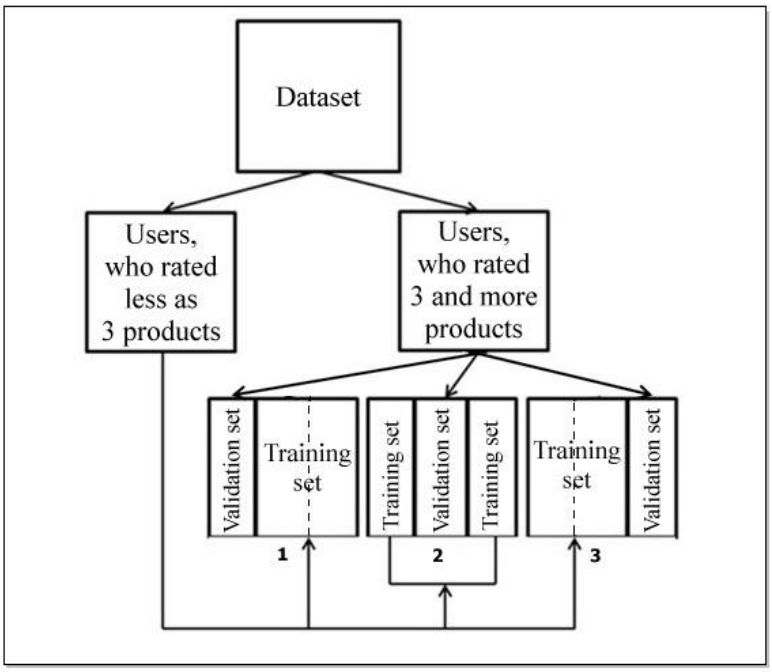

Figure 2. *Example of the Cross-Validation Model,* $\boldsymbol{K = 3}$.

Two types of experiments were done: experiments of rating prediction and that of product recommendation. Two different datasets and several recommendation methods were used for each type of experiments, performed with the MyMediaLite software.

In the case of rating prediction, *Jester* and *MovieLens* datasets were used. Effectiveness of the methods was estimated by *RMSE*, *MAE* and *NMAE* measures. Each measure defines the efficiency from one specific point of view, so it is necessary to take into account totality of these measures. The results are ranked by each measure and the average place of each method by all measures is calculated. In this way, the best methods are determined for each dataset. A summary of the experimental results is presented in Table 2.

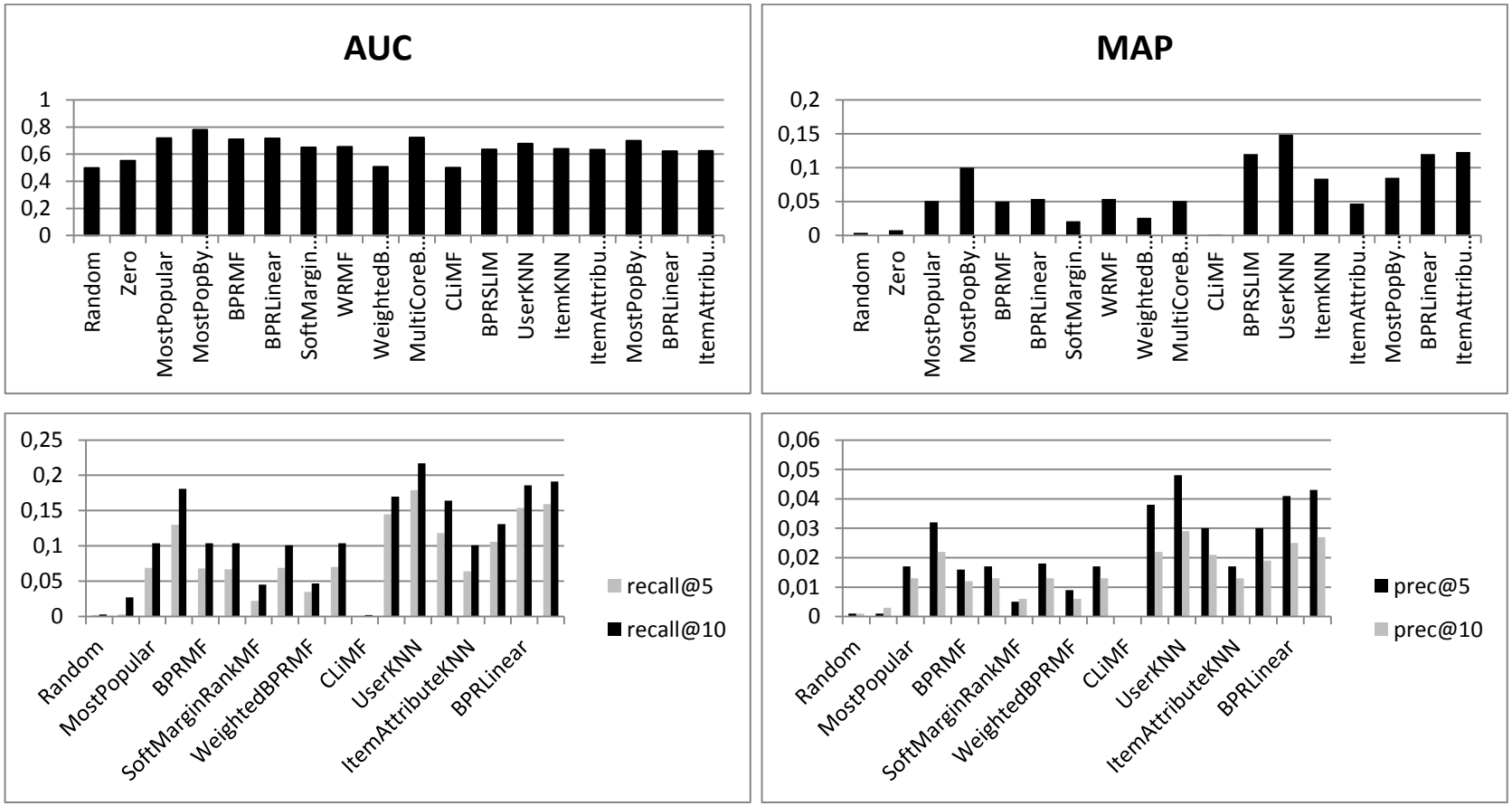

Figure 3. *Performance of recommendation methods in Sapnai.Net dataset.*

Table 2. *Results of cross-validation in different datasets*.

Method	Final place (Manoknyga.lt dataset)	Final place (Sapnai.net dataset)	Final place (MovieLens dataset)	Final place (Jester dataset)
Random	**17**	**10**	**11**	**12**
Zero	**16**	**8**	-	-
MostPopular	**9**	**2**	-	-
MostPopularByAttributes(C)	**3**	-	-	-
BPRMF	**12**	**3**	-	-
BPRLinear	**8**	**8**	-	-
SoftMarginRankingMF	**14**	**6**	-	-
WRMF	**10**	**5**	-	-
WeightedBPRMF	**15**	**7**	-	-
MultiCoreBPRMF	**11**	**4**	-	-
CLiMF	**18**	-	-	-
BPRSLIM	**4**	**1**	-	-
UserKNN	**1**	-	-	-
ItemKNN	**7**	-	-	-
ItemAttributeKNN	**13**	-	-	-
MostPopularByAttributes(A)	**6**	-	-	-
BPRLinear	**5**	-	-	-
ItemAttributeKNN	**2**	-	-	-
SlopeOne	-	-	**5**	**4**
GlobalAverage	-	-	**10**	**11**
ItemAverage	-	-	**8**	**9**
MatrixFactrorization	-	-	**1**	**2**
UserAverage	-	-	**9**	**9**
UserItemBaseline	-	-	**7**	**6**
CoClustering	-	-	**6**	**3**
LatentFeatureLogLinearModel	-	-	**4**	**6**
BiasedMatrixFactorization	-	-	**1**	**10**
SVDPlusPlus	-	-	**3**	**1**
SigmoidCombined AsymmetricFactorModel	-	-	-	**5**

In the case of product recommendation, *Manoknyga.lt* and *Sapnai.net* datasets were used. Effectiveness of the methods was estimated by *AUC*, *prec@5*, *prec@10*, *recall@5*, *recall@10*, *MAP*, *NDCG*, and *MRR* measures. The results are ranked according to each measure and the average place of each method is calculated according all measures. In this way, the best methods are determined for each dataset. An exemplary efficiency of methods in Sapnai.net dataset is presented in Fig. 3. A summary of experimental results is presented in Table 2.

The analysis has showed that there is no universal recommendation method, what could be suitable for any dataset. The methods analyzed experimentally demonstrate different results in each dataset. For example, the best method in the *Jester* dataset was *MatrixFactorization*, while the *CoClustering* method was only the 6th one. In the *Movielens* dataset, the best results were demonstrated by the *SVDPlusPlus* method, and the *CoClustering* method took the 3rd place.

Note, that for low density datasets, the recommendation methods, based on determination of the most popular products, demonstrate best results. This note is not unexpected, because the problem of empty ratings appears here and this type of methods eliminates this problem. The experimental research has showed that universal methods (like *UserKNN* and *ItemKNN*) are suitable for small datasets only. These methods require a large amount of computing resources, so it was possible to carry out experiments only with a relatively small *manoknyga.lt* dataset. For experiments with larger datasets computer resources were insufficient. This fact justifies the need to improve methods in the way of resource reduction. The improvement should be directed not to more efficient implementation of the existing methods, but to a modification of the existing methods for specific areas, i.e. to increasing of specialization of the methods.

4. A New Recommendation Method and the Area of its Applications

In this chapter, a new recommendation method is proposed and experimentally examined. This method determines the specifics of user groups when the user-item matrix is of high density. The new method is suitable for datasets with a large number of users and a relatively small number of products. This type of datasets is very popular in specialized online stores and in the specific web directories.

UserKNN and *ItemKNN* methods are considered as universal in various scientific publications. These methods are based on the *k* nearest neighbor (*KNN*) principles. Experimental researches in Chapter 3 have showed that these methods require large computing resources, especially if many data about users and products are collected. That justifies the need to improve the methods in the way of resource reduction. The discovery process of *k* similar users is a complicated task, what requires a large amount of computing resources. The main idea of a new recommendation strategy is to group

similar users into several clusters and to find a cluster, whose users are most similar to the target (new) user. We suggest here finding not k similar users, but a larger group of them. The size of this group is not known in advance.

Suppose that we have a user-item matrix $V = \{V_{ij}, i = \overline{1,m}, j = \overline{1,n}\}$ of m users $a_1, \dots, a_m$ and n products $b_1, \dots, b_n$. Let a user evaluate products using the scale of integer numbers $\{u_{min}, \dots, u_{max}\}$, consisting of n_u elements $n_u = u_{max} - u_{min} + 1$. The meaning of V_{ij} indicates the evaluation by the *i*-th user of the *j*-th product. $V_{ij} \notin \{u_{min}, \dots, u_{max}\}$, if the *i*-th user has not evaluated the *j*-th product.

Let us consider the matrix $V = \{V_{ij}, i = \overline{1,m}, j = \overline{1,n}\}$ as fully filled with the evaluations (it consists of the data on m users, who have rated all the n products). This matrix can be produced out of some available user-item matrix, where the number of users is larger than m, by picking the users, who rated 100% of products, and, in this way, to form a set of m users who rated all the products. Thus, the method is designed to generate recommendations, if the evaluation density of data is large enough. *Jester I*, *Jester II*, and *Jester III* are the available datasets suitable to form matrix V, that is fully filled with evaluations.

Using the matrix V, it is possible to classify the users into similar users' clusters $C_1, C_2, \dots, C_k$ that comprise a different number of users:

$C_1 = \{a_1^1, a_2^1, \dots, a_{m_1}^1\}$ with m_1 users,

(...)

$C_k = \{a_1^k, a_2^k, \dots, a_{m_k}^k\}$ with m_k users,

where $m = \sum_{l=1}^{k} m_k$, a_i^l is the *i*-th user of the *l*-th cluster, $A = C_1 \cup C_2 \cup \dots \cup C_k$; $C_1 \cap C_2 \cap \dots \cap C_k = \emptyset$.

Each cluster C_l, $l = 1, k$, has a center:

$$X_l = (x_1^l, \dots, x_n^l),\ x_j^l = \frac{\sum_{a_i \in C_l} V_{ij}}{m_l}, \quad (1)$$

where m_l is the number of users in the l-th cluster C_l. The dimensionality of the cluster center is *n* because it is equal to the number of products.

When a new user (target user) a_N joins the system, s randomly selected products $b_{N_1}, b_{N_2}, \dots, b_{N_s}$ are presented for his evaluation from the set of products $\{b_1, \dots, b_n\}$.

Here N_i is the number of product order between 1 to n; $s < n$, and $V_N = (V_{NN_1}, \ldots, V_{NN_s})$ are ratings of the products $b_{N_1}, b_{N_2}, \ldots, b_{N_s}$ by the new user.

After obtaining the new evaluations $V_{NN_1}, \ldots, V_{NN_s}$ of s products $b_{N_1}, b_{N_2}, \ldots, b_{N_s}$, lower dimensional cluster centers are selected in each cluster C_l, $l = \overline{1,k}$, based on the products $b_{N_1}, \ldots, b_{N_s}$ only:

$$X_l^N = \left(x_{N_1}^l, \ldots, x_{N_s}^l\right). \quad (2)$$

The dimensionality N_s of the cluster center here is lower than *n*, because it is equal to the number of products N_s evaluated by the user a_N.

Then the Euclidean distances

$$\rho(V_N, X_l^N) = \sqrt{\sum_{i=1}^{s}(V_{NN_i} - x_{N_i}^l)^2}, l = \overline{1,k} \quad (3)$$

between the ratings V_N and lower dimensional cluster centers are calculated. The user a_N is allocated to the cluster, where distance (3) is lower, i.e. $a_N \in C_{l^*}$, where $l^* = arg \min_{l=1,k} \rho(V_N, X_l^N)$.

Then, it is possible to offer the best rated products of the users from the cluster C_{l*} to the user a_N.

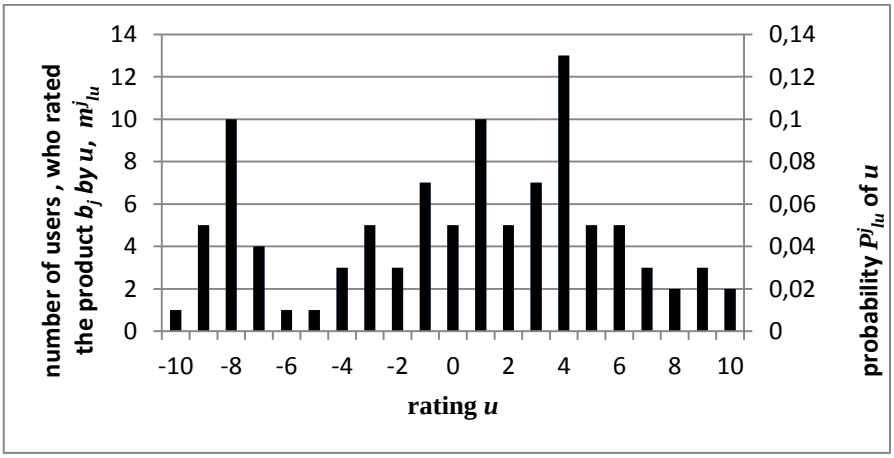

Figure 4. *Example of the distribution of product rating in the cluster C_l.*

Each product $b_{j.}$ in the cluster C_l has the distribution of rating that shows how many users of the cluster C_l provided the rating u, $u \in \{u_{min}, \ldots, u_{max}\}$, for this product (see

Fig. 4). The height of the column m_{lu}^{j} illustrates how many users of the cluster C_l defined the rating u for the product $b_{j.}$ Note that $\sum_{u=u_{min}}^{u_{max}} m_{lu}^{j} = m_l$.

Fig. 4 defines the function of the distribution density of a particular rating u. The scale on the right side of Fig. 4 shows the probability that the users of the cluster C_l will provide the evaluation u of the product $b_{j:}$

$$P_{lu}^{j} = P\left(V_{ij} = u, \text{if } a_i \in C_l\right) = \frac{m_{lu}^{j}}{m_l}. \quad (4)$$

Note that $\sum_{u=u_{min}}^{u_{max}} P_{lu}^{j} = 1$.

Using formula (5), it is possible to calculate the average rating given by the users of the cluster C_l for the product b_j:

$$\bar{V}_l^{\,j} = \frac{1}{u_{max} - u_{min} + 1} \sum_{u=u_{min}}^{u_{max}} P_{lu}^{j}\, u. \quad (5)$$

The product with the highest average rating in the cluster C_l is recommended for the new user a_N. If this product has already been offered for this user, then the system recommends the product with the second in size rating, and so on.

After the new user has evaluated the recommended product, the total number of his evaluated products increases, it means: $s = s + 1$.

The calculations above are repeated starting from formula (2), in which the evaluations of s products by the new user are used to calculate the lower dimensional cluster centers in each cluster C_l, $l = \overline{1,k}$, based on the products $b_{N_1}, \dots, b_{N_s}$.

After the new user has finished the work (after using and rating the offered or chosen products), the matrix V is extended to a new row with the ratings of this user, $m = m + 1$.

The method proposed to create recommendations is based on clustering of users. In order to get the optimal recommendations, it is necessary to determine the proper number of clusters k. The experimental research should disclose dependences of the recommendation efficiency on the number of clusters. In addition, this research should reveal whether the user classification is an acceptable way in the creation of recommendations to the new users.

For experiments, the first Jester data set is used. The density of evaluations amounts up to 75%, so each user has approximately evaluated 75 products out of 100. Moreover,

7200 users have rated all the 100 products.

In order to draw objective conclusions, calculations that disclose the new user's behaviour after rating the product s are done a lot of times and the average results obtained. Let us fix the value of s. Then, for each user from the validation set, 100 experiments were done with randomly selected product sets $\{b_{N_1}, \dots, b_{N_s}\}$. The average of the results over all randomly selected product sets and all users from the validation set is found out for different s. Let us denote the average rating of the offered product $s+1$ over all the users by $\bar{u}_{s+1}$, when the ratings of the previous s products are known. The results, gained during the experiments, are presented in Fig. 5, Fig. 6 and Table 3.

From the diagrams of Fig. 5, we see that in some starting interval $[1;\ s(\bar{V}_{max})]$ the meaning of $\bar{V}$ is growing. The maximal growth increases with an increase in the number k of the user clusters, however, the best result (the highest value of $\bar{V}_{max}$) is obtained not with the highest or lowest number of the clusters. Therefore, it is some optimal number of clusters. In the interval $[s(\bar{V}_{\max});\ 99]$, we observe a decrease of $\bar{V}$.

In the case $s = n-1$, the value of $\bar{u}_{s+1} = \bar{u}_{n-1}$ becomes the average of ratings of all the products of all the users. When we have a small amount of the rated products by the new user, we cannot be quite sure that the decision on a proper cluster is really good. However, when s reaches a particular size, the cluster, which the user is assigned to, does not change.

The efficiency of the proposed method is compared with that of three other methods: *Random* method (point of baseline); *MostPopular* method (when the users' clustering process is not used and it is considered that all users in a dataset are similar); with a collaborative filtering-based (*UserKNN*) method. The comparision criterion is the average rating of a recommended product. *Jester* I dataset was used for experiments. The dependence of the average rating of recommended product on the number s of rated products is presented in Fig. 7.

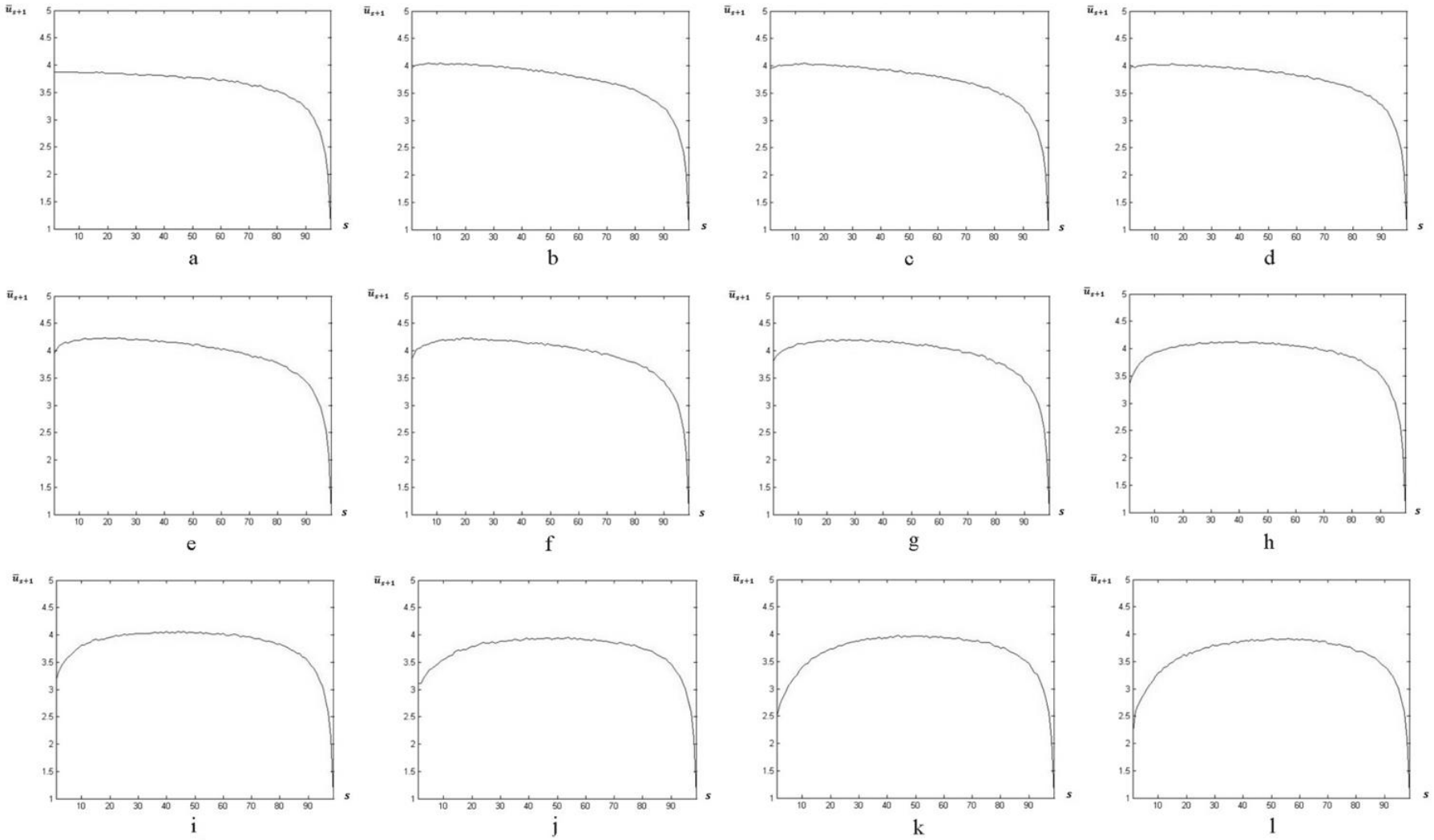

Figure 5. *Dependence of the average rating* $\bar{u}_{s+1}$ *of the offered* $(s+1)$ *product on the number* k *of clusters and on the amount* s *of the already rated products.*

Table 3. *Experimental results of the dependence on the number of clusters.*

Clusters count k	**Starting value $\bar{u}_2$**	**Equalization point s^***	**Maximal value $\bar{u}_{max}$**	**Point of $\bar{u}_{max}$ (s_{opt})**	**$\bar{u}_{max}$-$\bar{u}_2$**	**Final value $\bar{u}_{n-1}$**	**Middle value of interval $[1; s^*]$**
1	3,877	-	3,877	1	0,000	1,081	-
2	3,949	36	4,048	11	0,099	1,068	4,034
3	3,936	38	4,045	13	0,109	1,051	4,020
5	3,929	44	4,032	16	0,102	1,062	4,012
7	3,939	60	4,198	20	0,259	1,067	4,148
10	3,948	68	4,235	24	0,287	1,081	4,196
15	3,851	75	4,232	22	0,387	1,080	4,165
20	3,801	79	4,204	28	0,403	1,073	4,164
30	3,677	86	4,210	26	0,534	1,062	4,160
50	3,355	92	4,131	39	0,796	1,081	4,120
70	3,169	94	4,054	46	0,884	1,081	4,035
100	3,116	94	3,940	50	0,825	1,074	3,936
150	2,495	97	3,976	44	1,481	1,067	3,968
200	2,190	98	3,916	56	1,726	1,062	3,904

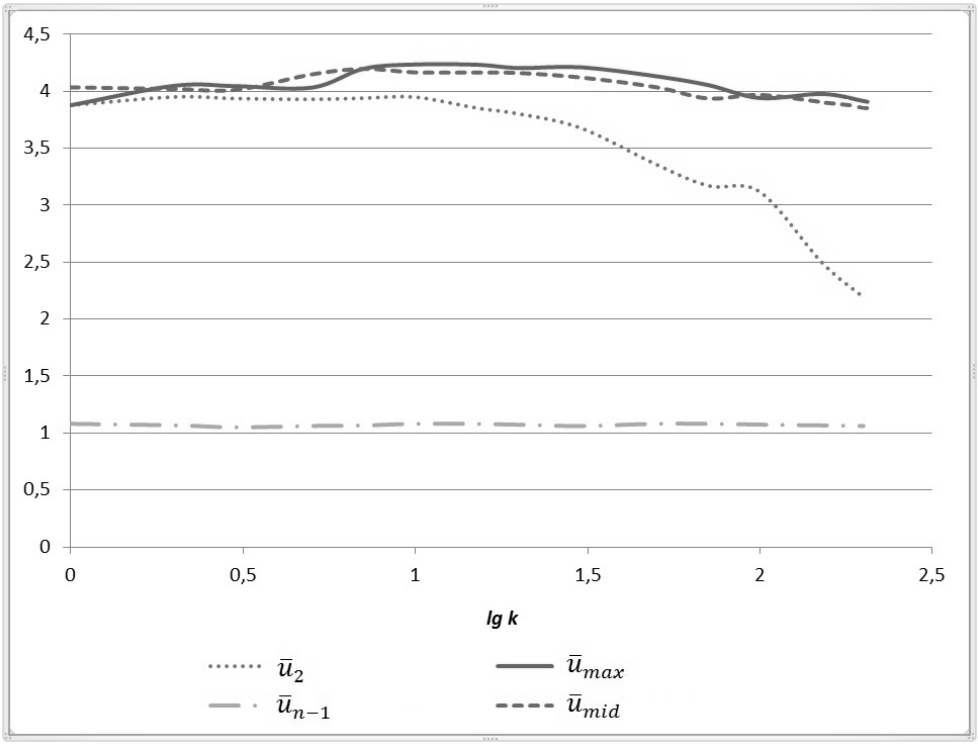

Figure 6. *Dependance of $\bar{u}_2$, $\bar{u}_{max}$, $\bar{u}_{n-1}$ and $\bar{u}_{mid}$ on the number of clusters k.*

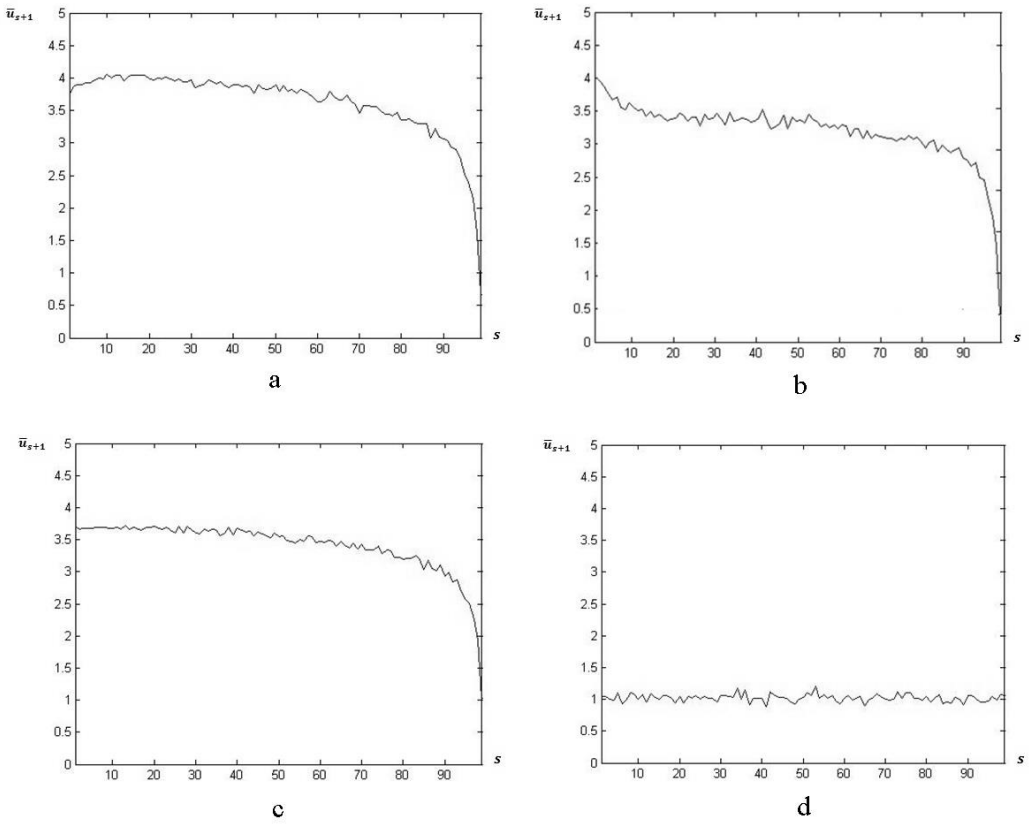

Figure 7. *Comparision of methods: **a)** the proposed method, **b)** the collaborative filtering based (UserKNN) method, **c)** Most Popular method, **d)** Random method.*

The proposed method and the collaborative filtering method are more effective each other in different intervals of s. The new method demonstrates better results, when a longer history of user‘s ratings is known. Therefore, it would be possible to merge both methods and gain a better recommendation.

Another advantage of the proposed method is faster computations as compared to the collaborative filtering method. Of course, the process of users’ clustering takes a lot of time, but this process can be moved to the background of computing, using grids or supercomputers (a real time mode is not necessary).

5. Summary of the Results and Conclusions

Social networks and internet shops are now experiencing a rise. To better evaluate the need of users in social networks and internet shops and to recommend suitable products, these users are analyzed in full.

In the dissertation, a new recommendation method is proposed and experimentally examined. This method determines the specifics of user groups when the user-item matrix is of high density. The new method is suitable for datasets with a large number of users and relatively small number of products. This type of datasets is very popular in specialized online stores and in the specific web directories.

The main idea of the new recommendation method is to group similar users into several clusters and to find a cluster, whose users are most similar to the target (new) user. We suggest here to find not some fixed number of users similar to the targert user, but a large group of them of size not predefined in advance.

The research leads to the following conclusions:

1. The proposed method determines the specifics of behavior of user groups. The experimental research with thefirst Jester database has shown that the optimum number k of user clusters belongs to the interval [7; 30]. The best result is gained as k=10, where the maximal average rating of the offered product over all the users increases up to 9,2 %, as compared with the case, where there is no clustering. The essential increase in the number of clusters is not reasonable. If 200 clusters are used, the maximal average rating of the offered product over all the users is higher only by 1 % than that in the case where there is no clustering.

2. User clustering speeds up the generation of recommendation. The new method generates recommendations two times faster than the traditional collaborative filtering method. On the other hand, the new method generates better recommendations, when a new user has rated more than 5 % of products.
3. There are no universal recommendation methods, the results depend on the specifics of a dataset. The methods analyzed experimentally demonstrate different results in each dataset. For example, the best method in the *Jester* dataset was the *MatrixFactorization* while the *CoClustering* method took only the sixth place. In the *Movielens* dataset, the best results were shown by the *SVDPlusPlus* method, while the *CoClustering* method was third. The proposed method and the collaborative filtering method are more effective in different intervals of products rated by users. New method demonstrates better results, when a longer history of user‘s ratings is known. Therefore, it would be possible to merge both methods and gain better recommendations.
4. The effectiveness of the proposed method depends on the density of the user-item matrix. The best recommendations are obtained when the history of product evaluations by the new user contains about 25 % of all the products in the database.

List of Publications on the Topic of Dissertation

The articles published in the peer-reviewed periodical publications:

1. **Rapečka, Aurimas**; Marcinkevičius, Virginijus. Knygų paklausos prognozavimo elektroniniame knygyne galimybės. *Informacijos mokslai*, ISSN 1392-1487. Accepted.

2. **Rapečka, Aurimas**; Dzemyda, Gintautas. A new recommendation model for the user clustering-based recommendation system. *Information Technology and Control*, 2015, Vol. 44, No. 1, ISSN 1392-124X, p. 54–63. (Impact Factor 2014: 0,623)

3. Kurasova, Olga; Marcinkevičius, Virginijus; Medvedev, Viktor; **Rapečka, Aurimas**; Stefanovič, Pavel. Strategies for big data clustering. *ICTAI 2014: 2014 IEEE 26th International Conference on Tools with Artificial Intelligence* [Proceedings], 10-12 November, 2014, Limassol, Cyprus., IEEE Computer Society, 2014. ISSN 1082-3409, p. 740–747.

4. **Rapečka, Aurimas**; Marcinkevičius, Virginijus; Dzemyda, Gintautas. Rekomendacinės sistemos algoritmų veikimo elektroninio knygyno duomenų bazėje analizė. *Informacijos mokslai*, 2013, t. 65, ISSN 1392-0561, p. 45–55.

5. Kurasova, Olga; Marcinkevičius, Virginijus; Medvedev, Viktor; **Rapečka, Aurimas**. Duomenų tyrybos sistemos, pagrįstos saityno paslaugomis. *Informacijos mokslai*, 2013, t. 65, ISSN 1392-0561, p. 66–74.

The articles published in the conference proceedings:

6. **Rapečka, Aurimas**; Dzemyda, Gintautas. A new recommendation method for the user clustering-based recommendation system. *EMC2015: Engineering Management and Competitiveness: 5th International Symposium* [Proceedings], June 19–20, 2015, Zrenjanin, Serbia, ISBN 9788676722563. p. 328–332.

7. Kligienė, Stanislava Nerutė; **Rapečka, Aurimas**. Challenges of digital era: potential and pitfalls of social media. Ethics and trust in collaborative cross-domains. *COLLA 2011*: *The First International Conference on Advanced Collaborative Networks, Systems and Applications* [Proceedings], June 19–24, 2011, Luxembourg, ISBN 9781612081434. p. 34–39.

8. **Rapečka, Aurimas**; Dzemyda, Gintautas. Rekomendacinių sistemų ir jose naudojamų rekomendavimo algoritmų apžvalga. *XV Kompiuterininkų konferencijos mokslo darbai*. Vilnius, 2011, ISBN 9789986342618. p. 175–185.

About the Author

Aurimas Rapečka was born on the 3rd of June, 1986 in Zarasai district, Lithuania. In 1992-2000 he attended the primary school of Degučiai, in 2004 he graduated from the Zarasai Ąžuolas gymnasium. He received a Bachelor's degree in Electronic Engineering from Vilnius Gediminas Technical University in 2008 and a Master's degree in Communication and Information from Vilnius University in 2010. From 2010 to 2014 he was a PhD student of Vilnius University, Institute of Mathematics and Informatics. Also, from 2009 till now he is working as a researcher at Institute of Mathematics and Informatics, Vilnius University.

REKOMENDACINIŲ SISTEMŲ SOCIALINIUOSE TINKLUOSE EFEKTYVUMO DIDINIMAS

Tyrimų sritis

Tobulėjant technologijoms ir daugėjant aktyvių interneto vartotojų vis daugiau produktų ir paslaugų perkeliama į virtualią aplinką. Neišeidamas iš namų klientas gali internete įsigyti prekę ar pasinaudoti paslauga, taip taupydamas savo laiką. Tačiau čia atsiranda naujų problemų.

Pirmiausia – kaip išsirinkti prekę tada, kai visi siūlomi produktai panašūs, o patirties mažai; ir antra – kaip surasti reikiamą produktą tarp daugelio kitų, visai nereikalingų. Šioms problemoms spręsti plačiai taikomos rekomendacinės sistemos (RS). Didžiajai daliai metodų, taikomų rekomendacijoms kurti, reikalingos vartotojų, produktų ir vartotojų vertinimų produktams aibės. Šios aibės plačiai kaupiamos elektroninėse parduotuvėse ir socialiniuose tinkluose, kur žmonės gali bendrauti tarpusavyje, išsakyti savo nuomonę ir tokiu būdu tiesiogiai ar netiesiogiai vertinti produktus ir paslaugas. Todėl manoma, kad tiek elektroninės parduotuvės, tiek socialiniai tinklai yra tinkamiausios terpės taikyti RS.

Darbo aktualumas

Yra sukurta daug metodų, taikomų rekomenduoti prekėms ar paslaugoms. Kiekvienas iš jų turi privalumų ir trūkumų. Pavyzdžiui, plačiai paplitusiais universaliais metodais (pavyzdžiui, *k* artimiausių kaimynų) galima pasiekti gerų rezultatų su didžiąja dalimi duomenų rinkinių, tačiau dėl savo veikimo principų (koreliacijos koeficiento apskaičiavimo tarp konkretaus ir kiekvieno kito vartotojo įverčių viso duomenų rinkinio produktams) šie metodai reikalauja didelių skaičiavimo resursų. Jei skaičiavimo resursų ištekliai arba skaičiavimų trukmė ribota, šių metodų rezultatyviai pritaikyti nepavyksta. Ši problema ypač aktuali realaus laiko rekomendacinėms sistemoms, veikiančioms internete.

Praktikoje išskiriami du pagrindiniai rekomendacinių sistemų tipai: rekomenduojančios pagal skaitinius įverčius ir rekomenduojančios pagal „naudojo ar nenaudojo“, tai yra dvejetainį įverčių formatą (0 ir 1). Pastarasis tipas yra populiaresnis,

nes duomenims surinkti nereikia tiesioginių vartotojų atsiliepimų (nereikalaujama įvertinti žiūrėtą filmą, suvartotą produktą ir pan.). Pasaulyje labiausiai paplitę k artimiausių kaimynų tipo metodai, kuriais pasiekiami rezultatai esti pakankamai geri su dauguma duomenų rinkinių. Greitesni ir dideliems duomenų rinkiniams galimi taikyti metodai nėra labai universalūs ir geriausius rezultatus pasiekia taikomi konkretaus tipo, užpildymo ar dydžio duomenų rinkiniams. Tai patvirtino ir disertacijos 3 skyriuje pateikti atliktų eksperimentų rezultatai.

Siekiant sumažinti k artimiausių kaimynų metodų skaičiavimų apimtis, panašius vartotojus galima suskirstyti į grupes klasterizavimo metodais. Tačiau svarbu nustatyti optimalų klasterių skaičių, kuris gali smarkiai skirtis įvairiuose duomenų rinkiniuose. Kita problema – naujo vartotojo priskyrimas prie tinkamo klasterio.

Vartotojų klasterizavimu ir minėtų problemų sprendimu paremtas ir disertacijos ketvirtojoje dalyje pristatomas autoriaus sukurtas metodas, per rekomendacijų kūrimo procesą įvertinantis ir vartotojų grupių specifiką, kai vartotojų įverčių produktams matrica yra tankiai užpildyta.

Darbo tikslas ir uždaviniai

Disertacijos tikslas – sukurti naują rekomendacijų kūrimo metodą, įvertinantį vartotojų grupių specifiką, kai vartotojų įverčių produktams matrica yra tankiai užpildyta.

Šiam tikslui pasiekti sprendžiami tokie uždaviniai:

1. Atlikti analitinę rekomendacinių sistemų veikimo principų apžvalgą.
2. Susisteminti žinias apie metodus rekomendacijoms kurti ir jų efektyvumą.
3. Atlikti eksperimentinį rekomendacinėse sistemose dažniausiai taikomų metodų efektyvumo tyrimą.
4. Sukurti naują rekomendavimo metodą, tinkamą taikyti duomenų rinkiniams, kai vartotojų įverčių produktams matrica yra tankiai užpildyta, ir įvertinantį vartotojų grupių specifiką kuriant rekomendacijas.

Mokslinis naujumas

Šiame darbe pateiktas ir ištirtas naujas rekomendavimo metodas, įvertinantis vartotojų grupių specifiką, kai vartotojų įverčių produktams matrica yra tankiai užpildyta.

Šis metodas tinkamas taikyti tokiems duomenų rinkiniams, kuriuose yra didelis vartotojų ir santykinai mažas produktų skaičius. Tokio tipo duomenų rinkiniai paplitę specializuotose elektroninėse parduotuvėse, taip pat įvairiuose specifiniuose interneto kataloguose.

Ginamieji teiginiai

1. Pateiktas rekomendavimo metodas įvertina vartotojų grupių elgsenos specifiką.
2. Vartotojų klasterizavimas pagreitina rekomendacijų generavimo procesą.
3. Nėra universaliai gerų rekomendavimo metodų – jų rezultatus lemia duomenų rinkinio specifika.
4. Pateikto metodo generuojamų rekomendacijų efektyvumą lemia vartotojų įverčių produktams matricos tankis.

Darbo rezultatų aprobavimas

Tyrimų rezultatai publikuoti 4 periodiniuose recenzuojamuose ir 3 kituose moksliniuose leidiniuose. Tyrimų rezultatai buvo pristatyti ir aptarti 9 nacionalinėse ir tarptautinėse mokslinėse konferencijose.

Disertacijos struktūra

Disertaciją sudaro 5 skyriai ir literatūros sąrašas. Disertacijos skyriai: Įvadas, Rekomendacinių sistemų ir jose taikomų metodų apžvalga, Populiariausių rekomendavimo metodų eksperimentinis įvertinimas, Siūlomas rekomendavimo metodas ir jo taikymo galimybių tyrimai, Rezultatų apibendrinimas. Papildomai disertacijoje pateikta: paveikslų, lentelių, naudotų žymėjimų ir santrumpų sąrašas bei priedai. Visa disertacijos apimtis yra 113 puslapių su priedais, juose pateiktas 31 paveikslas ir 10 lentelių. Disertacijoje remtasi 74 literatūros šaltiniais.

Rezultatų apibendrinimas ir išvados

Interneto socialiniai tinklai ir įvairios elektroninės parduotuvės šiuo metu tampa vis populiaresnės. Siekiama geriau įvertinti vartotojų poreikius ir rekomenduoti jiems tinkamus produktus ar paslaugas, todėl vartotojai ir jų elgsena yra įvairiopai analizuojami. Disertacijoje pateiktas rekomendavimo metodas, įvertinantis vartotojų grupių elgsenos specifiką, kai vartotojų įverčių produktams matrica yra tankiai užpildyta. Tokio tipo duomenų rinkiniai paplitę specializuotose elektroninėse parduotuvėse, taip pat įvairiuose specifiniuose interneto kataloguose. Naujo metodo tikslas – sugrupuoti panašius vartotojus į tam tikrą klasterių skaičių ir ieškoti klasterio, kuriam priklausantys vartotojai yra panašiausi į naują vartotoją, kuriam reikia sugeneruoti rekomendaciją. Čia ieškoma ne panašiausių vartotojų fiksuoto skaičiaus, o visos jų grupės, kurios dydis iš anksto nėra žinomas.

Tyrimai leido padaryti šias išvadas:

1. Pateiktas rekomendavimo metodas įvertina vartotojų grupių elgsenos specifiką. Eksperimentinis tyrimas su *Jester* I duomenų rinkiniu parodė, kad optimalus klasterių skaičius k šio duomenų rinkinio vartotojams priklauso intervalui $[7; 30]$, o geriausi rezultatai gaunami, kai $k = 10$. Tada pastebimas 9,2 % vidutinės rekomenduojamo produkto įverčio reikšmės didėjimas lyginant su populiariausių prekių rekomendavimu, kai neatsižvelgiama į pirkėją. Vis dėlto vartotojų klasterizavimas su dideliu klasterių skaičiumi nėra efektyvus – tyrimas parodė, kad didžiausia vidutinė rekomenduojamo produkto įverčio reikšmė, kai yra 200 klasterių, vos 1 % didesnė už to įverčio reikšmę vieno klasterio atveju.
2. Vartotojų klasterizavimas pagreitina rekomendacijų generavimo procesą. Pateiktu metodu produktų rekomendacijas tūkstančiui vartotojų pavyko sugeneruoti du kartus greičiau nei su įprastu bendrojo filtravimo metodu, realizuotu toje pačioje programinėje aplinkoje. Be to, gaunamos geresnės rekomendacijos nei su bendrojo filtravimo metodu, kai naujo vartotojo vertinimų produktams istorija apima daugiau kaip 5 % visų duomenų rinkinio produktų. Tolesnių tyrimų objektas galėtų būti šių metodų sujungimas – rekomendacijos būtų generuojamos iš pradžių vienu metodu, o po to kitu.

3. Nėra universaliai gerų rekomendavimo metodų – jų rezultatams daro įtaką duomenų rinkinio specifika. Geriausi rezultatai gauti taikant skirtingus metodus eksperimentiškai tirtiems duomenų rinkiniams. Tiriant *Jester* duomenų rinkinį, tinkamiausias metodas buvo *MatrixFactorization*; *CoClustering* metodas buvo tik šeštas. O štai tiriant *MovieLens* duomenų rinkinį tinkamiausias metodas buvo *SVDPlusPlus*, tačiau *CoClustering* metodas buvo trečias. Kadangi naujasis ir bendrojo filtravimo metodai yra efektyvesni vienas už kitą, kai vartotojų vertintų produktų skaičiaus reikšmės skirtingos, nagrinėtinas šių metodų sujungimas – rekomendacijos būtų generuojamos iš pradžių vienu metodu, o po to kitu.
4. Pateikto metodo generuojamų rekomendacijų efektyvumą lemia vartotojų įverčių produktams matricos tankis. Nustatyta, kad geriausios rekomendacijos gaunamos tada, kai naujo vartotojo vertinimų produktams istorija apima bent apie 25 % visų duomenų rinkinio produktų.

Trumpai apie autorių:

Aurimas Rapečka gimė 1986 m. birželio 3 d. Zarasų rajone. 1992-2000 metais lankė Degučių pagrindinę mokyklą, 2004 m. baigė Zarasų „Ąžuolo“ gimnaziją, 2008 m. Vilniaus Gedimino technikos universitete įgijo elektronikos inžinerijos bakalauro laipsnį. 2010 m. Vilniaus universitete įgijo komunikacijos ir infomacijos magistro laipsnį. Nuo 2010 m. iki 2014 m. buvo Vilniaus universiteto Matematikos ir informatikos instituto doktorantu, nuo 2009 m. iki dabar dirba Vilniaus universitete, Matematikos ir informatikos institute.

Aurimas Rapečka

REKOMENDACINIŲ SISTEMŲ SOCIALINIUOSE TINKLUOSE EFEKTYVUMO DIDINIMAS

Daktaro disertacija
Fiziniai mokslai, informatika (09 P)
Redaktorė Jorūnė Rimeisytė

Aurimas Rapečka

INCREASE OF THE EFFICIENCY OF RECOMMENDER SYSTEMS IN SOCIAL NETWORKS

Doctoral Dissertation
Physical Sciences, Informatics (09 P)
Editor Janina Kazlauskaitė