Slovak University of Technology in Bratislava

Faculty of Informatics and Information Technologies

FIIT-5208-72091

Veronika Gondová

User modeling in the domain of e-commerce

Master Thesis

Study programme: Information Systems
Study field: 9.2.6. Information Systems
Training workplace: Institute of Informatics, Information Systems and Software Engineering, FIIT STU Bratislava
Thesis supervisor: prof. Mária Bieliková

May 2018

Declaration of Honor

I honestly declared that this thesis was written independently by me under professional supervision of prof. Mária Bieliková. All references have been clearly cited.

Bratislava, May 2018 .....

Veronika Gondová

Acknowledgement

My special thanks belong to my supervisor prof. Mária Bieliková for all her time, human touch, professional support and helpful advices during the work on this thesis.

I wish to thank my classmate and friend, Zuzana Bobotová, for collaboration on defining the problem of event abstraction, the lecturer and PeWe member Dr. Michal Kompan for inspirative discussions and advisement and PeWe member Peter Gašpar for technical support and many inspirations.

My thanks belongs to the entire PeWe group, thanks to which I have received many new ideas and questions to think about.

I also wish to thank company ZľavaDňa, for the opportunity to work on this interesting thesis in the cooperation with domain experts. My special thanks belongs to Peter Sedlák, Matej Moravčík and Monika Tihanyiová for their time, expert advices and many inspirative discussions and Robert Kasanický for technical support and a lot of productive discussions.

I would also like to expect many thanks to my boyfriend, patents, sister, family and friends who helped me a lot and supported me during my studies and also during my work on this thesis.

Veronika Gondová

Anotácia

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný program: Informačné systémy

Autor: Bc. Veronika Gondová
Diplomová práca: Modelovanie používateľa v doméne e-obchodu
Vedúci diplomovej práce: prof. Ing. Mária Bieliková, PhD.
máj 2018

Web je zdrojom obrovského množstva informácií a služieb. Jednotlivé služby sú prevádzkované webovými sídlami. Konzumenti služieb ich využívajú s cieľom uspokojiť svoje potreby. Na mieru uspokojenia do veľkej miery vplýva výber vhodného zdroja informácií. Nakoľko web poskytuje obrovské množstvo informácií, používateľ nemá priestor na preskúmanie všetkých alternatív, dokonca často preskúma len nepatrný zlomok dostupných informácií.

Jeden z prostriedkov na zvyšovanie miery uspokojenia používateľa je odporúčanie. Hlavným zdrojom informácií pre odporúčanie na webe sú stopy používateľov webu, ktoré za sebou v jednotlivých službách zanechávajú. Miera uspokojenia z odporúčaného produktu je reprezentovaná v podobe presnosti odporúčania. Na presnosť odporúčania vplýva niekoľko faktorov, pričom medzi najhlavnejšie patrí množina dostupných dát o používateľovi a prostredí.

V našej práci sa venujeme modelovaniu používateľa ako základného prostriedku pre zvýšenie presnosti odporúčania. V práci sme navrhli postup tvorby modelu používateľa cez rôzne typy predikcií. Vzhľadom na povahu dát sme navrhli proces abstrakcie stôp používateľov cez vzory v správaní a ich využitie v predikcii demografických charakteristík. Výsledky našej práce boli implementované a overené nad datasetom z e-obchodu.

Annotation

Slovak University of Technology Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES
Degree course: Information Systems

Author: Veronika Gondová
Master Thesis: User modeling in the domain of e-commerce
Supervisor: prof. Mária Bieliková
May 2018

The web is the source of a huge amount of information and services. Individual services are operated by the websites. Consumers use them to meet their needs. The level of satisfaction is greatly influenced by the choice of a suitable source of information. Because the web provides a huge amount of information, the user does not have the opportunity to explore all alternatives.

One of the means to increase user satisfaction is a recommendation. The main source of information for web recommendation are footprints of web users. The satisfaction rate of the recommended product is represented as the precision of the recommendation. One of the most significant factor, that influence precision of recommendation is the available dataset that contains information about users and environment.

In our work, we are focusing on user modelling as the basic means to increase the precision of the recommendation. We proposed a method for user model creation via different types of prediction. Based on data character, we have suggested abstraction of users' footprints through patterns in behaviour and their usage in demography prediction tasks. The results of our work were implemented and verified with e-commerce dataset.

Content

1 Introduction

The web is the source of a large amount of information and services that are linked to each other. Providers publish them to satisfy the customers' needs. Maslow in [103] divided the human needs into 7 categories as followed: physiological needs, safety needs, belongingness and love needs, esteem needs and self-actualization. The web is the tool to satisfy several types of needs. Satisfaction in psychology corresponds to the level of contentment with work that was done. In the context of Web, it reflects the level of enjoyment of obtained information or services. The level of satisfaction is generally mostly influenced by the source of information.

Because the Web provides a large amount of information, human (user, customer) can not explore every possibility to find the best one. The time limitation and information overload are two main causes. Nowadays in the context of Web personalization, the providers are trying to increase customers' satisfaction to the highest possible level. One of the means to increase user satisfaction is recommendation [146].

The recommendation systems in compliance with [122] focused on individual persons, that do not have enough experience to choose the best alternative from provided ones. The recommendation is one of the means to support the decision-making process (e.g. film selection, music selection or product buying). There are two basic recommendation approaches - non-personalized and personalized recommendation. The non-personalized recommendation is based on most popular items and personalized recommendation focuses on individual users' preferences. The result of recommendation is presented as a list of ranked items. This list is created based on user preferences [124].

There are several methods of recommendation that are based on human trust to group opinion, opinion of closest persons or comparison with well-known products [100]. The most well-known recommendation method includes collaborative filtering and content-based recommendation. With the combination of this methods, the hybrid recommendation is formed. Burke in [29] defined six basic recommendation approaches that contain: collaborative filtering, content-based recommendation, knowledge-based recommendation, demography-based recommendation, community-based recommendation (or social-recommendation) and utility-based recommendation.

The prime idea of collaborative filtering was based on a recommendation of items, that users with similar taste liked in past [125]. There are two basic collaborative filtering types - memory-based and model-based collaborative filtering [129]. Memory-based collaborative filtering uses a statistical method to find the group of most similar users or items (so-called neighbours). In general, it can be divided into user-based (user ratings are estimated in accordance to ratings of other users) and item-based (user ratings are estimated based on user previous ratings to similar items - items are similar if more users rated them in the same way) collaborative filtering. Model-based collaborative filtering generates the user rating model to predict the future user ratings. This model is often created using machine learning techniques.

The main thought of content-based recommendation is a recommendation of items that correspond to user preferences [116], where the item similarity is computed on items characteristics (e.g. genre, keywords, text similarity and so on). The result of recommendation represents the top N most sim-

ilar items - similar to items that user positively rate in the past. Knowledge-based recommendation use domain knowledge to compare item traits with user preference [147]. This recommendation approach is based on information obtained from user web behaviour, user and item characteristics and information provided by domain experts.

In the demography-based recommendation, the user similarity is computed based on selected demography characteristics. Besides typical demography characteristics, it can work with information such as the presence of a pet in the family [81]. Community-based recommendation work with user friends preferences [19]. Presumption of this recommendation approach is the trust to friends more than to other users - similar but unknown [139]. Utility-based recommendation works with a utility function that is domain specific. In specific domains it can outperform other recommendation approaches [67].

The number of research thesis about recommendation improvement increased in the last decade. The recommendation improvement can be performed via different approaches including recommendation methods selection, the recommendation hyper-parameter tuning, text-mining, web-mining, user modelling or user profile design. Kopeinik et al. on in [80] discussed a set of six recommendation algorithms in the domain of e-learning. The evaluation using six different datasets showed the hybrid recommender (that contains from cognitive-based recommender and popularity-based recommender) as the best.

Similar studies were in the domain of e-commerce. Paraschakis et al. in [112] focused on a comparison of several types of recommenders in the domain of e-commerce. Results of their work showed that traditional recommenders are better than complex recommendation algorithms. The best results achieved the kNN collaborative filtering and association rules mining (described in [119]). The comparison of two recommendation approaches - Most frequent item and Associaton rules based recommendation, in the e-commerce was also realized by Sarwar et al. in [129], where the best results achieved Association rules-based method with the Center-based neighbourhood. The comparison of collaborative filtering method using 3 e-commerce datasets (food, fashion and films) was performed by Huang et al. in [69], where the best result for 2 datasets (food and fashion) obtained algorithm Link-analysis and for Film domain dataset item-based collaborative filtering. The Link-analysis is a type of collaborative filtering that is based on Hypertext-Induced Topic Selection [70].

Another approach to recommendation improvement is a study of data-mining methods selection. This approach was discussed by Gibert et al. in [52], where the selection of one from 27 described data-mining methods was explored. Their proposal also includes some kind of recommendation that recommends data-mining methods in accordance with the context of usage. The recommendation result was evaluated via expert assessment. Segrera and Moreno in [132] focused on usage of classifiers in the context of recommendation. They analyze followed types of classifiers: Bayesian classifier, kNN classifier and Decision Three classifier. Besides individual classifiers they also focused on multi-classifiers created via bagging, boosting and stacking approaches. From the mentioned, the best results achieved stacking.

Another approach to recommendation improvement is based on additional information supplemen-

tation (via user behaviour analysis). The information supplementation corresponds to different methods including prediction of user behaviour, user behavioural pattern mining or user modelling. These individual methods can be combined or followed each other. Grbovic et al. in [56] focused on prediction of a user following purchases based on user shopping patterns. This work was an inspiration for [79], where the prediction of time and price were complemented. In the comparison with random prediction, suggested method achieved significantly better results - RMSE 0.3806 for price prediction and RMSE 0.4272 for time prediction.

The user behavioural analysis was the key issue of [149], where the local customers and visitor was distinguished. The experimental results showed that local customers focus on discounts and visitors on price, time of usage and type of products. The study of demography characteristics' impact on shopping behaviour [108] showed that the importance of product attributes differ from males and females. For females were the most important factors price and discount and for males, it was price and product category. Besides demography characteristics, the user behaviour is also influenced by user motivation. The e-commerce customer motivation was explored by Fronimos and Kourouthanassis in [48], where the customer motivation types were suggested. Based on analysis of 45 research thesis the five basic motivation types of the customer were created. This motivation types are followed: apathetic, convenience, enthusiast, traditional and hedonic customer.

In view of the above, the success of recommendation is influenced by several factors including data we can work with (user data and environment data), methods to data interpretation and selected recommendation methods. In our work, we focus on data processing for recommendation to bring new information with the potential to increase the success of recommender and in this way to increase the user satisfaction level. This information is obtained via machine learning methods that are used for user behavioural analysis. We focus on user modelling. In psychological theory [34], the person can be described as a set of persons' characteristics, traits and features. This idea is used in the domain of user modelling in the online environment, where the user model is usually created as a set of persons' characteristics [41]. The human characteristics such as user demography, personality traits and user environment are ones of the factors that influence the user behaviour [12]. This fact is used in user characteristic predictions based on user information behaviour.

We suggested a user model based on a combination of user characteristics and user behavioural patterns in Web to increase the success of recommendation. User characteristics, in general, include demographic characteristics, domain characteristics, web literacy, personality traits, the user goals, interests or motivation. One way to user characteristics prediction is based on user behaviour in the monitored environment. We are mainly dedicated to demography characteristic prediction and user modelling based on the combination of user behaviour and user characteristics.

Our work is dedicated to user modelling in the context of Web. Primary we focused on e-commerce. We analyze user footprints data obtained from commerce projects. The proposal evaluation was realized using the dataset from Slovak e-commerce company ZlavaDna[1]. This dataset was rich in different data types that include: transaction data, user click-stream data, user demography, products textual description and customers communication with the call centre. Those e-commerce

data was from the period of five years. Additional information was achieved via questionnaires provided on the e-commerce website.

Part of our work is associated with project HIBER that goal is a better interpretation of user footprints in the Web environment [21, 22]. Project APVV-15-0508, HIBER - Human Information Behaviour in the Digital Space is based on interdisciplinary cooperation between FIIT STU and FiF UK. With this cooperation, we have created the questionnaire focused on different user characteristics. This questionnaire was published on ZlavaDna website and was completed by 4 443 users. The questionnaire textual content and questionnaire evaluation algorithm were made by Department of psychology FiF UK. The questionnaire was presented in the form of e-commerce deal that was propagated through newsletters. Design of questionnaire evaluation was made with help of ZlavaDna graphic team (marketing, web developers). The questionnaire with its evaluation was created as the independent module that was deployed on Crowdex[2](system created on FIIT STU to manage the micro-work during experiments).

1.1 Open research problems and goals

As we mentioned, the success of recommendation is influenced by several factors including data we can work with. However, the available dataset may not have enough information and that's the space for our work. The main goal of our work is, therefore, the answer a question:

RQ Main: How do we need to preprocess e-commerce data to get additional information for the recommendation?

Since this question is quite general and extensive, we look at it from multiple perspectives.

In general, there are three types of data in e-commerce datasets - transaction data, activity data and entity profiles (including user profiles and item profiles). Transaction data and activity data are usually recorded in the form of click-stream. In the first step, we, therefore, focused on click-stream data with which each e-commerce work. Our goal was to find the best abstraction method for using this kind of data in the machine learning tasks. The main research question of the first step was:

RQA: What is the best abstraction method for click-stream data to use this higher-level representation as input into the machine learning tasks?

As research has shown, the demography is one of the factors that influence people's decision-making [108] (which is responsible for the results of the recommendation). At the same time, one of the main types of recommendation is the demographic recommendation [12, 81]. That is why we in our second step focused on demographic characteristics. Our second research question was:

RQB: How to deal with feature engineering to effectively predict demographic characteristics?

[1]https://www.zlavadna.sk/
[2]https://crowdex.fiit.stuba.sk/

There are many recommendation types and most of them are based on user model [27, 160]. User modelling is, therefore, one of the most discussed themes in the context of recommendation. The last step of our research is focused on user modelling and exploration of user model characteristics. Our third research question was:

RQC: How the components of complex user model influence the result of recommendation?

1.2 Thesis structure

Our thesis is structured into three chapters and conclusion. Individual chapters were written as independent parts and that is why they have some common parts (e.g. dataset description or description of methods that are used in two or all three chapters). The reason for writing chapters as independent parts was our aim to use these chapters as a basis for publication of the results in the form of three research papers.

The main part of this thesis is written in English. However, since its previous versions (Diploma Project I and Diploma project II) were written in the Slovak language, some annexes are written in the Slovak language.

The chapters of our works are as follows:

User Feedback in E-commerce: Events Abstraction for Demography Prediction. This chapter is focused on the problem of event abstraction that has potential to become an effective source of information for machine learning tasks. In this chapter, we propose a method of event abstraction for machine learning tasks. This method is dedicated to domains with a textual representation of items (e.g. e-commerce, e-banks or journals). The abstraction method consists of two general parts – an abstraction of items and an abstraction of events. The abstraction of items is topic modelling problem that is connected with text prepossessing issue. The event abstraction is based on the pattern-recognition method that consists of association rules mining, sequence patterns mining and N-grams mining. We evaluate our abstraction method on e-commerce dataset. The evaluation was performed in two steps – the evaluation of item abstraction (comparison with explicit categories designed by a domain expert in the three machine learning tasks: association rules mining, sequence pattern mining and gender prediction) and the evaluation of event abstraction (comparison of our three pattern mining methods in the task of gender prediction).

The concept of item abstraction evaluation via association rules mining and sequence pattern mining was the result of cooperation with classmate Zuzana Bobotová and consultations with Dr. Michal Kompan within the Knowledge Discovery course. The evaluation itself, the experimentation as well as the chapter text are the result of our separate work.

Demography Prediction in E-commerce. This chapter discussed the problem of demography prediction in the domain of e-commerce. In this chapter, we propose a user model for demography prediction tasks. Our model is suggested for e-commerce domain and consists of 6 types of features –

transaction-based features, temporal-based features, rating-based features, domain-based features, technical-based features and activity-based features. We evaluated our model via prediction of 4 demography characteristics – net income, partner relationship, a child in family and region of user residence. The evaluation was performed in two steps – the evaluation of contribution of individual types of features to demography prediction task (comparison of transaction-based features with combination of transaction-based features and temporal-based features/activity-based features/domain-based features in the 4 demography prediction tasks) and the evaluation of our user model in comparison with other state-of-the-art models suggested for demography prediction in the domain of e-commerce.

User Model for Recommendation in E-commerce. The last chapter focus on the problem of user modelling in the domain of e-commerce. In this chapter, the user model for a recommendation was suggested. This user model is based on transaction data and user click-stream data that are available in the domain of e-commerce. The suggested user model is defined as triplets that includes user preferences, user traits and patterns in user behaviour. We evaluate our user model via item recommendation task. The evaluation was performed in two steps – the evaluation of individual parts of our user model (the impact of these characteristics on the results of the recommendation) and the evaluation of personality-based characteristics for which we focused more closely (the impact of personality-based characteristics in combination with other characteristic types on the result of the recommendation).

2 User Feedback in E-commerce: Event Abstraction for Demography Prediction

The Web provides a large number of services aimed often to information presentation or providing useful means for fulfilling various needs e.g. by shopping. Customers use them to satisfy their needs. The process of satisfaction is generally connected with the activity of customer in a service. The activity of a service user is obviously recorded in the form of events or logs.

Events are defined as any users' feedback recorded in the system (e.g. service). Events consist of implicit and explicit feedback generated by the user. They may include users' ratings, transactions, page crawling or users' eye movement. Events correspond to unique situations that are connected to certain time-stamp and certain user. They are often linked to the certain domain-specific item (e.g. product in the e-shop, a genre in movie service, learning object in e-learning system). However, in some cases, connection to the domain-specific item can be missing (e.g. view of the homepage, search by query or usage of navigation objects).

An item in our work is defined as a basic unit of web service offer. Item is domain specific. It corresponds to deal in e-commerce, a movie in a movie database, paper in newspapers or insurance in an insurance company and so on.

Events are a low-level representation of activity [85]. They are often associated with certain problems. The main problem is that users can generate a large number of different events (e.g. they can visit many different web pages or products). Events are also time-specific. The number of events and time-specification are the reason, why they cannot be directly used for machine learning tasks such as pattern recognition or user classification. These machine learning tasks are based on rules recognition (rules in a form of patterns in pattern recognition task and rules in a form of features' combination in classification task) that are common across overall dataset or across samples in a certain category (in classification task). Events as unique records, therefore, cannot be directly used for search of common rules. This is the reason for a creation of a high-level representation of events in the process of event abstraction [102]. The process of event abstraction is based on a description of events via their common features (e.g. categories). The number of these features is smaller (in comparison with the number of events) and, therefore, it can be used for identification of more usable rules.

There are two main ways how event abstraction can be performed - focusing on items that are connected to a certain event and focusing on events themselves. Item-based abstraction is designed as categorization problem of low-level items into high-level categories. In the context of textual representation of items (which we are focusing on), the problem of categorization is perceived as text mining issue) [11] (it can be solved by different text mining approaches). The event-based abstraction is based on a generalization of events. The process of generalization is performed by choosing of representative events or set of events from overall dataset [102]. The main disadvantage of item-based abstraction are events that are not directly connected with a domain-specific item (e.g. view of the homepage) - those events cannot be abstracted by item-based abstraction. On the other hand, the main disadvantage of event-based abstraction is the uniqueness of events -

found events or set of events are still too unique to be used in machine learning tasks (they can not generate usable rules - the individual rules are connected with a very little amount of events). The solution of this problem is a combination of these two approaches.

Abstraction of items in the form of categorization is an essential part of every larger web service. Offers in e-commerce are divided into categories based on their nature and their parameters; music, films or books are divided into genres and products in e-bank are divided based on their usage. However, in some cases, the item can be associated with multiple categories without identification of the main category (It is potential problem in classification task, where we categorize items into exactly one class.). Another problem in the context of machine learning is the discriminatory power of explicit categories that can be deficient - this is the base for our first research question, in which we assume that explicit categories created by domain expert provide the best distribution of themes across dataset from the view of the real user.

RQ1: Can latent categories achieve comparable results with explicit categories in machine learning tasks (e.g. pattern mining, classification or clustering)?

The problem of event abstraction is up to date as evidenced by the number of papers focusing on this issue which have been written in the recent years. With the increase of the data generated through Web, the increase of need for knowledge obtained from data was recorded. Event abstraction as the process of knowledge discovering is, therefore, necessary. Event abstraction is nowadays applied in the domains such as networking, web servers, medicine tools or eye-tracking. We often encounter with using event abstraction without its explicit definition [43] (this is typical for different machine learning tasks e.g. classification, where event abstraction is represented as part of feature engineering task). As we can see in 2.1, the most usable abstraction method is based on the finding of pairs, triplets or n-grams.

Our method of event abstraction is based on pattern recognition. Pattern recognition is performed in three ways - sequence pattern mining, association rules mining and n-grams mining. The last step of our method is the choice of their best combination considering to particular machine learning task. The comparison of these techniques to each other is the place for next research question.

RQ2: What is the optimal combination of pattern miming methods (association rules mining, sequence pattern mining and n-grams mining) for demography prediction task?

In this chapter of our work, we present following contributions:

- exploring the automatic category obtaining to determine the properties of items
- proposal of the events' representation in the form of high-level patterns
- exploring the different pattern mining method efficiency in the task of demography prediction

2.1 Related work in field of event abstraction

The solutions of event abstraction can be generally divided into two groups – item-based abstraction and event-based abstraction. The item-based abstraction is defined as text mining problem. The

text mining is a part of data mining field that rapidly increased in the recent years [59]. The main goal of text mining field is to analyze information to discover hidden patterns (including hidden trends and outliers) [11]. Despite that, text mining consists of many issues. Classification, clustering and topic modelling are the most appropriate solutions for item-based abstraction.

2.1.1 Item-based abstraction based on text categorization

The problem of text categorization is defined as a problem of choosing the best class/classes from a set of predefined classes [73, 165]. The choice of class/classes is based on classification algorithms such as Naive Bayes, Decision Tree, K-Nearest Neighbourhood, Support Vector Machine (SVM) or Maximum Entropy that have demonstrated significant results in the domain of text classification [137]. The main types of text classification by Aggarwal et al. [11] are decision tree classifier, rule-based classifier, probabilistic classifier (e.g. Naive Bayes classifier), linear classifier (e.g. Support Vector Machine and Neural Network classifier), proximity-based classifier and meta-algorithm classifier (e.g. ensemble classifier). Text classification is generally usable for e-mail or news filtering [89, 131], language identification [30, 144], genre classification [90, 142], sentiment analysis, article selection, document organization and retrieval [2, 131], opinion mining, e-mail classification and spam filtering [31, 33, 91, 135] or recommendation [104]. The main disadvantage of the text classification is the need for labelled training dataset that is not always available.

The choice of optimal machine learning algorithm for text categorization was a key idea of several papers. Yang in [165] focused on a comparative evaluation of few classifiers, where KNN, LLSF and WORD were used as a baseline. The best results obtained KNN, LLSF and NNet methods. These algorithms were followed by CLASSI, DTree, SWAP-1, CHARADE and RIPPER algorithm. The worst performance obtained NaiveBayes classifier and Rocchio algorithm. Joachims in [73] compared Support Vector Machine classifier with other well-known classifiers (Naive Bayes classifier, Rocchio algorithm, C4.5 method and KNN classifier). The experiment showed SVM as best performing classifier in the group of compared classifiers – the SVM classifier achieved the most robust behaviour. Yang and Liu in [166] performed comparative study across five text categorization method: SVM, kNN, NNet, LLSF and NB classifier - the methods were compared based on robustness. Their results showed that SVM, LLSF and KNN significantly outperformed NB and NNet in the case of the small number of positive training instances per category. Sebastiani in [131] discuss three text classification issues: document representation, classifier construction and evaluation of classification. The comparison of machine learning algorithms for text classification was also the main idea of Pawar and Gawandes' work [114]. Their work was dedicated to comparison of six classification algorithms: K Nearest Neighbor, Rocchios' Algorithm, Decision Trees, Naïve Bayes Algorithm, Back propagation Network and Support Vector Machine. In the comparison of supervised machine learning algorithms, SVM classifier was recognized as the most effective classifier.

Recent research in the field of text classification is dedicated to problematic languages classification [58, 169], graph-based categorization [58, 101, 155], improvement of existing algorithms [47], feature engineering [84, 92, 169], transfer of classifiers [9] or multi-level text categorization [57]. Zhu

et al. in [169] are focused on feature selection methods - they proposed a new method of feature selection based on IG algorithm. This method was evaluated on Chinese text classification corpus, where obtained better results than simple IG algorithm. Li in [92] also focused on feature selection problem that was solved by class-based and importance weighted document frequency. Preliminary experiments shown effectiveness of suggested method. The problem of feature engineering was key idea of [135], where Chi-Square feature ranking method obtained better results in comparison with other state-of-the-art solutions. Malliaros and Skianis in [101] proposed a new graph-based approach for text categorization, where each document is represented by a graph that corresponds to the relationships between the different terms. The results of work showed that the new graph-based approach outperforms existing approaches. The graph-based representation was also used by Hadni and Gouiouez in [58] for Arabic text categorization, where a novel method of the graph-based approach using BabelNet knowledge resource was developed. Fragos and Skourlas in [46, 47] focused on K-Nearest Neighbor algorithm, that was extended to lf-igf KNN developed for medical articles classification. The problem of multi-label text classification was discussed in [3], where ensemble learning using convolution and recurrent neural networks was proposed and evaluated.

2.1.2 Item-based abstraction based on document clustering

Document clustering is a method of unsupervised learning that creates groups of similar text documents without training dataset. The similarity of an object in the group is expressed by a similarity function. Hierarchical clustering and K-means algorithms are two widely used algorithms for document clustering [143]. In general, we can divide text clustering algorithms into few groups [11]: distance-based algorithms (they determine similarity of objects through similarity functions), word and phrase-based clustering (they found set of important words that are used for cluster searching), probabilistic document clustering and topic modelling (based on probabilistic model creation across all documents in the corpus - it is not only clustering method) and semi-supervised clustering [18, 107]. Distance-based clustering includes agglomerative and hierarchical clustering algorithms (based on incremental merging or dividing of clusters based on their similarity with one another), distance-based partitioning algorithms (e.g. k-means or k-medoids) and hybrid approaches. Word and phrase-based clustering consist of: clustering with frequent word patterns, leveraging word clusters for document clusters, co-clustering words and documents, clustering with frequent phrases. Document clustering is often used to search for similar documents (e.g. articles in newspapers or ratings in e-commerce) and implicit subject identification (domain of e-mails and articles). Other application domains are document organization and browsing [36, 42], corpus summarization [17, 126] and document classification (used for improvement of the quality in supervised learning algorithm) [17, 107]. In the context of item-abstraction, undefined cluster name is a fundamental weakness of this approach. One of the possible solutions is cluster representation based on common words or keywords.

Recent research in the context of text clustering focused on clustering of short text documents [10, 94, 121, 133], feature engineering [13], improvement of clustering algorithms [159], distributed and parallel text clustering [88, 162], semantic graph-based approaches [16, 72] or interactive cluster-

ing [109]. The comparison of clustering algorithms for short text documents was main idea of [121], where K-means, SVD and graph-based approaches were compared - the results showed that graph-based algorithms perform best. The problem of short text documents was also solved by Seifzadeh et al. in [133], where the statistical semantic approach was proposed. Feature engineering was a key issue of research in [13], where citation-specific features in the context of scientistific texts clustering were presented. This citation-based representation of documents in experiments outperforms the full-text clustering approach in two scientistific journal datasets. Xiong et al. in [159] suggested improvement of K-means algorithm in the domain of text clustering by optimization method of initial cluster centres. The optimization of initial cluster centres caused improvement in the accuracy and stability of clustering. Bai and Jin in [16] suggested a semantic graph-based structure for text representation that has a potential to optimize similarity calculation. The results confirm the significant improvement of accuracy in the context of Chinese text clustering.

2.1.3 Item-based abstraction based on topic modelling

Topic modelling is a statistical model that was developed for extracting of latent topics from the collection of documents [23]. This approach is often confused with document clustering. The main difference is a type of clustering. While document clustering is based on so-called hard clustering (each document is associated with one cluster), topic modelling stand on soft clustering (each document has probabilities across all clusters) [11]. The topic modelling is not strictly clustering approach, it is very often used for dimensionality reduction problem. Latent Semantic Analysis, Latent Dirichlet Allocation, Hierarchical Dirichlet Process or Non-negative Matrix Factorization are algorithms of topic modelling that are often used in domains of journals, newspapers and websites [40, 63]. Wang and Blei in [148] used statistical topic modelling as a part of collaborative topic modelling - the method developed as a combination of statistical topic modelling and traditional collaborative filtering for research papers recommendation. Results of this work showed that usage of statistical topic modelling can improve traditional recommender algorithms. The problem of user similarity computation, that is an essential part of recommendation was discussed in [141] where a method of user interest matrix creation through topic modelling approach was suggested. The domain of micro-blocks in Twitter was examined by Hong and Davison in [63] which focused on a problem of classification using Latent Dirichlet Allocation.

Research papers dedicated to topic modelling in the last years mainly focused on issues such as: developing of new topic modelling approaches [14, 62], hierarchical topic modelling [138], discovering of author interests [161], multimedia topic modelling [151], scalability and stability of algorithms [164, 167] or feature selection [5]. Arora et al. in [14] presented new topic modelling approach with provable guarantees that is more practical than other provable approaches. The proposed approach had a potential to become effective and robust to violations. The experiments' results showed that suggested approach achieve comparable results to the best MCMC approaches. The question of LDA interpretability was discussed in [62] where WikiLDA approach was presented. WikiLDA consists of two general components - finding the most relevant Wikipedia concepts for documents in corpus and usage of Generalized Pólya Urn (GPU) to involve semantic relatedness into the process

of LDA. The text classification most improved via WikiLDA in the domains with hard separable classes. The problem of author interest identification was key issue of [138], where interest drift model was presented. Suggested model thanks to the sensitivity to the ordering of words in texts achieve better results than other state-of-the-art topic models. Chen et al. in [5] focused on the problem of feature selection for LDA. Since the LDA does not directly consider feature selection, input feature selection in the form of genetic algorithm was proposed. This approach improves the F1 score by 0.76

2.1.4 Item-based abstraction based on text mining algorithm combination

Researchers in the last years also have been focusing on the combination of this approaches (combination of text classification, document clustering a topic modelling). The reason for the creation of combined methods was a development of an effective method that has a potential to benefit from advantages of individual approaches. Chen and Zhang in [6] proposed a method of LDA-KNN that improves the similarity calculation in classification issue. LDA-KNN does not compute similarity only between feature words, but it also includes semantic similarity. Semantic similarity is computed using Latent Dirichlet Allocation. This method was compared with other state-of-the-art methods based on a combination of classification and topic modelling and suggested method achieved the best performance in automatic text classification on Chinese dataset. The weakness of this approach is its time efficiency. The combination of text classification and topic modelling algorithm LDA was also designed in [96], where Liu et al. suggested method combined LDA for feature extraction and SVM for text classification. Suggested method achieved better performance and shorten the training time of classification. Integration of document clustering and topic modelling was key idea of [158]. The integration was designed as multi-grain clustering topic model that includes two components - mixture component (discovering of latent groups) and topic model component (mining multi-grains topics). The experiment results confirm the effectiveness of the proposed model.

2.1.5 Event-based abstraction

Categories created in item-based abstraction can be subsequently used for the event-based abstraction. However, abstraction of items is not always necessary. Domains with a small number of items, permanent items and general items do not require item abstraction. The main intention of event-based abstraction is the generation of a more general representation of events with higher discriminatory force. One of the first need for event abstraction was connected with system debugging issue [85, 86]. The debugging of the distributed application using event abstraction was suggested by Kunz in [85, 86]. The abstraction was designed as a grouping of low-level event sets into one high-level event. Due to that idea, the two event set structures was proposed - complete precedence abstractions and contractions. The event abstraction was lately also used for including timing properties as extension of programming language [64], abstraction of continuous systems [51], sensor data integration [97], live-event modeling [127] or process mining [102]. All these applications of event abstraction require reduction of low-level events with conservation of knowledge included in

the original events. The main difference between this applications and applications in machine learning approaches such as classification is the need for discriminatory force.

The event abstraction is an essential part of machine learning approaches that work with data in event form. However, it is often presented as a part of feature engineering, rather than an individual issue. Duong at all. [43] proposed session abstraction in the form of n-grams that was presented as part of feature engineering for gender prediction task. The event abstraction is often based on general event abstraction methods that are also applicable in other domains (e.g. process mining or sensor data processing). However, the impact of chosen event abstraction method on the discriminatory force in machine learning task is not very discussed. One of the widely used event abstraction techniques is event clustering and finding of sub-sequences (in the form of pairs, triplets or n-grams). Besides this, the new event abstraction method was developed. Tax et al. in [145] suggested a method of event abstraction based on a generation of feature vector representations using XES extensions. George at all. [50] presented IL-MINER method that is able to discover event patterns without a priori knowledge of event abstractions.

2.2 Our proposal for pattern-based events abstraction

Events abstraction is an essential part of machine learning problems which are focused on data in event form (because of a large number of different events that are not appropriate for machine learning problems). As we mentioned in related work there are some approaches for events abstraction. Events are often associated with certain items typical for a domain – this is the place for item-based abstraction. However, item-based abstraction may not be enough - that is why event-based abstraction was suggested. There is also a potential for usage of categories created in item-based abstraction as an input into the event-based abstraction.

In order to abstract users' events, we propose the method of events abstraction based on the users' behaviour pattern recognition. Our method was designed for domains with the textual representation of items. In the case of usage with other representations, e.g. multimedia, items should be firstly preprocessed to the textual form.

The main goal of our method is a creation of descriptive features (in the form of patterns on latent categories level) for presented event set - these features thanks to the high level of abstraction have potential to be more appropriate for machine learning tasks (in comparison with row events). Our method can be directly used for machine learning tasks such as classification, clustering or pattern mining. In the context of e-commerce, it is generally usable for user segmentation, user modelling, personalized emails, personalized web page, customer purchasing power prediction or prediction of user traits (e.g. demography, personal traits or interests). Our method consists of two parts – item abstraction and event abstraction (see Figure 1). The item abstraction part can be also separately used for re-categorization of offer or improvement of information retrieval.

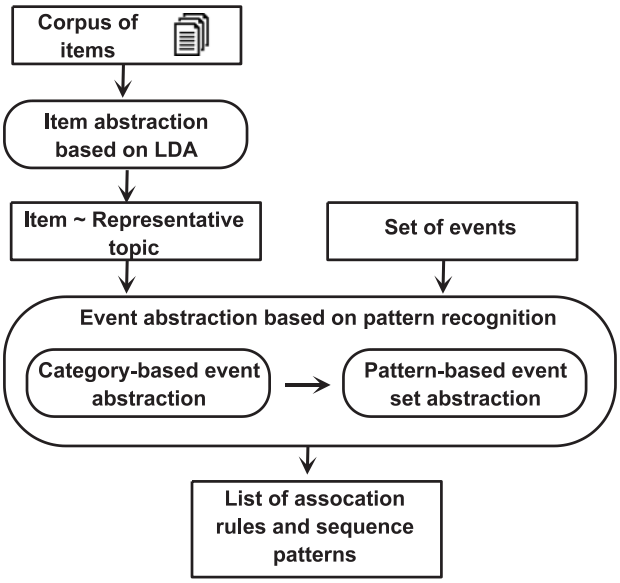

Figure 1: Pattern-based events abstraction method. The input of this method contains a corpus of items and event set. The output is composed of association rules and sequence patterns. The first component, item abstraction, generates latent categories for domain items (see section 2.2.1). Those latent categories are used as input into the event abstraction component. The event abstraction consists of two general parts - category-based event abstraction and pattern-based event set abstraction (see section 2.2.2).

2.2.1 Item abstraction based on Latent Dirichlet Allocation

In the context of web services, there are two main ways to category creation. The first mode is quite a bit more expensive and is based on domain expert assessment – however, people involvement requires more resources. Expert categories are connected with the accurate distribution of categories based on real document themes (themes from the view of a reader). On the other hand, second mode – latent categories, does not require experts involvement. The absence of expert involvement corresponds to cheaper and faster way to category creation. However, the accuracy of this approach is weaker (in the comparison with expert categories). In the problem of discriminatory force (that is essential in classification and pattern mining issues), categories obtained by text mining approaches can achieve comparable results with categories created by domain experts.

As we mentioned in the related works, the problem of category creation is the test mining issue – it can be solved by different text mining approaches. The most usable approaches include text classification, document clustering, topic modelling and their combinations. Given that our method of pattern-based users' events abstraction should be generally usable, it includes item abstraction as topic modelling problem. The text classification is a supervised learning and requires labelled dataset and that is not always available - this is the reason why text classification was not chosen. Document clustering does not require training dataset but it also does not directly support names of clusters. Names of clusters can be very helpful for classification explanation or pattern mining usage for marketing purposes. Topic modelling directly generate names of clusters in the form of representative topics. It also does not require labelled dataset. As we can see in related works, the

topic modelling in comparison with document clustering across multiple experiments often achieved better results. In our method, therefore, categories are perceived as latent categories created by Latent Dirichlet Allocation (LDA) – the most widely used topic modelling algorithm.

The input of LDA is a corpus of items, where each item corresponding to the one text document. This text document is represented as a vector of words or tokens. Before the LDA application, items should be pre-processed to this form.

Pre-processing of items includes:

1. text cleaning
   (a) removal of special characters
   (b) removal of HTML elements (that is not always necessary - but if documents are e.g. product of web crawling it can be important)
2. text tokenization - splitting the text into tokens, where a token represents single unit separated by space (token = word or punctuation mark)
3. token lemmatization - token conversion into the basic form of a word (the word without conjugation, inflexion, prefixes and suffixes)
4. removal of stopwords (frequently used words, that are not important in the context of the corpus)
   (a) domain-specific stopwords (in our application e-commerce specific stopwords e.g. coupon, price or discount)
   (b) general stopwords – stopwords specific for certain language
5. removal of diacritical marks
6. convert to lowercase

Although the pre-processing of texts is relatively easy for the English language (due to existing libraries), it can be quite difficult for other languages. In our application with the Slovak language, we use tools developed by text.fiit group[3] that provide services for Slovak text lemmatization and the list of Slovak stopwords. However, these tools are not standardized – we had to fairly detailed test their functionality (that includes the preparation of test examples). The lemmatization was chosen instead of stemming because of the context maintaining - lemmatization in the Slovak language better maintain the context of the document.

Pre-processed items in the form of the corpus can be used as input into the topic modelling algorithm. How related works showed, one of the widely used topic modelling algorithms is Latent Dirichlet Allocation (LDA). LDA is a three-level Bayesian model that represents an individual collection of items as a final set of topics. Each topic is represented as a set of basic token set probabilities [24]. Our method works exactly with the first most probable topic that represents a category of item. Method of item abstraction is visualized in Figure 2.

[3]http://text.fiit.stuba.sk/

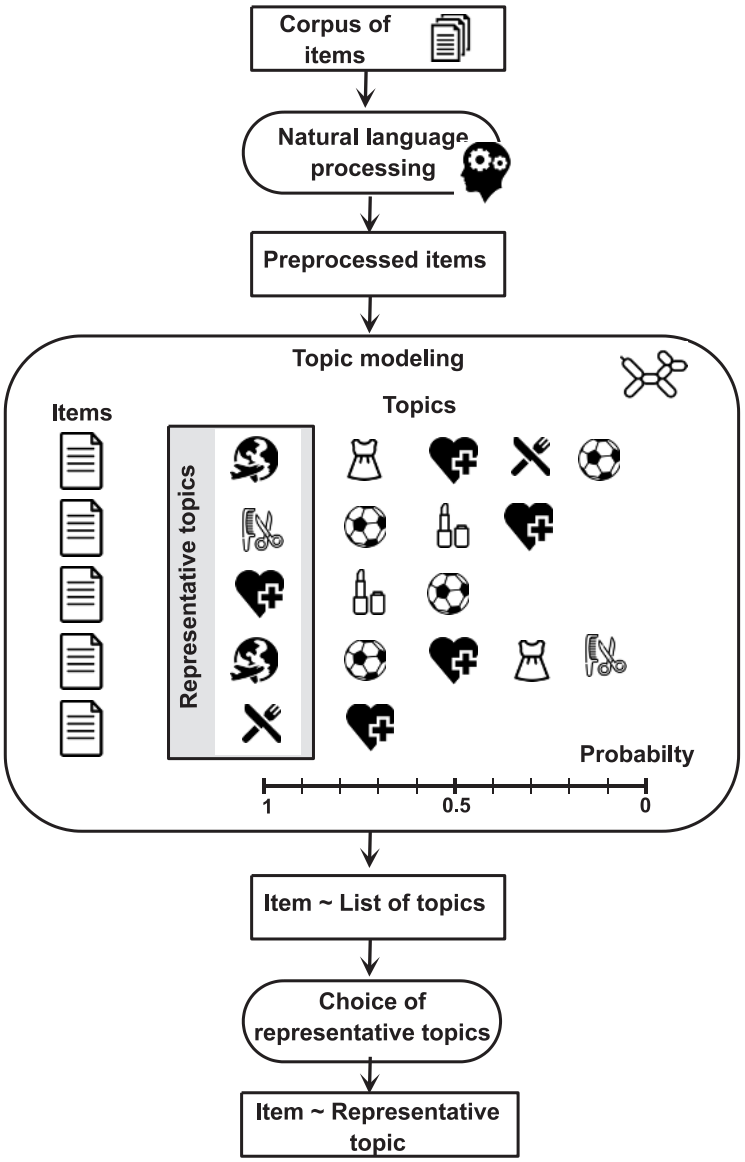

Figure 2: Method of item abstraction based on Latent Dirichlet Allocation. As we can see, the item-based abstraction consists of 3 steps: natural language processing, topic modelling and choice of representative topics. The input of this method is a corpus of items, where each item is represented through the text document. The output of natural language processing step is preprocessed corpus of items. Topic modelling consists of Latent Dirichlet Allocation algorithm that computes probabilities distribution across generated topics. The output of this step is set of items that are represented as lists of topics (that was created via LDA) with their probabilities. The topic corresponds to the latent category of item (e.g. food, sport, travel). The last step of item-based abstraction method is choice of representative topics - the representative topic is topic with the highest probability computed in topic modelling step.

2.2.2 Event abstraction using pattern recognition

The second component of our Pattern-based events abstraction method is event abstraction. Categories created in item abstraction together with low-level events are a basic input of event abstraction. The event abstraction composed of three steps: category-based event abstraction, pattern recognition and the choice of the best pattern combination with respect to the type of machine learning task (see Figure 3). Category-based event abstraction creates a high-level abstraction of individual events using latent categories created in item-based abstraction. Pattern recognition is used for abstraction of event-set. Pattern recognition is not dedicated to the individual events (in comparison with category-based event abstraction), it likely abstracts whole dataset of events - creates a new representation of events in the form of patterns. Abstraction of event-set reduces the number of events and choices the representative events.

Category-based event abstraction considers four types of events:

1. Events associated with exactly one item (e.g. view of item or rating of the item) – abstraction using category of item

2. Events associated with multiple items with dominant class (e.g. view list of items after search) – abstraction using dominant category

3. Events associated with multiple items without dominant class (e.g. purchase) – duplication of the event (up to the level of only one item per event) and abstraction using category of items

4. Events without association for items (e.g. view of a basket) - abstraction using event type

Abstraction of events based on item categories represents an effective way to achieve higher-level events. However, the number of events may be too high - there is the place for event reduction and selection of important patterns. As we mention in related works, there are some approaches for event-set abstraction. Our method includes event-set reduction using pattern-based methods: sequence patterns, association rules and n-grams. Pattern recognition is performed at the session level (set at 25 minutes).

The output of pattern-based event set abstraction step consists of three sets - set of association rules, set of sequence patterns and set of n-grams. The recognized patterns are used in the last abstraction step – in the choice of the best patterns combination due to the machine learning task. The choice of the best combination can be performed in different ways (using different metrics with respect to the chosen machine learning task e.g. classification, clustering or recommendation). In the task of classification, it can be performed via information gain metric. The chosen patterns can be directly used for marketing purposes or after mapping on individual users, they can be used as features in user classification problem.

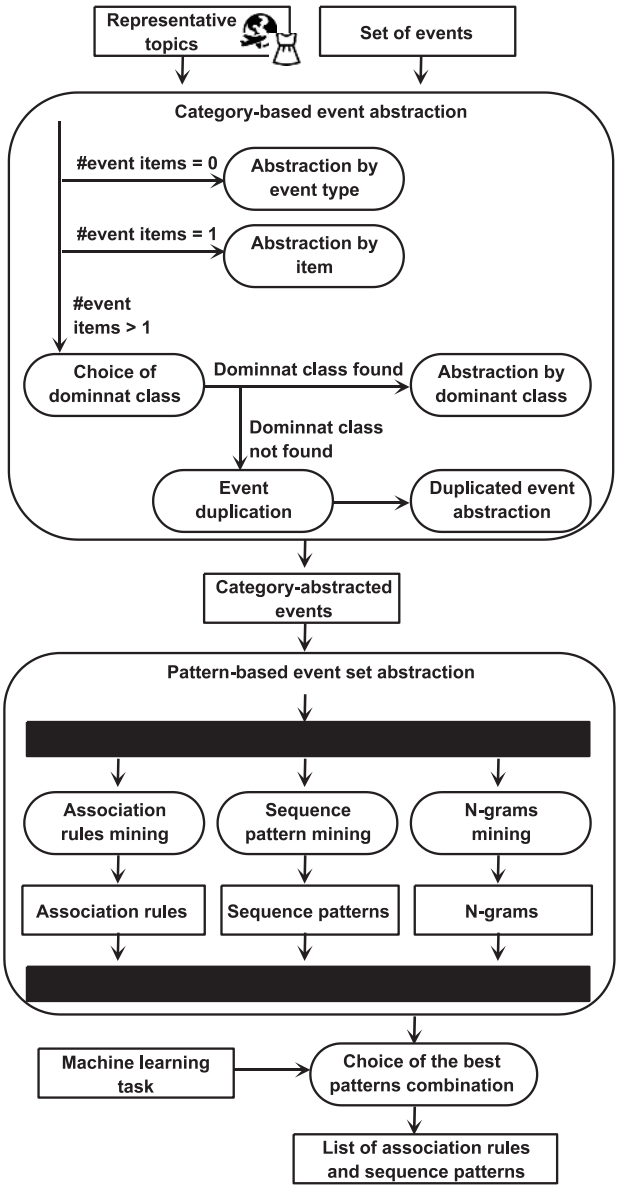

Figure 3: Method of event abstraction based on pattern recognition. The event-based abstraction is composed of three parts - category-based event abstraction, pattern-based event set abstraction and choice of the most representative patterns. The input to category-based event abstraction contains set of events and representative topics generated in item-based abstraction. The category-based event abstraction consists of 4 types of event abstraction based on representative topics - latent categories. The choice of event abstraction type is chosen based on domain items connected with the specific event. The output of category-based event abstraction is set of abstracted events. The next step, pattern-based event set abstraction, contains parallel pattern mining tasks - association rules mining, sequence pattern mining N-grams mining. The last step is choice of the best patterns combination. The input of the last step contains patterns obtained in the pattern-based event set abstraction step and machine learning task for witch our event abstraction method should be used. The output of this method is event set abstracted to the form of patterns.

2.3 Evaluation of Pattern-based event abstraction method

Due to the complexity of the proposed abstraction method, we evaluated our approach in three steps based on our research questions. In the first step, we focused on item abstraction based on Latent Dirichlet Allocation. Latent categories obtained from item abstraction were compared with explicit categories predefined by a domain expert in the task of pattern recognition. These two approaches were compared in the terms of pattern recognition metrics – support and confidence. The support in the evaluation expresses the occurrence frequency of the pattern set within the whole dataset. Support is expressed as Eq1. The confidence is an indicator of the degree of rule truthfulness - it is defined as Eq2. The comparison was performed through top 10 relevant recognized patterns. The relevance of patterns was determined based on specified rules. The evaluation also includes comparison via the number of generated patterns.

$$Support(A = B) = Probability(A \cup B) \tag{1}$$

$$Confidence(A = B) = \frac{Probability(A \cup B)}{Probability(A)} \tag{2}$$

In the second step, we focused on pattern-based event abstraction. We compared our three parallel pattern mining approaches using the gender prediction task. The comparative metrics were precision, recall and F1 score. The task of gender prediction was chosen because n-grams are often used method for event abstraction in demography prediction task [43].

$$Precision = \frac{|\{\text{relevant} \quad \text{documents}\} \cap \{\text{retrieved} \quad \text{documents}\}|}{|\{\text{retrieved} \quad \text{documents}\}|} \tag{3}$$

$$Recall = \frac{|\{\text{relevant} \quad \text{documents}\} \cap \{\text{retrieved} \quad \text{documents}\}|}{|\{\text{relevant} \quad \text{documents}\}|} \tag{4}$$

$$F1 \quad score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

2.3.1 Dataset description

Our work is dedicated to the domain of personalization of services available on the Web, primarily in the domain of e-commerce. The dataset used for evaluation was obtained from Slovak e-commerce company ZlavaDna[4]. However, ZlavaDna is not typical e-shop, more precisely it is discount portal - it offers deals in the form of discount coupons. The purchased coupons are subsequently applied through partners or providers. ZlavaDna is typical with short-time deals and diverse offer. Deals are generally offered during a couple of weeks or months (permanent deals in this domain does not exist). The average time of offer validity is 2.5 week. The time of bought coupon usage is also limited. E-commerce provides a diverse offer that is structured into 6 basic categories - food,

travel, services, goods, health and sports. Individual offers are interconnected exclusively through partners that provide these offers (e.g., one offer of one restaurant). This dataset includes followed data:

- Transaction data (3 million of purchases, 500 000 customers)
- User activity on the website (18.5 million of events)
- Customer general information
- Item general information
- Item categories

Since the ZlavaDna is an active private company, the dataset is not freely available.

Dataset preprocessing. As event part of our dataset was not preprocessed yet, it required some preprocessing steps. The event set preprocessing consisted of four steps:

1. Dataset cleaning
2. Mapping of sessions to signed customers
3. Outlier detection
4. Session complementing

The event set cleaning was necessary because of error and irrelevant events obtained in our dataset. The cleaning of event set included followed operations:

- Choice of the relevant time period and deletion of events outside this period (The choice of relevant period was necessary because some part of event set was collected during test phase that includes a large number of error events).
- Deletion of events without event type (if we do not know the type of event, this event is for our analysis irrelevant).
- Deletion of testing events and events for a recommendation (The purpose of this events is recommender system updating and evaluation - they are not in the centre of our interest).
- Deletion of events without user identifier because of they are inappropriate for machine learning problem such as classification of a user (what is one of our evaluation approaches).
- Deletion of events without connection to domain items - only for events which this connection should include and therefore they are perceived as error events.
- Deletion of deals with less than 5 view events - those deals was not really offered and these events are the output of e-commerce employer work (Deals before real presentation for customers are prepared and controlled by several different employees and their views are also recorded. We assume, then deals with less than 5 views were never published).

The result of event set cleaning was the event set without faulty and irrelevant events - the reduction ratio was 17%.

The second step of the event set preprocessing was a mapping of sessions to signed users. Since we work with dataset provided by e-shop, where sign-up is not necessary, we had to deal with the problem of logged-in and not logged-in users. The events had to be back mapped to the real logged-in user identifier. However, this process was connected with some problems - front-end and back-end events had to be mapped separately (because of their different structure) and there was also some situations in which mapping was not possible (e.g. if user purchased by card and after payment close the bank website - mapping of cookies to real user identifier was not recorded because it was normally performed after return to e-commerce website - 23% of purchases). The next problem of user mapping was events generated by e-commerce employees in the name of real customers - in this case, the event includes an identifier of real customer and cookie of an employee. The solution was removing all events with employees' cookies. The last part of this step was a deletion of events that were not mapped to the real user identifier (events performed by users without historical purchase in e-commerce).

A necessary part of preprocessing is outliers detection. By analyzing our dataset we found out that the distribution of user activity is unequal. For this reason, we decided to remove events from the most active and least active users. The goal of this removing was the deletion of outlier values, mostly generated by e-commerce employees (with high activity in system) and inactive users (with little activity in the system). The deletion of outliers improve the results of our analysis - it avoids distortion. Finally, we removed users with less than 5 events (2.2 million cookies) and users with more than 1000 events (240 cookies). The result of outlier removing was 18.5 million of events generated by 500 000 customers (the original dataset include 24 million of events generated by 2.8 million of cookies).

The last step of the event set preprocessing was session complementing. Because the events did not contain explicit session identifiers, they had to be complemented. The time period of session end after user inactivity was set on 25 minutes (in the domain of e-commerce commonly is used 20 o 30 minutes). The dataset includes 3.3 million of sessions with the average length 4.93 events.

2.3.2 Hyperparameter tuning

The evaluation includes four machine learning algorithms with number of parameters. The parameters' setting was often performed experimentally. This section will focus on the experimental set of parameters for our machine learning algorithms.

Parameters of Latent Dirichlet Allocation. The Latent Dirichlet Allocation (LDA) is the main part of item abstraction. Based on the input corpus of items it generates the set of topics with its distribution across item corpus. The most representative topics correspond to latent categories that are the result of item abstraction part of our method.

[4]https://www.zlavadna.sk/

The most important parameter of LDA is the number of topics to be modelled - this parameter was set experimentally. Other parameters were set as follows: $passes = 20$ a $probability = 0.01$. The number of topics was selected by progressive testing of the topics' number from 50 to 300 with an increase of 20. The evaluation was performed on the basis of the two rules we have defined:

1. The cluster generated by the LDA should contain as few basic categories[5]as possible.
2. The basic category should contain as many LDA clusters as possible.

The result score of parameter setting was calculated by a followed formula:

The LDA setting score

$$Score = \frac{\frac{Avg(TopicDistribution)}{|Topics|} + 1 - \frac{Avg(BasicCategoryDistribution)}{|BasicCategories|}}{2} \tag{6}$$

Where the topic distribution corresponds to the number of topics in the individual basic category and basic category distribution corresponds to the number of basic categories obtained in the certain topic. The result of choosing the best LDA setting was parameters: $passes = 20$, $probability = 0.01$ and $topics = 70$. The result setting indicates that the dataset is best described through 70 categories (despite the fact that e-commerce uses more categories).

Parameters of FP-growth algorithm for association rules mining. The association rules mining was used as evaluation task for first research question (evaluation of item abstraction in comparison with expert categories). It is also a part of our Pattern-based Events abstraction method that was evaluated in the second research question. The mining of association rules was performed via FP-growth algorithm.

The input FP-growth algorithm looks as follows: $\{S_1, S_2, \ldots, S_N\}$, where $S_1, S_2, \ldots, S_N$ are sets of events within the N sessions and S_X corresponds to one session of a particular user U. S_X is a set of unique events within a single session as follows: $S_X = \{E1_{C1}, E2_{C2}, ..., Ei_{Ci}\}$, where $E1_{C1}, E2_{C2}, ..., Ei_{Ci}$ are individual events in a session ($E1, E2, ..., Ei$ represent type of events and $C1, C2, ..., Ci$ are category identifiers). The output of this algorithm is a set of rules that look as follows: $\{E1_{C1}, ..., EM_{CM}\} => \{E1_{C1}, ..., EH_{CH}\}$.

The FP-growth algorithm requires a set of parameters: minimum support and minimum confidence. However, its parameters depend on the input dataset (mostly on their size or number of sets) – that is why they have to be determined experimentally. In general bigger datasets require smaller values of this parameters and on the other hand, smaller datasets bigger values of support and confidence. The choice of optimal parameters was based on the number of generated rules (we require at least several hundreds or thousands of generated rules). In our experiment, we started with $Support_{MIN} = 0.8$ and $Confidence_{MIN} = 0.8$.

[5]The basic categories were obtained from the basic distribution of the web service page (in our case 6 categories).

Table 1: The evaluation of association rules mining with different parameter settings.

MIN support	**MIN confid.**	**Expert categories**			**Latent categories**		
		Number of rules	Average support	Average confid.	Number of rules	Average support	Average confid.
0.05	0.8	0	0	0	0	0	0
0.05	0.5	1	0.124	0.506	3	0.121	0.619
0.04	0.8	0	0	0	0	0	0
0.04	0.5	5	0.095	0.552	5	0.101	0.589
0.03	0.8	0	0	0	1	0.079	0.804
0.03	0.5	9	0.074	0.567	13	0.079	0.614
0.2	0.8	0	0	0	1	0.058	0.804
0.01	0.8	25	0.051	0.573	29	0.058	0.602
0.005	0.8	15	0.015	0.832	34	0.018	0.868
0.005	0.7	103	0.015	0.753	62	0.018	0.811
0.005	0.6	245	0.015	0.694	137	0.018	0.719
0.001	0.8	1615	0.003	0.868	402	0.004	0.901
0.001	0.7	3402	0.003	0.806	741	0.004	0.828
0.001	0.6	4970	0.003	0.757	1449	0.004	0.737
0.001	0.5	6460	0.003	0.709	2297	0.004	0.668
0.0005	0.7	11324	0.002	0.825	1940	0.002	0.828
0.0005	0.6	16060	0.002	0.773	3963	0.002	0.734
0.0005	0.5	21355	0.002	0.717	6335	0.002	0.665
0.0001	0.8	104716	0.0003	0.905	9176	0.0004	0.896
0.0001	0.7	167117	0.0003	0.846	19550	0.0004	0.816
0.0001	0.6	244261	0.0003	0.784	43752	0.0004	0.721
0.0001	0.5	334198	0.0003	0.72	67545	0.0004	0.661

As we can see in Table 1, the lower parameters of minimum support and minimum confidence increase the number of generated rules. However, the decrease of this parameters also caused the decrease of average support and average confidence of found rules. The solution is, therefore, a compromise between the high number of generated rules and expected average support and confidence. The result of experimental settings showed that our optimal parameters were: $Support_{MIN} = 0.005$ or smaller and $Confidence_{MIN} = 0.5 - 0.8$ (see Table 1).

Parameters of Parallel PrefixSpan algorithm for sequence pattern mining. The sequence pattern mining just like association rules mining was used as evaluation task in first research question and also as part of our method in the evaluation of second research question. The sequence pattern mining was performed via Parallel PrefixSpan algorithm.

Mabroukeh a Ezeife in [99] explain how to use sequential pattern mining in the context of web mining. Based on their work, we have proposed input to Parallel PrefixSpan as a set of sequences. Each sequence corresponds to one session and has the following form $S =< E1_{C1}, E2_{C2}, ..., Ei_{Ci} >$, where $E1_{C1}, E2_{C2}, ..., Ei_{Ci}$ are individual events in a time sequence ($E1, E2, ...Ei$ represent a type of events and $C1, C2, ..., Ci$ are category identifiers). In case of event associated with multiple classes without dominant category, the sequence is as follows: $S =< E1_{C1}, E2_{C2}, ..., Ei_{Ci}, (Ei + 1_{Ci+1}, Ei+2_{Ci+2}), Ei+3_{Ci+3}, ..., Ei+m_{Ci+m} >$, where parenthesis $(Ei+1_{Ci+1}, Ei+2_{Ci+2})$ indicates the parallel running events. The output of the Parallel PrefixSpan algorithm is set of sequences looks like $S =< E1_{C1}, E2_{C2}, ..., Ei_{Ci} >$.

Table 2: The evaluation of sequence pattern mining with different parameter settings.

MIN support	**MAX pattern length**	**Expert categories**		**Latent categories**	
		Number of rules	Average support	Number of rules	Average support
0.3	5	0	0	1	0.309
0.2	5	2	0.288	3	0.295
0.1	5	8	0.177	11	0.177
0.05	5	23	0.104	39	0.097
0.05	10	23	0.104	39	0.097
0.01	5	351	0.023	524	0.024
0.01	10	442	0.021	2330	0.016
0.01	15	445	0.021	2977	0.015
0.01	25	445	0.021	2977	0.015
0.005	5	1228	0.011	1584	0.013
0.005	10	2175	0.009	6622	0.01
0.005	15	2270	0.009	41644	0.007
0.005	25	2272	0.009	46669	0.007
0.0005	5	56585	0.001	34709	0.002

The main parameters of this algorithm are minimum support and maximal pattern length. Those parameters depend on the size of the dataset and therefore they had to be set experimentally. The choice of optimal parameters was based on the number of generated patterns (we require at least several hundreds or thousands of generated patterns).

The experimental results show that optimal minimum support is 0.3 or lower. The maximal pattern length in general influences the uniqueness of found patterns and therefore it should be set on lower value (shorter pattern are more general - they are more appropriate for machine learning tasks such as classification). We tried followed values: {5, 10, 15, 25} (see Table 2).

Parameters of Random Forest for gender prediction task. The gender prediction was used as evaluation task in our both research questions. The prediction was realized by Random Forest algorithm. This algorithm includes several parameters that had to be set. However, not every parameter directly influences the result of classification. Some parameters influence the speed and efficiency of the algorithm. In our research, we focused only on the tuning of hyper-parameters that can make our model better. These parameters include the number of estimators, criterion, maximal depth, maximal features and minimum sample split.

The number of estimators corresponds to the number of trees that are generated in the forest. The higher values of trees in general increase the performance of the algorithm and makes the code slower. Max features parameter is the number of features that are considered when the algorithm is looking for the best split. It is the maximum number of features that individual tree can work with. The increase of maximum features in general also increase the performance of Random Forest. However, it is not always true - the increase of maximum features decreases the diversity of individual trees. Therefore, it is very important to find the optimal value of maximum features parameter. Minimum sample split is the minimum number of samples that had to be in the leaf node. Criterion corresponds to the function that measures the quality of split - it can be set to Gini

Table 3: The hyper-parameter tuning of Random Forests' parameters (AR - association rules, SP - sequence patterns, N - N-grams, T - patterns gained via topics, EC - patterns gained through expert categories).

Features	**Parameters**				
	Number of estimators	Criterion	Max depth	Max features	Min samples Split
AR(EC) + SP(EC)	500	gini	25	0.08	2
AR(T)	500	entropy	25	0.05	5
SP(T)	200	gini	25	log2	5
N(T)	500	entropy	25	0.08	3
AR(T) + SP(T)	500	entropy	25	0.1	2
SP(T) + N(T)	500	gini	15	0.1	3
AR(T) + N(T)	500	entropy	25	0.05	3
AR(T) + N(T) + SP(T)	500	entropy	25	0.1	4

impurity or information gain. Maximum depth limits the number of levels in an individual tree - it is the maximum depth of this tree. The default values of this parameters are followed: number of estimators = 10, criterion = Gini, maximum depth = without limitation, maximum features = the square root of the total number of features and minimum samples split = 2.

The hyper-parameter tuning was performed using grid search method and 10-fold cross validation. The evaluation metric was Recall macro. The tried parameters' values was followed: number of estimators = [10, 20, 50, 100, 200, 500], criterion = [gini, entropy], maximum depth = [5, 10, 15, 20, 25, undefined], maximum features = [square, log2, 0.1, 0.08, 0.06, 0.05, 0.04, 0.02] and minimum samples split = [2, 3, 4, 5]. However the values of parameters depend on input dataset, that had to be set separately for each features set. In the Table 3 can be seen result parameter settings after the hyper-parameter tuning.

2.3.3 RQ1: Can latent categories achieve comparable results with explicit categories in machine learning tasks?

The comparison of latent categories with expert categories was performed in three machine learning tasks - association rules miming task, sequence pattern mining task and gender prediction task. In this chapter, we will discuss latent categories performance in this three machine learning tasks.

Latent categories in association rules mining The comparison of latent categories with expert categories was performed in two ways:

1. Comparison via number of generated rules (see Assumption 2).
2. Comparison via support and confidence achieved in 10 relevant rules.

Since the goal of our abstraction method is usage of the abstracted event set for machine learning approaches, we evaluated the number of found association rules (the number of rules should be sufficient to cover the information contained in the dataset). In the first step, we, therefore, compared latent categories with expert categories via the amount of generated rules - we assume

Table 4: Found top 10 relevant rules (letters represent the type of event - V is a view of an item, L is a list of items with specific category and numbers represent categories and -1 is representation for a home page and -2 for an event without dominant class).

	Latent categories	**Expert categories**
1	$[V21, L-2, L-1] => [V50]$	$[V165, V87, V615] => [V57]$
2	$[V21, V15] => [V50]$	$[V1151, V87, V615] => [V57]$
3	$[V16, V21] => [V50]$	$[V1111, V1163] => [V615]$
4	$[V47, V15] => [V50]$	$[V1071, V57, L-1] => [V615]$
5	$[V16, L-2, L-1] => [V50]$	$[V1071, V1111] => [V615]$
6	$[V47, L-2, L-1] => [V50]$	$[V1111, V985] => [V615]$
7	$[V7, L-1] => [V50]$	$[V985, V57, L-1] => [V615]$
8	$[V21, V15, L-1] => [V50]$	$[V985, V1163, L-1] => [V615]$
9	$[V47, V21, L-1] => [V50]$	$[V1155, V57] => [V615]$
10	$[V16, V47] => [V50]$	$[V1151, V87, V615, L-1] => [V57]$

Table 5: Comparison of support and confidence of top 10 relevant association rules.

	Latent categories		**Expert categories**	
	Support	Confidence	Support	Confidence
1	0.020	0.823	0.009	0.813
2	0.013	0.878	0.009	0.802
3	0.012	0.913	0.008	0.818
4	0.011	0.879	0.008	0.807
5	0.011	0.830	0.008	0.826
6	0.010	0.848	0.007	0.855
7	0.009	0.897	0.007	0.842
8	0.008	0.920	0.007	0.813
9	0.007	0.946	0.007	0.848
10	0.006	0.937	0.007	0.827

that bigger number of rules is sufficient (see Table 1). It can be also a prerequisite for a better discriminatory power in the demography prediction task.

The Table 1 shows that latent categories are more appropriate for a smaller number of generated rules, otherwise expert categories generate more number of rules in lower parameter settings.

The second step of evaluation via association rules mining was the comparison of latent categories with expert categories using top relevant rules. The relevance of patterns was defined as follows. The relevant pattern in association rules mining contains at least two different categories of items (without undefined categories). The results of top 10 relevant patterns obtained by association rules mining are presented in Table 4 and the comparison of their metrics are shown in Table 5. The parameters was set as follows: $Support_{MIN} = 0.005$ and $Confidence_{MIN} = 0.8$.

As can be seen in Table 5 latent categories created with our method (item abstraction based on Latent Dirichlet Allocation) generate better rules than expert categories. As we assumed, expert categories are more detail and therefore also their the most popular rules achieved lower values of support and confidence then rules generated via latent categories. The result of comparison through association rules mining is that latent categories generate more efficient rules with higher support and confidence. On the other hand, it generates less amount of patterns that can be important for

Table 6: Found top 10 relevant sequence patterns (letters represent a type of event - V is a view of an item, L is a list of items with specific category and numbers represent categories and -1 is representation for a home page and -2 for an event without dominant class).

	Latent categories	**Expert categories**
1	$[[V50], [V50], [V50]]$	$[[V57], [V615]]$
2	$[[L-1], [V50], [V50]]$	$[[V615], [V57]]$
3	$[[V50], [V21]]$	$[[V57], [V87]]$
4	$[[L-1], [V50], [V50], [V50], [V50]]$	$[[V87], [V615]]$
5	$[[L-2], [V50], [V50], [V50]]$	$[[L615], [L615], [L615]]$
6	$[[V21], [V50]]$	$[[L615], [V615]]$
7	$[[V15], [V50]]$	$[[V615], [V615], [V615], [V615]]$
8	$[[V21], [V50], [V50]]$	$[[V615], [V87]]$
9	$[[V15], [V50], [V50]]$	$[[V1163], [V615]]$
10	$[[V9], [V50]]$	$[[V615], [L615]]$

machine learning task (depends on exact use-case).

Latent categories in sequence pattern mining The comparison of latent categories with expert categories was performed in two ways:

1. Comparison via number of generated rules (see Assumption 2).
2. Comparison via support and confidence achieved in 10 relevant rules.

Because of our goal to use found patterns for machine learning tasks, our evaluation deals not only with the quality of sequence patterns but also with a number of found patterns. For this reason, we looked at the two point of view again. In the first step, we compared latent categories and expert categories via the number of generated patterns - see Table 2. And the second step was dedicated to comparison through top 10 relevant patterns.

As we can see in Table 2, sequence pattern mining using latent categories, in general, generate a large number of patterns (what can be convenient for machine learning tasks). The average support of found patterns is a bit lower than in expert categories.

In the second step, we compared expert categories and latent categories via top 10 relevant patterns. The relevant pattern in this context contains at least two different items (difference in category or event type) or contains the same item at least 3 times. The results of top 10 relevant patterns obtained by sequence pattern mining are presented in Table 6 and the comparison of their support is shown in Table 7. The parameters was set as follows: $Support_{MIN} = 0.015$ and $Length_{MAX} = 5$.

The Table 7 shown that latent categories obtained by item abstraction generate sequence patterns with higher support - these patterns are better quality then patterns obtained via expert categories (As we assumed, expert categories generate more detail patterns with a lower value of support).

Table 7: Support comparison of top 10 relevant sequence patterns.

	Latent categories	**Expert categories**
1	0.094	0.026
2	0.065	0.826
3	0.034	0.019
4	0.033	0.019
5	0.033	0.019
6	0.033	0.017
7	0.027	0.017
8	0.022	0.017
9	0.017	0.016
10	0.017	0.015

Latent categories in gender prediction task The evaluation via gender prediction task was realized as a classification problem (binary classification into two classes - man, woman). The dataset includes 42 792 samples (15 765 women and 27 027 men). The problem of the unbalanced dataset was solved by down-sampling (to the size of the smaller class). The dataset was divided into the training set and testing set in the ratio 80:20. The training dataset was in the first step used for hyper-parameter tuning using grid search method and 10-fold cross validation (for more details see section Parameters of Random Forest). In the evaluation we mainly focused on metric Recall - our metric for hyper-parameter tuning was Recall Macro too.

The comparison was realized via two features sets - the first feature set contains patterns with topics. The second feature set consists of patterns created using expert categories. The number of features in latent category feature set was 5 410 and in expert category set 5 143. Because of a large number of features, we have chosen the most representative features using feature selection based on information gain metric. The limit parameter of feature selection was the importance of a feature more than 0.001. The feature selection reduced the original feature set as followed: latent category feature set - 124 features and expert category feature set 216 features.

As we can see in Table 8, the most important features generated via latent categories are simpler and achieved higher importance (in comparison with the most important feature generated via expert categories). One of the possible reason for this phenomenon is the number of selected features. The expert category feature set contains more features, and probably, therefore, its top features achieve lower importance values. The result of comparison of expert categories and latent categories in gender prediction task was followed: gender prediction using latent categories: precision = 0.71, recall = 0.67, F1 measure = 0.68 and gender prediction using expert categories: precision = 0.71, recall = 0.66, F1 measure = 0.68. The result of this two prediction was almost the same - the discriminatory power of latent categories is comparable to the discriminatory force of categories defined by the domain expert.

To sum it up, as we can see in Table 7 and Table 5 pattern mining in both approaches (association rules mining and sequence pattern mining) generate more quality top patterns using latent categories (obtained by out item-based abstraction method) rather ten expert categories. Expert categories are more detail and generate patterns with lower support and confidence.

Table 8: Top 10 most important features in gender prediction task -comparison of expert categories and latent categories obtained via Latent Dirichlet Allocation (letters represent a type of event - V is a view of item, L is list of items with specific category, Rc is e-commerce referrer, Rfb is Facebook referrer and numbers represent categories).

	Expert categories		**Latent categories**	
	pattern	importance	pattern	importance
1	V271	0.016	V53	0.016
2	V269	0.013	V4	0.016
3	V1007	0.011	L37	0.015
4	V1059	0.011	V0	0.014
5	V471	0.01	V66	0.014
6	V1007 V1007	0.008	V12	0.014
7	Rc B	0.007	V33	0.014
8	L1007	0.007	Rfb	0.013
9	V283	0.007	V37	0.013
10	V87	0.007	L33	0.013

The comparison by the number of generated patterns depends on pattern mining task. In association rules mining a larger number of patterns was generated using expert categories, while in the sequence pattern mining more patterns were generated with latent categories.

The comparison via gender prediction task shown the comparable discriminatory power of latent categories and expert categories. The gender prediction task proved the minimum lost of discriminatory information in comparison with expert categories. The latent categories are, therefore, the suitable solution of category absence problem. It can be also used as an alternative to manual expert categorization that is more expensive (the human expert involvement is very expensive).

2.3.4 RQ2: What is the best combination of pattern miming methods (association rules mining, sequence pattern mining and n-grams mining) for demography prediction task?

Association rules mining, sequence pattern mining and n-grams mining represent basic approaches for event pattern mining. The main purpose of the association rules mining is to find out what events are performed by users within individual sessions. On the other hand, the main goal of sequence pattern mining is the identification of frequent event sequences performed by users. The main goal of N-grams mining is the same as the goal of sequence pattern mining - the identification of frequent event sequences performed by users. The N-grams and sequence patterns are generally very similar. If the whole dataset of sequence patterns is found (the minimum support and confidence are equal to zero), the N-grams are a subset of sequence patterns. However, using parameters - minimum support and minimum confidence in the sequence pattern mining caused that set of N-grams is not a subset of chosen sequence patterns. The using of this limitation parameters can, therefore, generate different pattern sets - this is the reason for using of both approaches.

The main difference between sequence patterns and N-grams is in immediate sequence flow. While sequence pattern mining does not require immediate sequence, the N-grams requires it (e.g. if our session consists of $S =< E1_{C1}, E2_{C2}, E3_{C3} >$ the sequence pattern mining approach can generate

followed patterns: $SequencePatterns = \{< E1_{C1} >, < E2_{C2} >, < E3_{C3} >, < E1_{C1}, E2_{C2} >, < E1_{C1}, E3_{C3} >, < E2_{C2}, E3_{C3} >, < E1_{C1}, E2_{C2}, E3_{C3} >\}$ and N-grams generate: $Ngrams = SequencePatterns$ - $\{< E1_{C1}, E3_{C3} >\}$).

The comparison of pattern mining methods was realized in the same way as in RQ1 evaluation - binary classification into two classes - man, woman; 42 792 samples using down-sampling; train-test in the ration 80:20. The evaluation metric was recall. The hyper-parameter tuning is described in the section Random Forest Parameter.

The first step of the pattern mining methods comparison was the creation of 7 feature sets. They contains patterns created via individual pattern mining methods (association rules mining, sequence pattern mining and n-grams mining) and its combination. The feature selection was based on information gain metric.

The most important features generated by individual pattern mining methods are shown in Table 9. Top important features that are result of the combination of different pattern mining method are presented in Table 10. The Table 9 shown that sequence patterns features and N-grams features are similar - 6 from 10 patterns are identical. The main difference is in the importance of the top features. N-grams' feature, in general, achieved higher values of importance. The association rules generate more complex patterns with lower importance.

Table 9: Top 10 most important features in gender prediction task using stand-alone features datasets - without combination (letters represent a type of event - V is a view of item, L is list of items with specific category, B is view of basket, Rc is e-commerce referrer, Rfb is Facebook referrer and numbers represent categories).

	Association Rules		**Sequence Patterns**		**N-grams**	
	pattern	importance	pattern	importance	pattern	importance
1	L22 V22	0.013	V4	0.016	V53	0.019
2	L33 V45 L45	0.011	V0	0.015	V0	0.017
3	L4 L67 Rc	0.011	V53	0.013	V4	0.017
4	V37 L45 Rc V45	0.01	V22	0.012	V22	0.014
5	V12 V33 V45	0.009	V31	0.009	L37	0.012
6	L37 V45 Rc L45	0.009	V12	0.009	V52	0.011
7	V12 V37 V45	0.009	Rfb	0.009	L4	0.011
8	L0 B V0	0.008	L37	0.009	V66	0.011
9	L32 L45	0.008	Rc B	0.009	V31	0.01
10	V31 L45 V45	0.008	V37	0.009	V33	0.01

In the Table 10 we can see, that combination of pattern mining approaches often generate the top features more complex than individual methods. The top patterns created via the combination of sequence patterns and N-grams are more complex then features created by N-grams and sequence pattern mining separately. On the other hand, the complexity of top features in the context of association rules mining was decreased.

To conclude, the experimental results have shown that sequence pattern and N-grams achieved comparable results in gender prediction task - the recall for both approaches was 67% (see Table 11). Although the association rules mining generates more complex patterns (see Table 9), the results in the gender prediction task were lower - recall = 59% which is 8% less than sequence pattern or

Table 10: Top 10 most important features in gender prediction task using the combination of features created by Association Rules Mining (AR), Sequence Pattern mining (SP) and N-grams mining (N). The letters represent a type of event – V is a view of an item, L is a list of items with the specific category, B is a view of basket, S is a search of specific category, Rc is e-commerce referrer, Rfb is Facebook referrer, Rg is Google referrer and numbers represent categories.

	SP + AR		**SP + N**		**AR + N**		**AR + N + SP**	
	pattern	imp.	pattern	imp.	pattern	imp.	pattern	imp.
1	V0	0.015	B B	0.018	V0	0.02	B B	0.015
2	V4	0.015	L45 B	0.017	V4	0.019	Rc B	0.014
3	V22	0.015	Rc L45	0.017	L37	0.019	Rc L45	0.014
4	L37	0.014	Rc B	0.016	V66	0.018	L45 B	0.014
5	V66	0.014	L45 L45	0.016	V45 V45	0.018	R	0.013
6	V53	0.014	L45 L	0.015	V33	0.018	L L	0.013
7	V12	0.013	L37	0.015	L33	0.018	L45 V45 L45	0.013
8	V37	0.013	Rg V45	0.015	S45	0.017	L37	0.013
9	V33	0.013	L45	0.014	Rc L45	0.017	V45 V45	0.013
10	Rfb	0.0013	V45	0.014	V31	0.017	L37	0.013

N-grams. As we can see in Table 10 the combination of those approaches generate more complex patterns. The combination of sequence patterns and N-grams increased the recall by 3%. The result recall is 70%.

Table 11: The comparison of association rules mining, sequence pattern mining and N-grams in gender prediction task (AR - association rules, SP - sequence patterns, N - N-grams).

Dataset	**Features**	**Selected Features**	**Results**		
			Precision	Recall	F1
AR	3120	244	0.63	0.59	0.6
SP	3551	167	0.71	0.67	0.68
N	2371	140	0.71	0.67	0.68
AR + SP	5410	127	0.71	0.68	0.69
SP + N	5920	125	0.72	0.7	0.69
AR + N	4230	97	0.7	0.68	0.69
AR + N + SP	7779	164	0.7	0.66	0.67

2.4 Summary of event abstraction

Events abstraction is an essential part of many machine learning algorithms. In this chapter of our work, we present method of pattern-based users' events abstraction for domains with textual representation of items. The suggested method consists of two general parts – item abstraction based on Latent Dirichlet Allocation and event abstraction using pattern recognition.

Our method was evaluated using e-commerce dataset. Based on the comparison of top 10 relevant patterns, we can see that the proposed item abstraction method generate more quality top patterns than expert categories. Expert categories are more detailed and generate patterns with lower support and confidence. The comparison via gender prediction task shown the comparable discrim-

inatory power of item abstraction method and expert categories. The latent categories generated in item abstraction method are, therefore, the suitable solution of category absence problem and can be also used as an alternative to manual expert categorization.

Evaluation of event abstraction was based on the comparison of three pattern mining approaches - association rules mining, sequence pattern mining and n-grams. The experimental results have shown that sequence patterns and N-grams achieved comparable results in gender prediction task - the recall for both approaches was 67%. However, the combination of this approaches increased the recall by 3%, what suggests that the combination of these approaches can improve the performance of machine learning tasks.

3 Demography Prediction in E-commerce

Demographic characteristics are the basic attributes that influence a shopping behaviour of people [93, 120, 123]. However, characteristics such as age, gender, education or income belong among the type of data which people do not like to provide. Because the demography is one of the primary sources of information for marketing, targeted advertising and personalization of the Web [37, 81], service providers have to deal with users' reluctance to provide this type of data. There are several approaches to getting this data: different marketing campaigns, rewards in the form of benefits or machine learning techniques.

Our work is focused on demography acquisition in the form of demography prediction task. Demography prediction is perceived as a multi-class multi-task classification problem. In general, there are two main ways, how the classification performance can be improved – focusing on features and focusing on algorithms. Our work is dedicated to feature engineering (based on the feature engineering the our user model was suggested). The main goal of our work is, therefore, a creation of user model that improve the demography classification task. Suggested user model consists of the followed type of features: transaction-based features, activity-based features, temporal-based features, rating-based features, domain-based features and technical-based features. From these characteristics we focused more closely on temporal-based features, activity-based features and domain-based features – that was a place for our three research question.

RQ1: How improve the temporal-based features the classification performance when they are joined to the transaction-based features?

RQ2: What is the improvement of classification based on activity features in the comparison with classification based on transaction features?

RQ3: What is the improvement of classification performance, when domain-based features are added to the transaction-based features?

Since our goal, besides user model suggestion, also includes the evaluation of suggested model performance - we compared our model to other state-of-the-art models proposed for demography prediction task. Specifically, we focused on models created in the domain of e-commerce. Those models were selected from models proposed in the PAKDD 2015 Data Mining Competition. The selected models were: model suggested by Lu et al. in [98] and model proposed by Duong et al. in [43]. Both models in the PAKDD competition have placed in the top 10 solutions.

RQ4: How the suggested user model perform in the comparison with other state-of-the-art user models proposed for the domain of e-commerce?

In this chapter of our work, we present following contributions:

- the novel user model for demography prediction task in the domain of e-commerce
- an analysis of temporal-based features on demography prediction task
- comparison of features obtained from transaction data with the performance of features from

the web activity data

- an analysis of domain-based features on demography classification task
- comparison of suggested model with other state-of-the-art models proposed in the domain of e-commerce

3.1 Related works in demography prediction

In the context of services e.g. e-commerce, there are two main approaches to the demography prediction task. The first approach is based on textual documents created by a user (e.g. blocks, comments, e-mails or reviews). These methods analyze documents and looking for the association between writing style and demography characteristics [118]. However, this group of methods is not limited to writing – it can be also used for spoken words, where the conversational style is analyzed [49, 53]. The second approach is based on the analysis of user behaviour in the Web environment. This group contains methods dedicated to search behaviour [20, 55, 152], Web using [130, 136] or browsing history [65, 74, 118].

3.1.1 Demography prediction based on text documents

Herring et al. in [61] focused on gender prediction task based on the writing style of web blogs. Following this research, Herring and Paolilo in [60] analyzed the variation between gender and genre of web blogs. The domain of web blogs was also analyzed by Chen at all. in [4], where the optimization of KNN algorithm for gender prediction task was suggested. The optimization was based on latent semantic indexing. Bouadjenek et all. in [26] focused on prediction of gender and age via written text analysis. The method was based on gradient characteristics, that was used in the classification task.

The problem of prediction chat message authors' gender was described in [82], where the term-based and style-based classification methods were evaluated. Gender prediction in the context of email stream was discussed in [39], where the stylometric features and a word count features were used in Neural Network Classifier. The domain of e-mails was in the centre of interest in [7] too. The problem of multi-genre, short length, content-free e-mails was solved with psycholinguistic and gender-linked approach.

The main goal of Burger et al. in [28] was gender prediction using analysis of user posts on the social network Twitter. The prediction was based on many characteristics (e.g. post topic, post description or n-grams). The domain of web service Twitter was also analyzed by Pennacchiotti and Popescu [117] who focused on gender prediction task too. The gender prediction task was the key issue of [95], where the Russian text was analyzed. The problem of demography prediction based on spoken style was discussed by Garera and Yarowsky in [53] and Gillick in [53].

3.1.2 Demography prediction based on user behaviour analysis

The second group of methods is based on the analysis of user behaviour in the Web environment. The differences in Web usage behaviour are in the centre of interest for some time. Shaw and Gant in [136] examined the differences in the Internet behaviour of men and women. Despite the fact, that previous research proved the differences in Internet behaviour of men and women (men primarily use the Internet on gathering information and entertainment, while women on interpersonal communication), when all participants performing the same activity those differences were not recorded. During the experiment, the participants were tested for feelings of loneliness, depression and self-esteem. However, the result of this testing did not confirm the gender differences.

Despite the fact that gender differences in the same activity performing were not recorded [136], gender differences in the normal Internet use have been reported [45]. Fallows in [45] summarize recognized differences among men and women. Younger women used the Internet more than younger men. On the other hand older men used the Internet more than older women. Unmarried men used the Internet more likely than unmarried women. Men are generally slightly more frequent users of the Internet (in comparison with women). Men used the internet in the more variations. Women are oriented to the interpersonal communication and men rather search for information and use the Internet for recreation. The differences in behaviour among age groups was also discussed in [55], where the impact of web page design was examined.

Weber et al. in [152, 153] focused on the differences in the web search behaviour (e.g. differences among different income ranges or different ethnic groups). The application of this differences into the process of information retrieval led to a 1.4% increase in P@1 among all searches and increase of 7.1% in P@1 for queries with the larger entropy.

Because our work is a part of this second mode - it mainly focuses on user behaviour, we will now focus in more detail on works dedicated to demography prediction based on the behaviour. Murray and Durrell in [105] described the application of the Latent Semantic Analysis (LSA) technique for users' Internet usage data representation. The LSA vector space was subsequently used as an input to the three-layer neural model that was trained using the scaled conjugate gradient. The result of the work includes prediction of followed demographic characteristics: gender, age, income, marital status, education, children in the home. The experiment showed the minimum 60% statistical confidence across all prediction tasks.

Bi et al. in [20] focused on demography prediction task based on search phrases history. The demography prediction consists of prediction of age, gender and prediction of political and religious opinions. The prediction was performed on MyPrediction dataset that includes demography information and favourite items of users on social network Facebook. The favourite items were replaced by requests through the Open Directory Project. The model was evaluated using private search engine tool, where model in classification achieved followed results: age prediction - 77%, gender prediction - 84% in the metric AUC (the area under the ROC curve) in the classification based on favourite items and age prediction - 74%, gender prediction - 80% in the classification based on the transformation of favourite items into requests.

The main goal of Hu et al. in [65] was demography prediction based on Web usage logs (logs it the form of pairs: user - web page). The prediction was performed in two ways: prediction of gender and prediction of age. The method consists of two steps. The first step is a creation of regression model based on the labelled dataset and the second step is demography prediction via the Bayesian framework. The features used in classification consists of two groups: features based on the content (terms) and features based on the categories. The experimental results are followed: gender prediction - 79.7%, age prediction - 60.3% in the metric Macro F1. Kim [77] in the same way as Hu at all. in [65] used website content as the main source of information for demography prediction task. Predicted demographic attributes include gender, age, income and education. The prediction was realized via LDA and Logistic Regression algorithm.

The gender prediction using visited page history was the main issue of [74] too. The main idea of this work was the prediction of gender and age based on a web pages characteristics (e.g. words, HTML or hyperlinks). The work was oriented to the examination of different machine learning method performance. Suggested complex regression models achieved RMSE 9.97 in gender prediction task and RMSE 8.26 in age prediction task.

The work of Hu et al. [65] was an inspiration for Phuong et all. in [118], where the gender prediction based on browsing history was discussed. The features suggested by a Phuong et al. in comparison with Hu et al. was based on user behaviour (Hu et al. focused mainly on web page characteristics). The experimental results have shown the predictive power of high-level features such as categories and topics of visited web pages. Those high-level features in combination with time-based features and browsing patterns significantly improve the accuracy of prediction. Suggested model achieved 80.5% Macro F1.

Browsing history as the main source of information in demography prediction task was also used by Culotta et all in [35], where 6 prediction task were performed (prediction of gender, age, ethnic origin, education, income and offspring). The research was focused on users of social network Twitter. The main idea of work is in audience measurement that is used instead of the labelled dataset (e.g., it is estimated that 50% of the users, that visit some web page, has a bachelor's degree). The experimental results proved that data obtained via audience measurement are sufficient information for demography prediction task. Suggested model in the gender prediction task even overcomes the results of a classical supervised model.

Kakkar and Upadhyay in [75] suggested the method of users' age prediction based on users' web browsing history too. The browsing history in the form of features was used in the artificial neural network. ANN in binary classification into two groups adult and youngsters achieved 93.7% accuracy. Browsing history including social media was also analyzed by Goel et al. in [54], where the prediction of gender, age, income, education and race was suggested.

Users click stream in the context of demography prediction task was discussed by De Bock and Van Den Poel in [38]. Demography prediction task includes prediction of gender, age, level of education and category of profession. The performance of clickstream data using Random Forest classifier was compared with four algorithms: C4.5, CART, Bagging and AdaBoost. The experimental results showed that Random Forest classifier is the optimal algorithm for demography prediction based on

clickstream data.

The clickstream data was also analyzed by Ivanova in [71], where the demography prediction task using statistical methods was discussed. Demography prediction task contains following attributes: gender, age, education, marital status, income, employment status and also the residential area. Suggested features besides click stream features also include responses and interests of users. Features based on clickstream data were used by Atahan in [15] and Speltdoorn et all. in [140]. Atanah in the work focused on the predicted characteristics - age and income that was predicted via Bayesian classifier and Logistic Regression algorithm. The predicted characteristics of Speltdoorn et all. (gender, age, education and occupation) were predicted using ensemble semi-supervised learning algorithms: Tri-Training and Co-Forest.

3.1.3 Demography prediction in the domain of e-commerce

In the context of e-commerce, several papers dedicated to demography prediction task was presented. The papers in e-commerce mainly focused on feature engineering [1, 43, 98, 149], automatizing of features' acquisition [149] or big datasets [149].

Lu et al. in [98] focused on gender prediction based on viewing logs. The study was evaluated using the dataset from PAKDD 2015 competition. The dataset includes 30 000 records, where one record corresponds to one user session. The record contains followed information: session ID, begin time, end time and list of viewed items in the four abstraction levels. The suggested model consisted of 3 feature types: features based on user behaviour (session duration, number of products viewed during the session), time-based features (hours, parts of the day, days and weekends) and features based on viewing history (the most general category in session, first category in session, pairs of categories in session, individual categories in session). The problem of the unbalanced dataset was solved via cost-sensitive learning. The used classification algorithms were: SVM, FM and Random Forest algorithm. The best result 91.7% Macro F1 was achieved via SVM algorithm.

The gender prediction task using the dataset from PAKDD 2015 was also presented by Duong et al. in [43], where two types of features were suggested (basic features and advanced features). Basic features include temporal information (days in the month, months, days in the week, the start of the session in hours, end of the session in hours, session duration, average time of product viewing), number of viewed products and number of views per item or category. The advanced features include sequences in the form of n-grams, number of nodes in the abstraction level and the pair between two different abstraction levels. The problem of the unbalanced dataset was solved by a combination of sampling and cost-sensitive learning. The number of generated features (3 500 features) was reduced by information gain metric (2 500 features). The classification algorithms used for evaluation were: Random Forest, SVM and Bayesian networks. The best results 81.4% in Macro F1 was obtained by Random Forest algorithm.

Chaitanya et all. in [1] presented a novel approach to users' gender classification from the log files. The purposed method consists of two phases. The first phase is dedicated to feature engineering - it contains creation of features such as product category features, product prefix features or category

features. The second phase corresponds to classification based on features that were identified in the first phase. The method was evaluated using the dataset from PAKDD 2015 too. The experiment results showed 70% accuracy of the suggested method.

Cen in [32] suggested map-based gender prediction model that was dedicated to big e-commerce datasets. The prediction is based on the mapping the transition probability of the product categories with the presumption of transition process suitability for first-order Markov property. The suggested model was evaluated using the dataset from PAKDD 2015 Data Mining Competition. The experiment results showed the balanced accuracy 81.07%.

Wang et all in [149] looked at the demography prediction task as the multi-class multi-task classification problem. The classification was performed using transaction e-commerce dataset. The prediction includes 5 prediction tasks: prediction of gender, family status, income, age and the level of education. The work mostly focused on automatic feature acquisition that was performed through the proposed Structured Neural Embedding (SNE) algorithm. The main goal of SNE is automatic learning of transaction data representation and using of this representation in multi-task gender prediction.

3.2 Our user model for demography prediction task

The classification performance can be generally improved in two ways. The first way is based on the improvement of classification algorithms. The works focused on this type of improvement suggested new methods that outperform other classification algorithms, refined some existing algorithms or create new classification methodology based on a combination of some existing approaches (e.g. the combination of classification algorithms with clustering).

The second way is based on the feature engineering task. The feature engineering includes processes of feature creation, feature expansion and feature selection. In general feature engineering is based on domain knowledge of data. However, it is not a linear process. In many cases, suggested features do not carry the expected discriminatory power of information. The process of feature improvement is, therefore, an inherent part of feature engineering. The result features composed model that is used in the machine learning task (e.g. classification).

Our work deals with the second type of improvement that is based on the feature engineering. The result of our work is a user model, that was created considering the domain of e-commerce and e-commerce data availability. Suggested user model is general (because it should be usable for different predictive tasks - e.g. prediction of net income, prediction of child in family) and includes 7 562 features that can be divided into six types:

1. transaction-based features
2. temporal-based features
3. rating-based features
4. domain-based features

5. technical-based features
6. activity-based features

The types of features were selected with regard to the results of related works and with cooperation with domain experts. The following sub-chapters describe the individual types of features in detail.

3.2.1 Transaction-based features

Transaction-based features represent basic features that can be used by each e-commerce with transaction database. In general, this type of features is, therefore, the most widely-used type of features in our model.

- information about user purchases
- information about types of payments
- information linked to the purchased products
- distribution of purchases across different levels of category abstraction

Information about user purchases. Features created from transaction data contain amount-based features (number of purchased products, number of paid carts, number of unpaid carts, number of products in the cart, number of purchased gift products, number of canceled purchases), frequency-based features (number of purchases per year, number of purchases per month, frequency of purchases, time of the first purchase, time of the last purchase) and price-based features (paid amount, average purchase price, average price of products in the cart, average price of purchased gifts). All these features can be directly achieved from e-commerce transaction database. There is only one exception - features linked to the gift purchases, that does not have to be available in each e-commerce.

Information about types of payments. The group of features linked to the payment methods include followed features: number of payments through a bank account, number of payments through payment card, number of payments via Slovak 4 most popular bank companies, number of payments using mobile payment applications and the percentage proportion of individual types of payments. These features are also directly available in each e-commerce that provides different payment methods.

Information linked to the purchased products. Features linked to the products are e-commerce specific. They are based on the information obtained from the text description of the products or information available in the profiles of individual products. This group of features includes: family-based features (number of purchased baby-friendly products, number of purchased products that text is baby-friendly - text includes words such as baby or playground, the average number of people for whom the product is determined) and region-based features (the number of

local purchased products in the regions of Slovakia - Slovakia is divided into 8 regions, distribution of local purchases across regions in Slovakia, the most represented region in the local purchases). The region-based features are linked exclusively to local purchases such as food, local goods and local services. They do not include products such as travel trips or services for which people have to travel.

Distribution of purchases across different levels of category abstraction. These features include followed features in three abstraction levels: the number of purchases in a category, the average purchase price in a specific category and the distribution of purchases across categories. Our model contains three levels of abstraction:

1. Basic categories – e-commerce offer is structured into 6 basic categories - food, travel, services, goods, health and sports.
2. Domain expert categories – 145 categories that were linked to individual products by a domain expert.
3. Latent categories – categories obtained by machine learning algorithm.

The process of product abstraction to latent categories is perceived as topic modelling problem. The latent categories are created by Latent Dirichlet Allocation algorithm (LDA) - the most widely used topic modelling algorithm. The input of LDA is a corpus of items (one item = one product description in e-commerce), where each item corresponds to the one text document. This text document is represented as a vector of words or tokens. Before the LDA application, items should be pre-processed to this form. Pre-processing of items includes text cleaning, text tokenization, token lemmatization, removal of stopwords and diacritical marks and finally conversion to lowercase.

Pre-processed items in the form of the corpus can be used as input into the topic modelling algorithm. LDA is a three-level Bayesian model that represents an individual collection of items as a final set of topics. Each topic is represented as a set of basic token set probabilities [24]. Our latent categories correspond exactly to the first most probable topic in the set of topics. The method of latent categories acquisition is visualized in Figure 4.

3.2.2 Temporal-based features

Temporal-based features contain basic information also obtained in the transaction-based features. The difference between transaction-based features and temporal-based features is in the time period. While transaction-based features are not limited by time of execution, the temporal-based features are linked to a specific time period (e.g. day or month). As well as transaction-based features, temporal-based features can be also used by every e-commerce with transaction database.

Temporal-based features include followed information: the number of purchases in the time period, paid amount in the time period, the average price of purchases in the time period, distribution of purchases across basic categories (6 categories) in the time period and percentage of distribution across time periods. The time periods obtained in temporal-based features are followed: a month

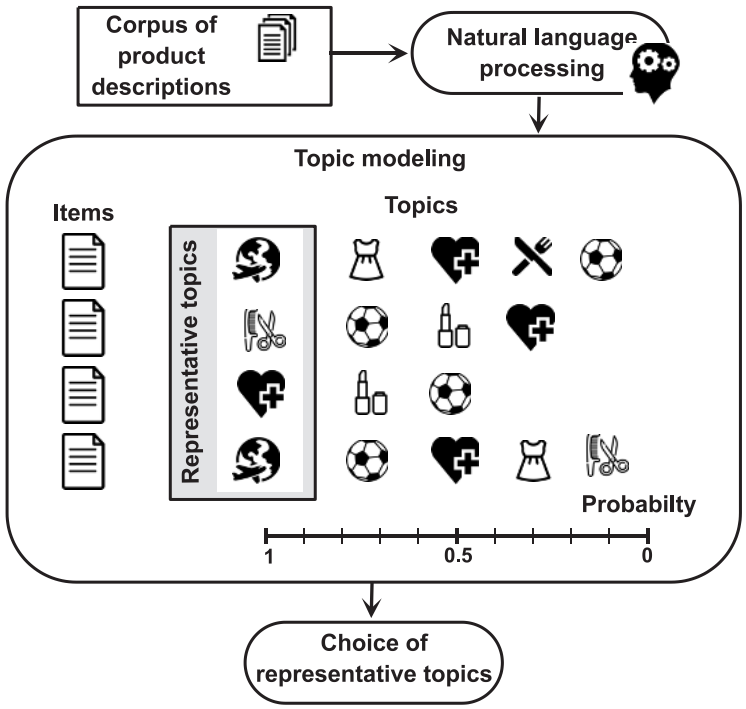

Figure 4: Method of latent categories acquisition. As we can see, the latent categories acquisition consists of 3 steps: natural language processing, topic modelling and choice of representative topics. The input of this method is a corpus of items, where each item is represented through the text document. The text document corresponds to product description in e-commerce. The output of natural language processing step is preprocessed corpus of items. Topic modelling consists of Latent Dirichlet Allocation algorithm that computes probabilities distribution across generated topics. The output of this step is set of items that are represented as lists of topics (that was created via LDA) with their probabilities. The topic corresponds to the latent category of item (e.g. food, sport, travel). The last step of this method is choice of representative topics - the representative topic is topic with the highest probability computed in topic modelling step. This topic corresponds to the latent category with which we work in our user model proposal.

in the year, day in the week, part of the day (morning, lunchtime, afternoon, evening, night), year season, Christmas time and e-commerce semesters (transaction database divided into semesters).

3.2.3 Rating-based features

Rating-based features draw from ratings provided for individual products. The type of rating is e-commerce specific. E-commerce often provides different ways of product rating. The rating can be binary (like - dislike or recommended - not recommended), discrete values (frequently three-valued scale, five-valued scale or ten-valued scale), continuous scale, text rating, detailed rating in the form of questionnaires or its combinations. User preference as some type of rating can be also expressed by mailing subscription, inserting into the favourites, inserting into the wish list, postponing for view later or recommending to friends.

Because of our dataset, our rating-based features include followed data: number of rated products, average rating, number of favourite products, number of subscribed mailings and number of unsubscribed mailings.

3.2.4 Domain-based features

This group of features consists of two types of features: features created due to the absence of demography information and features created due to the scoring purposes. All these features have been suggested by us, based on consultations with domain experts. The first group of features contains two features - gender estimation and estimation of the region where user lives. The second group includes three features - user e-commerce value, user activity and user preferences of luxury deals.

Gender estimation. Because of the gender absence, we suggested the process of its estimation. This process consists of three steps:

1. Identification of females based on the special suffix that is typical for female surnames.
2. Recognition of gender based on the first name via gender recognition tool[6].
3. Gender recognition based on the email address (it can also include user first name) via gender recognition tool.

Estimation of region where user living. In the absence of information about the user's residence, we propose a way of supplementing that information. The proposal is based on the assumption of the so-called local products. Local products are type of products, which generally depends on the location of the user, where he/she lives. In the context of e-commerce, there are offers such as food or a hairdresser, for which users usually do not travel.

Based on this idea, the following method for determining the user's residence was proposed:

```
local_products_GPS    # GPS of purchased local products

if count(local_products_GPS) >= 2:
    centroid = Centroid(local_products_GPS)
    median_of_distances = Median(DistacesToPoint(local_products_GPS, centroid)
    early_termination = False
    while median_of_distances > 20:    # 20km
        local_products_GPS = RemoveFurthermostPoint(local_products_GPS)
        if count(local_products_GPS) < 2:
            early_termination = True
            break
        centroid = Centroid(local_products_GPS)
        median_of_distances = Median(DistacesToPoint(local_products_GPS, centroid)
    if early_termination == False:
```

[6] https://genderize.io/

```
        median_latitude = Median(Latitudes(local_products_GPS))
        median_longitude =  Median(Longitudes(local_products_GPS))
        place_of_live = Point(median_latitude, median_longitude)
        user_region = PointInPolygon(place_of_live)
```

The value of 20 km was determined by the domain expert.

User value. User value corresponds to the actual value of a customer from the view of his/her buying behaviour. This value change in the time and it is influenced by followed parameters:

- overall paid amount
- number of purchases
- usage of payment card saving option
- time of the last purchase

The user value is expressed by following formula:

$$Value = CS * LP * 50 * \sum_{i=1}^{N} Price(Purchase_i) * TM(Purchase_i) \tag{7}$$

where N corresponds to the number of user purchases, $Price(Purchase_i)$ represents the price of purchase i, CS references to the use of payment card saving option and it can be expressed as followed:

$$CS = \begin{cases} 1.05 & \text{if user used the card saving option} \\ 1 & \text{else} \end{cases} \tag{8}$$

The influence of last purchase time is expressed by LP and it is equivalent to:

$$LP = \begin{cases} 1.1 & \text{if } Month(ActualTime - Time(LatPurchase)) \leq 2 \\ 1 & \text{else} \end{cases} \tag{9}$$

The time of individual purchases is reflected by $TM(Purchase_i)$. This time multiplier can be expressed by the followed formula:

$$TM(Purchase_i) = 1.1 - \lceil Month(ActualTime - Time(Purcahe_i)) \rceil \tag{10}$$

Individual constants were set according to the user behaviour analysis and consultation with domain experts.

User activity. Users' activity reflects the level of user activity in the e-commerce environment. It covers almost all the activities that the customer can do in the e-commerce environment. This score is calculated on different types of activities that customer performs and their contributions

Table 12: Types of user activities with their weights for User Activity calculation.

	Activity	**Weight**
1	Registration	4
2	User Log in	1
3	User log out	1
4	Search	1
5	View of the list of products	2
6	View of the product profile	3
7	Add product into the cart	1
8	Remove product from the cart	2
9	Reservation	4
10	Purchase	5
11	Purchase cancellation	3
12	Coupon download	3
13	Product rating	5
14	Change of profile	2
15	Mailing subscription	2
16	Mailing un-subscription in	2
17	Payment card saving	1
18	Payment card deletion	2
19	Add product to the favourites	1
20	Remove product from the favourites	1

to this score. User activity can be expressed by the following formula:

$$Activity = \sum_{i=1}^{M} Multiplicity(Activity_i) * Weight(Activity_i) \tag{11}$$

where M represents the number of different activities, $Multiplicity(Activity_i)$ expresses the number of $Activity_i$ occurrence in the context of user shopping behaviour and $Weight(Activity_i)$ corresponds to the weight of $Activity_i$. Table 12 shown individual activities, that can be performed in the e-commerce system and its weights.

The weights of activities were determined based on the user behaviour analysis and following consultation with the domain experts. The activity weight reflects the users' mental activity that must be developed to perform a given activity. Thus, the user activity represents a cumulative score that increases with activity in the system. These weights may vary between different e-shops - with respect to the page layout and organization of e-commerce application. However, our weights can be perceived as a basis in the case of this feature usage.

Preferences of luxury deals. Preferences of luxury deals reflect the level of products' luxury - products that were purchased by a user. It can be expressed by following formula:

$$Activity = \sum_{i=1}^{M} Multiplicity(ProductType_i) * Weight(ProducType_i) \tag{12}$$

where M corresponds to the number of product types, $Multiplicity(ProductType_i)$ indicates the number, how many times the $(ProductType_i)$ was purchased by a user and $Weight(ProducType_i)$ expresses the weight of the given product type. However, because the luxury of individual products is e-commerce specific, individual wights are not provided in this work.

3.2.5 Technical-based features

This group of features contains information available using user agent software. These features include a number of IP address, number of different devices, number of different operating systems, number of different browsers, number of cookies linked to the user, usage of top 10 most popular devices, usage of top 10 most popular operating systems, usage of top 10 top popular browsers. In addition to the information obtained via the user agent, this group also includes information about payment card saving (binary attribute True or False).

3.2.6 Activity-based features

Activity-based features contain information obtained from the users' activity recorded during the e-commerce website usage. These features are available only for e-commerce with web activity logging system. This group of features can be divided into following types of features:

- basic activity-based information
- temporal-based information
- information linked to the viewed products
- distribution of activity across different levels of category abstraction
- patterns in behaviour

Basic activity-based information. Basic activity-based features contain event-based features (number of recorded events, number of searches, number of product views, number of product list views - e.g. homepage, number of basket views), session-based features (number of recorded sessions, number of sessions with view of basket, number of session with purchase, average duration of session) and frequency-based feature (frequency of activity). These features are not e-commerce specific and therefore they can be recorded in each e-commerce.

Temporal-based information. Temporal-based features composed of following information linked to the certain time period: number of product views recorded in a time period, the percentual distribution of product views across time periods and number of product views distributed across 6 basic categories in the considered time period. Time periods considering in time-based features are followed: a month in the year, day in the weak, year season, Christmas time and part of the day (morning, lunchtime, afternoon, evening, night).

Information linked to the viewed products. Features linked to the viewed products include the same features as features linked to the purchased products in transaction-based features (see section 3.2.1). The difference is in their association with products. In transaction-based type, features are associated with the purchased product, and on the other hand, in activity-based type, features are associated with viewed products.

This group of features includes: family-based features (number of viewed baby-friendly products, number of viewed products that text is baby-friendly, the average number of people for whom the viewed product is determined) and region-based features (number of local viewed products in the regions of Slovakia, distribution of local viewed products across regions in Slovakia, the most represented region in the local viewed products).

Distribution of activity across different levels of category abstraction. These features include followed information in three abstraction levels: a number of product views in a category and the distribution of product views across categories. Our model in the same way as in transaction-based features contains three levels of abstraction :

1. Basic 6 categories.
2. Domain expert categories.
3. Latent categories.

Process of product abstraction to latent categories is described in section 3.2.1 (also see Figure 4).

Patterns in behaviour. The last part of activity-based features are patterns in user behaviour. User behaviour is recorded in the form of events. Patterns are therefore obtained via event abstraction method. Latent categories (see section 3.2.1 and Figure 4) together with low-level events are a basic input of event abstraction method.

The event abstraction composed of three steps: category-based event abstraction, pattern recognition and the choice of the best pattern combination with respect to the type of machine learning task (see Figure 5). Category-based event abstraction creates a high-level abstraction of individual events using latent categories. Pattern recognition is used for abstraction of event-set. It creates a new representation of events in the form of patterns. Abstraction of event-set reduces the number of events and choices the representative events. The output of pattern-based event set abstraction step consists of three sets - set of association rules, set of sequence patterns and set of n-grams.

Table 13: Solving the problem of dataset unbalance via sampling and cost-sensitive learning for different demography prediction tasks.

Demography characteristic	**Predicted classes**	**Sampling method**	**Class costs**
Net income	2 classes < 1 400, >= 1 400	undersampling	1:1.2
Partner relationship	2 classes have, do not have	oversampling	1:1
Child in family	2 classes yes, no	oversampling	1:1
User residence region	8 classes region of Slovakia	combination of oversampling and undersampling based on the 2. most numerous classes	1 for all

The recognized patterns are used in the last abstraction step - in the choice of the best patterns combination due to the machine learning task. The choice of the best combination is performed via information gain metric.

3.3 Evaluation of user model in demography prediction task

The performance of suggested user model was evaluated via demography prediction task. Specifically, we have focused on 4 demography characteristics: net income, partner relationship, a child in family and region of user residence. All predictions are perceived as user classification tasks. As the reference algorithm was chosen Random Forest algorithm - thanks to its robustness, efficiency and its results in related works. However, the result performance of our model (see section 3.3.5) in comparison with other models was also evaluated using SVM and KNN classifier.

The problem of dataset unbalance was solved through a combination of sampling and cost-sensitive learning. The choice of sampling method and class costs is shown in Table 13. Because the complexity of the suggested model, the feature selection task had to be performed. The feature selection was based on information gain metric. It was performed using Random Forest classifier with 250 trees (other parameters was set to default values). Relevant features were obtained as features with importance larger than 0.001 (computed based on information gain metric).

The dataset was split in 80:20 ratio (train:test). The train dataset was used for hyper-parameter tuning. Hyper-parameter tuning was performed via grid-search method using 10-fold cross validation. The cross validation evaluation metric was Recall Macro. Tuned parameters are presented in the Table 14. The result of parameter tuning for net income prediction is presented in Table 15, for partner relationship prediction in Table 16, for child in family prediction in Table 17 and for region of user residence prediction in Table 18.

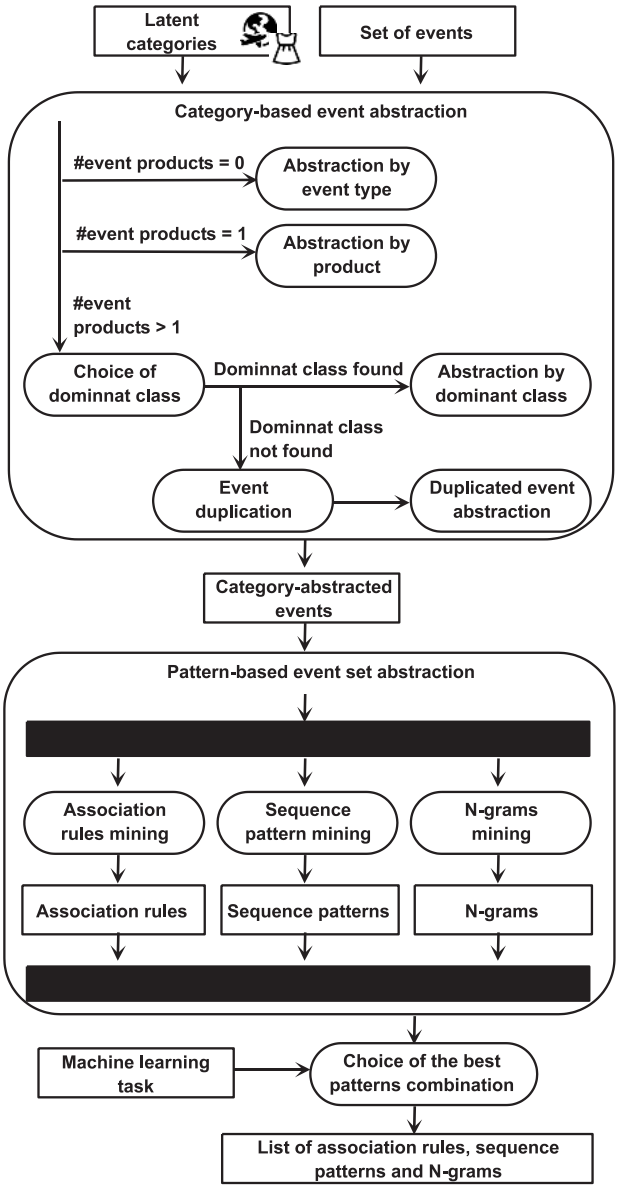

Figure 5: Method of event abstraction based on pattern recognition. The event-based abstraction is composed of three parts - category-based event abstraction, pattern-based event set abstraction and choice of the most representative patterns. The input to category-based event abstraction contains set of events and latent product categories. The category-based event abstraction consists of 4 types of event abstraction based on latent categories. The choice of event abstraction type is chosen based on products connected with the specific event. The output of category-based event abstraction is set of abstracted events. The next step, pattern-based event set abstraction, contains parallel pattern mining tasks - association rules mining, sequence pattern mining, N-grams mining. The last step is choice of the best patterns combination. The input of the last step contains patterns obtained in the pattern-based event set abstraction step and machine learning task (in our case classification) for which event abstraction method should be used. The output of this method is event set abstracted to the form of patterns.

Table 14: Parameters in hyper-parameter tuning.

Algorithm	**Parameter**	**Default value**	**Tried values**
Random Forest	maximal depth of tree	None	5, 10, 15, 20, 25, None
	number of trees in forest	10	10, 20, 50, 100, 250, 350, 500
	minimal samples to split a node	2	2, 3, 4, 5
	criterion for split quality evaluation	gini	gini, entropy
	maximal number of features considered in the best split search	sqrt	sqrt, log2, 0.1, 0.09, ..., 0.02
	using of bootstrap features	True	True, False
SVM	kernel	rbf	linear, poly, rbf, sigmoid
	penalty parameter C	1	0.5, 1, 1.5, 2
	degree of poly kernel	3	1, 2, 3, 4, 5
	gamma kernel koefficient	1 / number of features	1/B, B=25, 50, 100, 150, 200, 300
	coef0 for kernel function	0	0, 1, 2, 3, 4, 5
kNN	number of neighbors	5	odd-number between (2, 250)
	weights	uniform	uniform, distance
	p - parameter for Minkowski metric	2	1, 2, 3, 4
	metric	minkowski	chebyshev, minkowski

Table 15: The result of hyper-parameter tuning for net income prediction task.

Algorithm	**Parameter**	**Models**		
		Our	**Duong**	**Lu**
RF	max. dept	20	25	None
	num. of trees	250	250	250
	min. sample split	5	5	5
	criterion	entropy	gini	entropy
	max. features	0.09	0.02	log2
	bootstrap	True	True	True
SVM	kernel	poly	poly	poly
	C	1	1.5	1
	degree	2	1	1
	gamma	1/150	1/300	1/25
	coef0	2	2	1
kNN	num. neighbors	127	245	77
	weights	uniform	distance	distance
	p parameter	1	4	2
	metric	Mink	Mink	Mink

Table 16: The result of hyper-parameter tuning for partner relationship prediction task.

Algorithm	**Parameter**	**Models**		
		Our	**Duong**	**Lu**
RF	max. dept	20	15	15
	num. of trees	250	500	500
	min. sample split	3	3	3
	criterion	entropy	entropy	entropy
	max. features	0.09	sqrt	0.02
	bootstrap	True	True	False
SVM	kernel	poly	poly	poly
	C	0.5	1	1
	degree	1	2	1
	gamma	1/300	1/300	1/200
	coef0	0	1	1
kNN	num. neighbors	65	63	77
	weights	distance	distance	uniform
	p parameter	3	1	1
	metric	Mink	Mink	Mink

Table 17: The result of hyper-parameter tuning for child in family prediction task.

Algorithm	**Parameter**	**Models**		
		Our	**Duong**	**Lu**
RF	max. dept	10	None	None
	num. of trees	350	350	250
	min. sample split	5	2	5
	criterion	gini	gini	entropy
	max. features	0.07	0.1	sqrt
	bootstrap	False	False	False
SVM	kernel	poly	poly	poly
	C	1.5	1	1.5
	degree	5	3	3
	gamma	0.02	0.01	0.01
	coef0	5	2	3
kNN	num. neighbors	131	21	119
	weights	distance	distance	distance
	p parameter	1	4	1
	metric	Mink	Mink	Mink

Table 18: The result of hyper-parameter tuning for region of user residence prediction task.

Algorithm	**Parameter**	**Models**		
		Our	**Duong**	**Lu**
RF	max. dept	25	None	None
	num. of trees	500	500	250
	min. sample split	3	2	5
	criterion	gini	entropy	gini
	max. features	sqrt	sqrt	log2
	bootstrap	False	True	False
SVM	kernel	poly	poly	poly
	C	1.5	0.5	1
	degree	3	2	5
	gamma	1/300	0.01	1/200
	coef0	5	2	5
kNN	num. neighbors	63	85	47
	weights	distance	uniform	distance
	p parameter	2	2	2
	metric	Mink	Mink	Mink

The prediction was evaluated using metric Recall Macro (see Equation 8). In the comparison, we also report metrics: Precision Macro (see Equation 7) and F1 Macro (see Equation 9). Macro metric was calculated as the unweighted mean of metrics computed for individual classes. Macro metric, therefore, does not take into account problem of classes imbalance - this is particularly important when trying to predict infrequent classes.

$$Precision = \frac{|\{\text{relevant} \quad \text{documents}\} \cap \{\text{retrieved} \quad \text{documents}\}|}{|\{\text{retrieved} \quad \text{documents}\}|} \tag{13}$$

$$Recall = \frac{|\{\text{relevant} \quad \text{documents}\} \cap \{\text{retrieved} \quad \text{documents}\}|}{|\{\text{relevant} \quad \text{documents}\}|} \tag{14}$$

$$F1 \quad score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{15}$$

3.3.1 Dataset description

Our work is dedicated to the domain of e-commerce. Our dataset was obtained from private Slovak e-commerce company ZlavaDna[7]. ZlavaDna is a discount portal where the products in the form of discount coupons are offered. The coupon can be subsequently applied in real product providers. This dataset is specific with short-time deals and diverse offer. The average time of deal activity is 2.5 week. The time reserved for coupon usage is also limited. The offer is divided into 6 basic categories: food, goods, health, services, sports and travel. Individual short-time offers are associated to each other only through providers.

The dataset obtained from ZlavaDna for research purposes (but protected by NDA) includes the following data:

- Transaction data (3 million of purchases, 500 000 customers)
- User activity on the website (18.5 million of events)
- Item general information
- Item categories
- Customer general information
- Customer ratings

Dataset preprocessing. Because the user activity in the form of events was not preprocessed yet, we had to perform some preprocessing steps. The first step of preprocessing was cleaning the

[7] https://www.zlavadna.sk/

Table 19: The result of demography characteristics acquisition.

Demography characteristic	Original amount	Amount after acquisition
Partner relationship	3 428	12 224
Child in family	0	8 869
User residence region	58 825	91 668
Net income	9 236	11 325

dataset. In this phase the error and irrelevant events were deleted. As irrelevant events were events that:

- contained time-stamp outside the expected range
- did not contain the information about activity type (such as the view of offer, purchase or rating)
- contained a test label
- did not include a user identifier
- did not contain association to deal (if that type of event should)
- were linked to deal with less than 5 view events.

The result of activity cleaning was the reduction of the original event set by 17%.

The second step of activity preprocessing was cookies - user ID mapping. Because in ZlavaDna sign-up is not necessary, we had to solve the problem of logged-in and not logged-in users. The activity events had to be backwards mapped to the real logged-in user identifiers.

The last step of dataset preprocessing was session identification. Because the recorded events did not contain explicit session identifiers, they had to be identified. The session identifier timeout was set on 25 minutes. The dataset included 3.3 million of sessions with the average length 4.93 events.

Acquisition of demography information. Because the demographic characteristics are type of information that customers do not like to provide and many e-commerce (includes e-commerce, which provided our dataset) do not require them, we had to proposed the process of demography acquisition. In our case the demography acquisition consists of two steps:

1. demography acquisition via a questionnaire on the e-commerce website
2. demography acquisition through social network scraping

The result of demography information acquisition is presented in Table 19. The parameters of the labelled dataset for demography prediction task are shown in Table 20.

Table 20: The labelled dataset for demography prediction task (we considered only persons with more than three purchases and activity during last year).

Demography characteristic	**Number of samples**	**Predicted classes**	**Class ratio**
Net income	9 236	2 classes < 1 400, >= 1 400	15:4
Partner relationship	9 286	2 classes have, do not have	10:9
Child in family	6 771	2 classes yes, no	8:5
User residence region	58 825	8 classes region of Slovakia	34:11:10:14 :14:12:12:12

Table 21: The comparison of transaction-based features (TRAN) and a combination of transaction-based features with temporal-based features (TRAN + TEMP) in the demography prediction tasks using Random Forest algorithm.

Predicted Characteristic	**Types of features**	**Metrics**		
		P	**R**	**F1**
net income	TRAN	53.2	55.3	49.5
	TRAN + TEMP	57.4	63	54.2
partner relationship	TRAN	55.5	55.6	55.6
	TRAN + TEMP	55.6	55.5	55.3
child in family	TRAN	58.4	57.6	57.5
	TRAN + TEMP	59.8	57.8	57.1
region of user residence	TRAN	35.7	35	35
	TRAN + TEMP	36.7	36	36

3.3.2 RQ1: How the temporal-based features improve the classification performance when they are joined to the transaction-based features?

Temporal-based features are one of the widely-used types of features using in user modelling task. They are often used for user classification or personalized recommendation. As can be seen in related works, they are also popular in demography prediction tasks. Their performance in the demography prediction task is discussed in our first research question (see RQ1).

The evaluation was realized as user classification task using Random Forest algorithm. In the evaluation, we compared the performance of transaction-based features (our baseline) and a combination of transaction-based features with temporal-based features. Thus, we examined the contribution of temporal-based features in the demographic predictions (see Table 21).

As can be seen in Table 21 the temporal-based features improved the net income prediction task by 7.7% and region of user residence prediction task by 1% in metric Recall Macro. In other prediction tasks (prediction of partner relationship and prediction of a child in the family), transaction-based

features and combination of temporal-based features with transaction-based features achieved comparable results. In general, we can say that the temporal-based characteristics have the potential to improve the results of the demographic forecast, but this improvement is bound to the predicted characteristic.

Despite the fact, that majority of top 20 relevant features in net income prediction task and region of user residence prediction task consisted of transaction-based features, temporal-based features were able to improve the performance of this predictions. Ones of the most important temporal-based features in net income prediction task included paid amount during the winter time and paid amount outside the Christmas time. Paid amount during the winter time was ranked as the 9th most important feature and paid amount outside the Christmas time as the 10th most important feature.

In the context of region of user residence prediction task, the top 20 most important features includes followed temporal-based features: average price paid outside of Christmas time (10th position), paid amount outside the Christmas time (12th position), average price paid during the winter time (19th position) and paid amount during the winter time (20th position).

3.3.3 RQ2: What is the improvement of classification based on activity features in the comparison with classification based on transaction features?

In recent decades in the context of e-commerce, users' activity monitoring is rapidly growing. The recorded users' activity is often used for system personalization and different types of predictions (e.g. prediction of e-commerce yield, customer purchasing behaviour prediction or customer characteristics prediction). In our second research question we, therefore, discussed a comparison of prediction based specifically on transaction data and predictions based specifically on activity-based data.

The performance of suggested features was evaluated via user classification task using Random Forest classifier. In the first step of evaluation we have compared the performance of transaction-based features with the performance of activity-based features - see Table 22. And in the second step we have evaluated the contribution of activity-based features to transaction-based features in demography prediction task - also see Table 22.

Our expectation for the first comparison was the equivalent predictive power of transaction-based features and activity-based features. However, as Table 22 shows, activity-based features achieved weaker results than transaction-based features. We assume that this is mainly due to our dataset - our dataset contains 4 years of transaction-based data and 1.5 years of activity-based data. Based on this, we assume that 1.5 years of activity-based data is too short to use these data as a main source of information for demography prediction tasks (also with respect to user activity in the system where the average user buys 1.2 times per year).

Nevertheless, the Table 22 also shows that activity-based features as a complement of transaction-based features have the potential to improve demography prediction task. This improvement was observed in net income prediction task and region of user residence prediction task. In net

Table 22: The comparison of transaction-based features (TRAN), activity-based features (ACT) and a combination of transaction-based features with activity-based features (TRAN + ACT) in the demography prediction tasks using Random Forest algorithm.

Predicted Characteristic	Types of features	Metrics		
		P	R	F1
net income	TRAN	53.2	55.3	49.5
	ACT	52.6	54.5	43.3
	TRAN + ACT	57.6	62.5	54.2
partner relationship	TRAN	55.5	55.6	55.6
	ACT	50	50	48.2
	TRAN + ACT	55.7	55.7	55.6
child in family	TRAN	58.4	57.6	57.5
	ACT	54.2	52.9	51.2
	TRAN + ACT	57.8	57.4	57.4
region of user residence	TRAN	35.7	35	35
	ACT	33.2	27.3	27.4
	TRAN + ACT	39	39	38.9

income prediction, the combination of activity-based features and transaction-based features (in comparison with only transaction-based features) increased metric Recall Macro by 7.2%. The improvement in the region of user residence prediction was 4%.

The top 20 most important features in net income prediction task included followed activity-based features: frequency of activity (11th position) and the number of views (18th position). In the region of user residence prediction, the top 20 most important features also included: frequency of activity (8th position) and the number of sessions (20th position). The majority of the most relevant features was formed by transaction-based data.

3.3.4 RQ3: What is the improvement of classification performance, when domain-based features are added to the transaction-based features?

Domain-based features consist of 5 features: gender estimation, user residence estimation, user value, user activity and user luxury preferences (for more information see section 3.2.4). Their contribution to demography prediction tasks was compared to transaction-based features that are our baseline feature type (see Table 23).

As can be seen in Table 23 the domain-based features improved the demography prediction in followed tasks: prediction of net income - improvement by 7.3%, prediction of partner relationship - improvement by 4.1% and prediction of a child in the family- improvement by 1.8% in metric Recall Macro.

The net income prediction task was mostly influenced by domain-based features: user activity (8th position) and gender estimation (14th position). As the relevant features in feature selection were also selected: user luxury preferences and estimation of user residence. Partner relationship prediction was influenced by gender estimation (4th position), user activity (9th position) and also

Table 23: The comparison of transaction-based features (TRAN) and a combination of transaction-based features with domain-based features (TRAN + DOM) in the demography prediction tasks using Random Forest algorithm.

Predicted Characteristic	Types of features	Metrics		
		P	R	F1
net income	TRAN	53.2	55.3	49.5
	TRAN + DOM	56.7	62.6	53.4
partner relationship	TRAN	55.5	55.6	55.6
	TRAN + DOM	59.7	59.7	59.7
child in family	TRAN	58.4	57.6	57.5
	TRAN + DOM	60	59.4	59.5
region of user residence	TRAN	35.7	35	35
	TRAN + DOM	35.6	35.3	35.3

user luxury preferences and user residence estimation. The same domain features influenced the prediction of a child in the family too.

3.3.5 RQ4: How suggested user model perform in the comparison with other state-of-the-art user models proposed for the domain of e-commerce?

The final evaluation of our model performance was based on comparison with models suggested in PAKDD 2015 competition. Specifically, we compared with the model suggested by Lu et al. in [98] and model proposed by Duong et al. in [43]. The comparison was realized as demography prediction task where 4 characteristics were predicted: net income, partner relationship, a child in family and region of user residence.

Since each model has a fairly large number of features, the feature selection task had to be done. The result distribution of features in our model after feature selection is presented in Table 24. The ordered list of top 10 selected features for individual models are shown in followed tables: for net income prediction in Table 25, for partner relationship prediction in Table 26, for child in family prediction in Table 27 and for region of user residence prediction in Table 28.

As can be seen in Table 24 net income prediction was mostly realized on the basis of temporal features. However, the first positions were occupied by gender estimation and transaction-based features (see Table 25). Activity-based features were placed in the second half of the selected features list. From domain-based features, gender estimation, user e-commerce value and user activity were selected as relevant. From rating-based features was relevant the number of rated items and from technical-based features were relevant followed features: number of cookies, number of operating systems and number of used browsers.

The user luxury preferences were not chosen as a relevant feature - we can, therefore, assume that the user luxury preferences are not directly related to net income. List of relevant features also includes 4 types of payments, information linked to 11 latent categories and information linked to 6 expert categories.

Prediction of partner relationship in the same way as a prediction of net income is mainly built on temporal-based features (see Table 24). The first most relevant feature was gender estimation with importance 0.138 and the second most relevant was average price paid outside the Christmas time with importance 0.005 (see Table 26). The partner relationship prediction is mostly influenced by gender prediction, then transaction-based features, temporal-based features and finally activity-based features. From domain-based features, the partner relationship prediction is also influenced by user e-commerce value and user activity.

Partner relationship is also influenced by an average number of persons for which the offer is determined and the number of baby-friendly purchases. From features related to categories, 5 latent categories and 2 expert categories were selected as relevant.

The child in family prediction was mostly influenced by gender estimation - see Table 27 (the first place with importance 0.152). The prediction was built on transaction-based features, temporal-based features and activity-based features (see Table 24). The child in family prediction just like partner relationship prediction was also influenced by the average number of persons for which the offer is determined, the number of baby-friendly purchases and features linked to 5 latent categories and 1 expert category.

The region of user residence prediction was mostly influenced by features linked to purchases of local items in regions and views of local items in regions (see Table 28). From features that were not linked to regions, the average price paid in the category food achieved the highest importance. One of the most important features was also the estimation of user location (from domain-based features). The gender estimation was not included in the relevant features list. Relevant features also include features linked to 8 latent categories and 6 expert categories.

Table 24: The result distribution of individual types of features in the list of relevant features after feature selection task (TRAN: transaction-based features, TEMP: temporal-based features, RAT: rating-based features, DOM: domain-based features, TECH: technical-based features and ACT: activity-based features).

Predicted charac.	**Types of features**					
	TRAN 716	**TEMP 540**	**RAT 5**	**DOM 5**	**TECH 36**	**ACT 720 + patterns**
net income	61	158	1	3	4	84
partner relation.	54	153	1	3	4	56
child in family	55	145	0	3	3	39
region of user residence	68	154	0	4	4	54

The final comparison in net income prediction task is presented in Table 29. We can see that our model outperform other state-of-the-art models. Our model reached 60.6% in metric Recall Macro

Table 25: Top 10 most relevant features for net income prediction task.

	Models		
	Our	**Duong**	**Lu**
1.	gender estimation	average number of viewed products in session	average time of session
2.	paid amount outside the Christmas time	average time of session	number of views in category: travel
3.	overal paid amount	number of nodes in expert categories level	average number of viewed products in session
4.	paid amount during winter	number of views in category: travel	number of views in category: metropolis
5.	purchases frequency	average time per product in session	number of views in category: wellness
6.	paid amount in morning purchases	number of nodes in basic categories level	number of views in category: health and beauty
7.	average price of purchases	number of views in category: sport	number of views in category: goods
8.	time of last purchase	number of views in category: health and beauty	number of views in category: aquaparks
9.	time of first purchase	number of views in category: goods	number of views in category: food
10.	average price of items in cart	number of views in category: hiking	number of views in category: car

which is 6.7% more the Duongs' model and 7.1% more than Lus' model. The best performing algorithm for our model was SVM classifier. On the other hand, the best performing algorithm for both PAKDD models was Random Forest classifier.

The comparison via partner relationship prediction task is shown in Table 30. The best result in this task was also achieved by our model. The final score of our model was 64.6% in Recall Macro - Duongs' model lagged by 11.2% and Lus' model by 12.1%. The best performing algorithm for each model was SVM classifier.

As we can see in Table 31, the best performing model in a child in family prediction task was our model that obtained score 64.5% Recall Macro. The Duongs' model obtained score 53.9% that is 10.6% less than our model. Lu model score was 53.7%. The best algorithm for both our model and Lus' model was SVM classifier and for Duongs' model kNN classifier.

The comparison in the region of user residence prediction task shown that the best model for this task is also our model (see Table 32). The Recall Macro of our model was 28.4% which is 8.4% more than Duongs' model and 7.5% more than Lus' model. The best performing algorithm in this

Table 26: Top 10 most relevant features for partner relationship prediction task.

	Models		
	Our	**Duong**	**Lu**
1.	gender estimation	average time of session	average time of session
2.	average price paid outside the Christmas time	number of views in category: travel	number of views in category: travel
3.	purchases frequency	average number of viewed products in session	number of views in category: goods
4.	time of last purchase	number of nodes in expert categories level	number of views in category: sport
5.	time of first purchase	number of nodes in basic categories level	session started at 14:00
6.	number of purchases per year	number of views in category: goods	number of views in category: health and beauty
7.	paid amount outside the Christmas time	number of views in category: sport	average number of viewed products in session
8.	average price paid during the winter	number of views in category: health and beauty	number of views in category: skiing
9.	number of card payment	number of views in category: food	session started in morning
10.	average price of items in cart	session started at 14:00	number of views in category: wellness

task was Random Forest algorithm.

As can be seen in Table 29, Table 30, Table 31 and Table 32 our model outperform Duongs' and Lus' model in all demography prediction tasks. Based on these results, we can conclude that in the context of e-commerce with less user shopping frequency (e-commerce with a large number of new customers but fewer returning customers), models created on activity-based data are not as effective as models created on a combination of transaction-based data and activity-based data. In the comparison of Duongs' model and Lus' model with our activity-based features (see Table 33) our activity-based features outperform both models in the region of user residence prediction task - the result score was 7.3% more than Duongs' model and 6.4% more than Lus' model. The results in other demography prediction tasks were comparable.

Table 27: Top 10 most relevant features for child in family prediction task.

	Models		
	Our	**Duong**	**Lu**
1.	gender estimation	average time of session	average time of session
2.	purchases frequency	number of nodes in expert categories level	number of views in category: travel
3.	time of last purchase	number of views in category: travel	session started at 14:00
4.	number of purchases per month	number of viewed products in session	number of views in category: health and beauty
5.	overal paid amount	number of views in category: goods	number of views in category: sport
6.	average price of items in cart	number of nodes in basic categories level	number of views in category: goods
7.	frequency of activity	number of views in category: health and beauty	session started in morning
8.	average price paid outside the Christmas time	number of views in category: sport	average number of viewed products in session
9.	percentage of card payments	number of views in category: wellness	number of views in category: wellness
10.	paid amount during the summer	activity in April	session started in afternoon

Table 28: Top 10 most relevant features for region of user residence prediction task.

	Models		
	Our	**Duong**	**Lu**
1.	percentage of purchases in BA region	average time of session	average time of session
2.	number of purchases in BA region	number of nodes in expert categories level	number of views in category: goods
3.	the most represented region	number of views in category: goods	number of views in category: travel
4.	percentage of purchases in KE region	average number of viewed products in session	average number of viewed products in session
5.	number of purchases in KE region	number of views in category: travel	number of views in category: journals and books
6.	number of viewed products in BA region	number of nodes in basic categories level	number of views in category: sport
7.	percentage of purchases in ZA region	average time for product views in session	session started in afternoon
8.	percentage of purchases in BB region	number of views in category: sport	session started in morning
9.	number of viewed products in KE region	number of views in category: health and beauty	activity in Tuesday
10.	number of purchases in BB region	number of views in category: journals and books	number of views in category: health and beauty

3.4 Summary of demography prediction in e-commerce

Demographic characteristics are the basic attributes that influence a shopping behaviour. However, customers do not like to provide them, and this is, therefore, a place for demographic prediction. In this chapter of our work, we present a user model based on e-commerce data. This user model was suggested for demography prediction tasks and consists of 6 general types of features – transaction-based, temporal-based, rating-based, domain-based, technical-based and activity-based features.

Our model was evaluated using 4 prediction tasks: prediction of net income, prediction of partner relationship, prediction of a child in family and prediction of user residence region. Based on the comparison of transaction-based features (that are our baseline features) with other feature types, we can see the importance of temporal-based features, activity-based features and domain-based features. The temporal-based features in RQ1 improved the net income prediction task by 7.7% in Recall Macro. In RQ2, thanks to the combination of activity-based features and transaction-based features, the net income prediction was also improved by 7.2%. The improvement by 4% in Recall Macro was also recorded in the region of user residence prediction task. As RQ3 showed, domain-

Table 29: The comparison of suggested model with models proposed by Duong and Lu via net income prediction task.

algorithm	Models								
	Our			**Duong**			**Lu**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
RF	55.7	59.1	53	52.4	53.9	42.1	52.1	53.5	43.2
SVM	56.1	60.6	52.3	52.3	53.6	41.8	51.9	53.2	41.6
KNN	55.4	59.4	51.5	50.4	50.7	39.8	49.6	49.5	49.2

Table 30: The comparison of suggested model with models proposed by Duong and Lu via partner relationship prediction task.

algorithm	Models								
	Our			**Duong**			**Lu**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
RF	63.6	62.9	62.5	53.8	52.8	49.4	51.9	51.4	48.4
SVM	65.1	64.6	64.4	54.6	53.4	49.9	53.2	52.5	49.8
KNN	64.4	64	63.9	48.9	49.2	44.9	49.5	49.7	44.7

based features mostly influenced net income prediction - improvement by 7.3%, partner relationship prediction - improvement by 4.1% and prediction of a child in the family - improvement by 1.8% in metric Recall Macro.

The comparison with 2 state-of-the-art models, which have been placed in the top 10 in the PAKDD 2015 competition, showed the robustness of our model. In general, our model outperforms these models in each prediction task. In net income prediction task, our model reached 60.6% in Recall Macro which was 6.7% more than Duongs' model and 7.1% more than Lus' model. The final score of our model in partner relationship prediction was 64.6% in Recall Macro - Duongs' model lagged by 11.2% and Lus' model by 12.1%. In a child in family prediction, our model obtained score 64.5%, Duongs' model 53.9% (10.6% less than our model) and Lus' model 53.7%. The Recall Macro of our model in the region of user residence prediction was 28.4% which was 8.4% more than the Duongs' model and 7.5% more than Lus' model.

Based on these results, we can conclude that in the context of e-commerce with less user shopping frequency (e-commerce with with a large number of new customers but fewer returning customers), models created on activity-based data (such as models suggested by Lu and Duong in PAKDD 2015)

Table 31: The comparison of suggested model with models proposed by Duong and Lu via child in family prediction task.

algorithm	Models								
	Our			**Duong**			**Lu**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
RF	63.2	63.5	61.8	51.9	51.5	50.4	54.4	53	50.6
SVM	64.1	64.5	62.6	53.4	52.7	51.7	55.6	53.7	51.3
KNN	63.9	64.3	62.6	55.2	53.9	52.7	50	50	42.8

Table 32: The comparison of suggested model with models proposed by Duong and Lu via region of user residence prediction task.

algorithm	**Models**								
	Our			**Duong**			**Lu**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
RF	53.2	28.4	28.7	22.4	20	18.5	24	20.9	19.7
SVM	49.2	27.8	27.6	20	18	16.3	22.5	19.7	18.3
KNN	47.5	26.5	25.8	21.6	19.2	17.7	22.1	19.5	18

Table 33: The comparison of activity-based features that are part of our suggested model with models proposed by Duong and Lu using Random Forest classifier.

demography characterics	**Models**								
	Activity-based			**Duong**			**Lu**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
net income	52.6	54.5	43.3	52.4	53.9	42.1	52.1	53.5	43.2
partner relationship	50	50	48.2	53.8	52.8	49.4	51.9	51.4	48.4
child in family	54.2	52.9	51.2	51.9	51.5	50.4	54.4	53	50.6
region of user residence	33.2	27.3	27.4	22.4	20	18.4	24	20.9	19.7

are not as effective as models created on a combination of transaction-based data and activity-based data. In the comparison of Duongs' model and Lus' model with the activity-based features (one part of our user model), our activity-based features outperform both models in the region of user residence prediction task - the result score was 7.3% more than Duongs' model and 6.4% more than Lus' model. The results in other demography prediction tasks were comparable.

4 User Model for Recommendation in E-commerce

Information overload, in the context of the ever-increasing amount of data produced on the Web, has become acute in recent decades. One of the solutions to this problem is the recommendation - more precisely personalized recommendation. There are several types of recommendation including well-known content-based recommendation, collaborative filtering and hybrid methods. Despite the fact that fairly large number of recommendation methods were developed, most of them are based on the user model. User model, in general, represents the preferences of the individual users. Represented preferences are domain specific - they correspond to the purchased or well-rated products in e-commerce, learning style in e-learning or read books in e-library [78].

There are several approaches to user model creation. Brusilovsky in [27] presented two basic user model types - Stereotype model and Overlay user model. The main goal of Stereotype model is user mapping into the predefined classes. Overlay user model represents the user preferences by adding the user-preferences layer into the domain model. Those two types of user models became the inspiration for the next works and nowadays form the basis of many research works in the field of user modelling.

Senot et al. in [134] represent the user model as a set of <concept, value> pairs. The value in this pairs represent the level of interest in specific semantic concept (e.g. category of domain item) and it is taken from the [0,1] interval. The importance of semantic concept can be measured by following types of information: Quantity of Affiliation, Quantity of Consumption and Quantity of Interest. The quantity of Affiliation corresponds to a level of content affiliation to a specific semantic concept (e.g. in the domain of e-commerce product "pizza" can be described by their categories {Food = 0.9, Restaurants = 0.8}). The quantity of Consumption represents the intensity of concept consuming (e.g. the higher is a time of product viewing or purchased products' amount the higher is the interest to categories Food and Restaurants too) and Quantity of Interest represents the level of user interest in a semantic concept.

Besides pairs, the most common model representations include vector models [160, 168], graph models [68, 157] and Bayesian network models [113, 163]. Yu et al. in [168] in focused on group recommendation and profile merging, where the user model in the form of vectors was suggested. Individual vectors of user model correspond to movie-specific characteristics such as actors, genres or keywords. The web usage mining in the form of active sessions and access patterns formed to the n-dimensional vectors in [160], where user modelling based on Latent Dirichlet Allocation model was discussed.

As can be seen in related works, the user model is usually used to reflect the relationship between the user and domain-specific item. In general, it expresses the level of interest and it can be represented as follows [78]:

$$Model_u = \bigcup (Item_u, Value_u) \tag{16}$$

where the $Item_u$ corresponds to the specific item in the domain and $Value_u$ corresponds to the

intensity of preferences.

Nevertheless, the user model is generally used to represent preferences, it may also contain other data that describe the user. Wu and Chen in [156] discussed the user personality as input into the personalized recommendation. Their work was dedicated to the implicit acquisition of personality traits defined in big-five inventory. The results showed the significant improvement of recommendation in ranking accuracy and rating prediction. The personality traits as part of active learning model were proposed in [44], where the cold start problem in recommendation was discussed. The cold start problem was also the key issue of [66], where the personality-based recommendation using the combination of ratings and personality traits was suggested [66].

The concept of mood-based recommendation was discussed in [150], where user model in the form of multiple vectors corresponded to individual user moods was proposed. The emotions in music recommendation were the key idea of [87], where the graph-based user model consists of emotions and music features was proposed. The demographic characteristics as part of user model are the bases for demography-based recommendation [37, 115]. Besides characteristics such as personality, moods, emotions or demography, the web usage mining in the form of behavioural patterns - association rules [128] or sequence patterns [8], can be part of a user model.

In our work, we focus on complex user models, that besides the user preferences also include other user characteristics. We suggested the user model that consists of follows types of characteristics:

- preference-based characteristics
- user traits
- pattern-based characteristics

Because of our user model complexity, this is a space for our first research question:

RQ1: How influence the individual parts of user model (preference-based characteristics, pattern-based characteristics and user traits) the result of the personalized recommendation in the domain of e-commerce with short-time deals?

Since the impact of personality traits on decision making is still actual research theme [106, 111] (also in the domain of e-commerce [25, 110]), the second research question is dedicated to this type of user characteristics.

RQ2: What is the impact of personality-based characteristics in combination with other characteristics types on personalized recommendation?

The contribution of this chapter of our work is following:

- we suggested the new user model for the personalized recommendation in the domain of e-commerce
- we analyze the effect of individual user model characteristics on personalized recommendation
- we closely analyze the impact of personality characteristics in combination with other characteristic types on recommendation task

4.1 Our proposal for user model components

In the context of e-commerce, the user model is typically formed by user preferences expressed by explicit or implicit feedback. The explicit feedback includes product ratings or labeling of the product as a favorite. The implicit feedback is derived from user activity and it includes the product purchases, product views or product retrieval. In this way, the user model is usually formed as a pairs <product, value> or set of vectors. Due to the e-commerce domain data availability, there is also a space for usage other types of information, such as demographic characteristics or item category information.

In order to reflect the complexity of user decision, where user choose the products not only based on his/her preferences, but also based on his/her traits and behavioural patterns, we suggested the user model UM that can be expressed as a triplet (our user model formal definition was inspired by [76]):

$$UM = (Pref, Trait, Pattern) \tag{17}$$

where $Pref$ corresponds to user preferences, $Trait$ to user traits and $Pattern$ to user behavioural patterns.

User preferences $Pref$ are defined as set of different types of user preferences and can be described as:

$$Pref = \{Categ, Temp, Reg, Tech\} \tag{18}$$

where the $Categ$ corresponds to the user interest in specific category expressed via explicit or implicit feedback, $Temp$ corresponds to user temporal preferences. Reg corresponds to user regional preferences and $Tech$ to user technical parameter preferences.

The user traits $Trait$ contains three types of traits: personality traits $Person$, demography traits $Demog$ and domain specific characteristics $Domain$. They can be expressed as:

$$Trait = \{Demog, Person, Domain\} \tag{19}$$

The last part of suggested user model - user behavioural patterns $Pattern$ include three types of patterns: association rules, sequence patterns and N-grams and is defined as:

$$Pattern = \{AR, SP, Ngram\} \tag{20}$$

In general, $\forall$ element $e \in \bigcup\{Pref, Trait, Pattern\}$ element e represents the set of weighted vectors vw as follows:

$$e = \{vw_1, vw_2, vw_3, ...vw_{n;n\in\mathbb{N}}\}; \quad vw_i \in VW \tag{21}$$

where VW corresponds to set of all vectors.

The individual vector vw_i is defined as a pair of value vector v and the vectors' weight w:

$$vw_i = (w_i, v_i); \quad i \in \mathbb{N} \tag{22}$$

where vector v corresponds to the set of values va:

$$v = \{va_1, va_2, va_3, ...va_{m;m=|v|}\} \tag{23}$$

The weight w reflects the importance of individual user characteristic. In our work, we used static vectors' weights set to ones (we rather focused on model parts). The values va in value vector v is defined as triplets:

$$va_i = (id_i, val_i, we_i); \quad i \leq |v|; \quad id_i \in Id; val_i \in Val; we_i \in We \tag{24}$$

where we_i corresponds to the weight of value va_i considering to identifier id_i. The We is a set of weights which model works with (in our case, we work with constant weights), Id is a set of characteristic identifiers and Val represents the set of available values. The Id and Val are characteristic-specific.

In the case of $Categ$ in $Pref$ we considered following types of feedback measures:

- user purchases - $Pur = \{PurB, PurNC, PurPD, PurP, PurAP\}$
  1. binary attribute - user had or did not have a purchase in specific category - $PurB$
  2. number of purchases in specific category normalized to <0,1> interval - $PurNC$
  3. percentage distribution of purchases through categories - $PurPD$
  4. the overall paid amount in specific category normalized to <0,1> interval - $PurP$
  5. average price paid per purchase in the given category normalized to <0,1> interval - $PurAP$
- user ratings - $Rat = \{RatB, RatNC, RatPD, RatAR\}$
  1. binary attribute - user rated or did not rated specific category - $RatB$
  2. number of rating in specific category normalized to <0,1> interval - $RatNC$
  3. percentage distribution of ratings through categories - $RatPD$
  4. the average rating value in specific category normalized to <0,1> interval - $RatAR$
- mailing subscriptions - $Sub = \{SubB, SubPD\}$
  1. binary attribute - user subscribed or did not subscribed specific category - $SubB$
  2. percentage distribution of subscriptions through categories - $SubPD$
- favourite products - $Fav = \{FavB, FavPD\}$
  1. binary attribute - user added or did not added specific category to favourites - $FavB$
  2. percentage distribution of favourite items through categories - $FavPD$
- viewed products - $View = \{ViewNC, ViewPD\}$

1. number of views in specific category normalized to <0,1> interval - $ViewNC$
2. percentage distribution of views through categories - $ViewPD$

The set of identifiers Id for $Categ$ characteristics is therefore expressed as:

$$Id = \bigcup\{Pur, Rat, Sub, Fav, View\} \times \bigcup\{BC, LC, EC\} \tag{25}$$

where BC corresponds to basic categories (in our case 6 basic categories), LC corresponds to latent categories and EX to expert categories. The Val is associated to Id based on interest measure type.

In the case of other types of preference characteristics - $Temp$, Reg and $Tech$, the Val and Id are as followed:

- For the temporal preferences $Temp$, the $Val =< 0, 1 >$ and $val_i \in Val$ corresponds to the number of purchases in specific time period normalized to <0,1> interval. The Id represents the specific time period:
  - day in week: $Id = \bigcup_{i=1}^{7} Dw_i$
  - day in month: $Id = \bigcup_{i=1}^{31} Dm_i$
  - hour in day: $Id = \bigcup_{i=0}^{23} H_i$
  - month in year: $Id = \bigcup_{i=1}^{12} M_i$
  - year seasons: $Id = \{spring, summer, fall, winter\}$
  - day parts: $Id = \{morning, afternoon, evening, night\}$
- For regional preferences Reg, set of identifiers $Id = \bigcup_{i=1}^{8} R_i$ represents the individual regions of Slovakia and set of values $Val =< 0, 1 >$ and $val_i \in Val$ can represents two values:
  1. the number of local purchases (purchases of local products and services for which customers do not travel) in specific Slovak regions (Slovakia consists of 8 regions) normalized to <0,1> interval
  2. the percentage distribution of local purchases across Slovak region
- And finally, for technical preferences $Tech$, the $val_i \in Val$ represents the number of purchases performed using specific technical parameters normalized to <0,1> interval. The Id represents set of technical parameters and can be defined as:
  $Id = \bigcup\{Top_{10}(browsers), Top_{10}(OS), Top_{10}(devices)\}$.

In the context of traits' characteristics $Trait$, the Val and Id combinations are followed:

- For demography traits:
  - $Val = \{0, 1\}$ and $Id = \{$gender$\}$,
  - $Val = \{0, 1\}$ and $Id = \{$child in family$\}$

- $Val = \{0, 1\}$ and $Id = \{\text{partner}\}$
- $Val = \{0, 1, 2\}$ - three income classes and $Id = \{\text{net income}\}$
- $Val = \{1940, 1941, ..., 2005\}$ and $Id = \{\text{year of birth}\}$
- $Val = \{1, 2, ..., 8\}$ and $Id = \{\text{region}\}$

- For personality traits the $Val =< 0, 1 >$ represents the percentile of the user in a given personality trait defined in the Big Five Personality model. Those percentiles were normalized to the Slovak population. The set of identifiers $Id = \{O, C, E, N, A\}$, where A is Agreeableness, N is Neuroticism, E is Extraversion, C is Conscientiousness and O is Openness to experience.
- In the case of domain characteristics the $Val =< 0, 1 >$ represents the user scoring in a specific domain characteristic. The set of identifiers corresponds to the considered domain characteristics and can be expressed as followed: $Id = \{\text{user value}, \text{user activity}\}$.

In pattern-based characteristics $Pattern$, the $Val = \{0, 1\}$ and represents the pattern occurrence in user historical behaviour. The Id differ across pattern types and can be expressed es followed:

- for Association Rules: $Id = identifiers(AR) \times \bigcup\{BC, LC, EC\}$
- for N-grams: $Id = identifiers(Ng) \times \bigcup\{BC, LC, EC\}$
- for Sequence patterns: $Id = identifiers(SP) \times \bigcup\{BC, LC, EC\}$

where BC corresponds to basic categories (in our case 6 basic categories), LC corresponds to latent categories and EX to expert categories of products. The AR represents the set of all association rules, Ng set of all N-grams and SP set of all sequence patterns. Function $Identiefiers()$ return the identifiers of given patterns (e.g. identifiers('a,b,c'; 'a') returns ('id 1'; 'id 2')).

An example of suggested user model is presented in Table 34.

Latent categories. Because our model is designed with different abstraction levels, we suggested the process of latent categories creation. This process is based on topic modelling and is performed via Latent Dirichlet Allocation (LDA) - one of the most widely used topic modelling algorithm. The input of LDA is formed by text documents (the product description preprocessed to the form of tokens), that are connected to the form of corpus. The result of LDA is a set of k vectors $\{v_1, v_2, ..., v_k\}$, where $k = |documents|$. The each vector v is represented as follows: $v_i = \{TP_1, TP_2, ..., TP_{n;n<=|topics|}\}$. The individual elements of vector v_i corresponds to the pairs: $TP_j = (T_j, P_j)$, where T_j is a identifier of topic j and P_j is the probability of topic T_j for item i. The latent category LC of item i is then defined as the most probable topic: $LC_i = T_m; P_m = max(P_i; P_i \in v_i)$.

Patterns in behaviour. In the context of e-commerce, user activity is often recorded in the form of logs or events. The user behavioural patterns are obtained based on recorded events. However, the low-level representation of events that are linked to the identification number of

Table 34: An example of suggested user model. The user model is defined as set of weighted vectors. Each value of vector is defined as triplet (value identifier, value, weight). All values are normalized to <0,1> interval.

User characteristic	**Vector Representation**
Categorical	Purchases count: (A, 0.5, 1), (B, 0.3, 2), (C, 0.5, 1), ... Purchases price: (A, 0.2, 2), (B, 0.4, 3), (C, 0.1, 1), ... Average rating: (cat. 12, 0.2, 1), (cat. 5, 0.1, 2), ...
Temporal	Day in week: (Mon, 0.1), (Tue, 0.1), ... Month in year: (Jan, 0.4, 1), (Feb, 0.5, 2), ...
Regional	Regions: (BA, 0.3, 2), (BB, 0.2, 3), (KE, 0.5, 2), ...
Technical	OS: (Windows 10, 0.3, 1), (Windows 7, 0.5, 2), ... Devices: (IPhone, 0.6, 1), (Samsung, 0.3, 2), ...
Personality	(Neuroticizm, 0.5, 1), (Extraversion 0.78, 1, ...
Domain	(User Value, 0.5, 1), (Usrer activity, 0.78, 1)
Demography	(Gender, 1, 2), (Child in family, 0, 1), ...
Association rules	(AR 12, 0, 1), (AR 13,1, 2), (AR 56, 0, 2), ...
Sequence patterns	(SP 55, 1, 1), (SP 17, 1, 1), (SP 10, 0, 2), ...
N-grams	(N 14, 0, 1), (N 13, 1, 2), (N 5, 0, 1), (N 4, 1, 1), ...

individual products is often inappropriate - this is the reason for category creation (e.g. latent categories obtained via topic modelling). The process of event abstraction is based on this higher-level categories. For more details see Figure 6.

Domain characteristic: User activity. The user activity reflects the recorded activity of a user in e-commerce website and can be expressed by the followed formula:

$$UA = \sum_{i=1}^{N} Occurrence(Activity_i) * Weight(Activity_i) \tag{26}$$

where N is the number of activity types (e.g. view of product, product retrieval or purchase), $Occurance(Activity_i)$ corresponds to the number of $Activity_i$ type occurrences and the wight of activity type $Activity_i$ is represented by $Weight(Activity_i)$. The weight of activity type expressed the level of effort that user had to perform for specific activity type. The weights in the same way as types of activities are e-commerce specific. In our case, these wights were determined based on the user behaviour analysis and domain expert consultations.

Domain characteristic: User value. The user value is time-specific attribute and reflects the current value of user in e-commerce. This value is influenced by several characteristics and can be

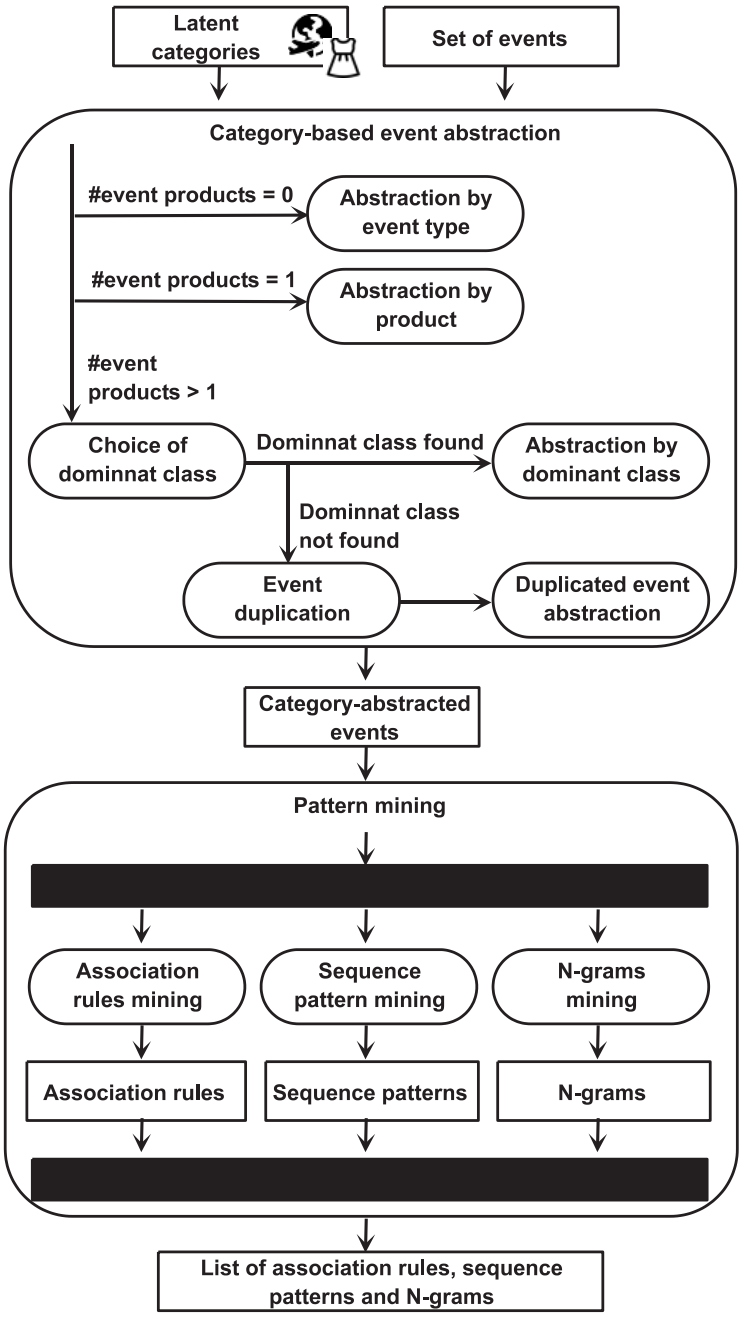

Figure 6: Method of pattern recognition consists of two steps: category-based event abstraction and pattern mining. The input of category-based event abstraction consists of latent categories and set of events. The category-based event abstraction is based on products linked to the specific event and its latent categories. The output of this abstraction is set of category-abstracted events. The next step, pattern-mining, contains parallel pattern mining tasks - association rules mining, sequence pattern mining, N-grams mining. The output of this method is set of recognized patterns.

defined as followed:

$$UV = 50 * CardS * LastPurch \sum_{i=1}^{N} Price(Purch_i) * TM(Purch_i) \tag{27}$$

where the $Price(Purch_i)$ corresponds to the price of purchase i, N represents the number of user purchases, $CardS$ reflects the usage of card saving and is equivalent to 1.05 if user used this option and 1 otherwise.

The $LastPurch$ represents the last purchase time and is equivalent to 1.1 if $Month(ActualTime - Time(LatPurch)) \leq 2$ and 1 othewise.

The time of specific purchase is reflected by $TM(Purch_i)$. This time multiplier is defines as:

$$TM(Purch_i) = 1.1 - \lceil Month(ActualTime - Time(Purch_i)) \rceil \tag{28}$$

Individual constants were set based on user behaviour analysis.

4.2 User model evaluation

The evaluation of suggested user model was performed as recommendation task, where baseline based on user purchases was expanded by user latent features - our user model characteristics. In this way, we work with equivalent weights of individual features. Our train and test dataset was based on user purchases which were divided by time identifier. The train dataset corresponds to the time interval two years (1.12.2015 - 1.12.2017) and test dataset to two months (purchases from interval 2.12.2017 - 2.2.2018). User model characteristics that are time-specific, e.g. the number of rating in categories or patterns in behaviour, were computed using a time interval of train dataset (there are also long-term characteristics, e.g. demography that is id not have to be computed).

From recommendation algorithm, the matrix factorization was selected. More precisely, we choose the hybrid matrix factorisation model that was described in [83]. Its implementation is available in LightFM[8]library. User model characteristics were used as user latent features. The evaluation was performed using followed parameter settings: loss function - Weighted Approximate-Rank Pairwise WARP [154], number of components - 30 and number of epochs - 30. Other parameters were set to default values.

The recommendation performance was evaluated using two metrics: Mean average precistion (MAP) and Normalized discounted cumulative gain (nDCG). Due to the random aspect of selected algorithm, we presented the average values of this metrics, that was calculated as the average of 50 algorithm running.

$$MAP = \frac{1}{Q} \sum_{q=1}^{Q} AveP(q) \tag{29}$$

[8]https://lyst.github.io/lightfm/docs/

Table 35: The result of baseline recommendation - hybrid matrix factorization using user purchases dataset.

User group	MAP	nDCG
Users with 1 purchase in train	1.22%	1.67%
Users with 2 purchases in train	1.03%	1.51%
Users with 3-5 purchases in train	9.18%	9.84%
Users with 6-10 purchases in train	6.33%	7.36%
Users with more than 10 purchases in train	4.82%	6.36%
All users	5.51%	6.4%

where Q corresponds to the number of requests and $AveP(q)$ is represented as followed:

$$AveP = \frac{\sum_{k=1}^{N} P@k}{|REL|} \tag{30}$$

where REL represents the relevant documents, N represents the number of recommended items and P@K reflects the Precision at K. Let $TP@K$ corresponds to true positives to position K, the $P@K$ can be expressed as followed:

$$P@K = \frac{TP@K}{K} \tag{31}$$

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{32}$$

where $IDCG_p$ represents the Ideal discounted cumulative gain at position p and can be expressed as followed:

$$IDCG_p = \sum_{i=1}^{|REL|} \frac{2^{rel_i} - 1}{\log_2(1+1)} \tag{33}$$

where $|REL|$ represents the ordered list of relevant documents (ordered by relevance) to position p.

The DCG_p represent the Discounted cumulative gain at position p and can be expressed by following formula:

$$IDCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(1+1)} \tag{34}$$

The result of our baseline model is presented in Table 35, where the results for individual user groups (based on a number of purchases) are presented.

4.3 Dataset Description

Our work is dedicated to the domain of e-commerce. Our dataset was obtained from private Slovak e-commerce company ZlavaDna[9]and is not freely available. ZlavaDna is discount portal, where customers buy products in the form of discount coupons. These coupons are subsequently used in

deal providers. ZlavaDna is specific with two characteristics: short-time deals (average deal activity is 2.5 week) and diverse offer (deals are structured into 6 basic categories: food, travel, health and beauty, sport, goods and services). This dataset includes following data:

- Transaction data (3 million of purchases, 500 000 customers)
- User activity on the website (18.5 million of events)
- Item general information
- Item categories
- Customer general information
- Customer ratings

Dataset preprocessing. Since the user activity in the form of click-stream contained a large number of error elements, the preprocessing had to be done. Dataset preprocessing includes following steps: dataset cleaning, cookies - user ID mapping and session identification. In dataset cleaning, the error and irrelevant events were deleted. Because in ZlavaDna sign-up is not necessary, we had to solve the problem of logged-in and not logged-in users. The activity events had to be back mapped to the real logged-in user identifiers. And finally, because the recorded events did not contain explicit session identifiers, they had to be identified. The session identifier timeout was set on 25 minutes. The dataset included 3.3 million of sessions with the average length 4.93 events.

Acquisition of demography information and personality information. Because our user model contains different types of information that are in the context of e-commerce not always available (in our dataset were not available), we had to suggest the process of its acquisition. The process of demography and personality data acquisition consists of followed steps:

1. demography and personality acquisition via questionnaire created as a result of project HIBER[10]on the e-commerce website
2. demography acquisition through social network scraping

The result of data acquisition is presented in Table 36.

Data for the recommendation. Because in our research questions we focused on different characteristic types comparison, we worked with exactly those users that had most of these characteristics. Precisely the following requirements had to be done:

- user completed the questionnaire and have a personality characteristics.

[9]https://www.zlavadna.sk/

[10]Project APVV-15-0508, HIBER - Human Information Behaviour in the Digital Space is based on interdisciplinary cooperation between FIIT STU and FiF UK [21, 22]. With this cooperation, we have created the questionnaire focused on different user characteristics. This questionnaire was published on ZlavaDna website and was completed by 4 443 users

Table 36: The result of demography and personality characteristics acquisition.

Characteristic	Original amount	Amount after acquisition
Partner relationship	3 428	12 224
Child in family	0	8 869
User residence region	58 825	91 668
Net income	9 236	11 325
Age	3 428	18 600
Personality traits	0	4 443

- user had at least 1 purchase in the period 1.12.2015 - 1.12.2017.
- user had at least 1 session activity in the period 1.6.2016 - 1.12.2017.

Based on these criteria the 2102 users where selected. These users were our train dataset. Test dataset includes 360 users with purchases in 2.12.2017 - 2.2.2018 time interval.

4.4 RQ1: How the individual parts of user model (preference characteristics, pattern characteristics and user traits) influence the result of the personalized recommendation in the domain of e-commerce with short-time deals?

As recent research showed every characteristic type of our model (preference-based characteristics, pattern-based characteristics and user traits) has a potential to increase the recommendation performance. However, real recommendation result is e-commerce specific. Our first research question focused on individual characteristic types and its impact on the recommendation in the context of e-commerce with short-term deals.

Because by increasing the number of user latent characteristics we can also increase the matrix sparsity, it is not always appropriate to use all characteristics - better is to look for the most appropriate combination. That is the reason why we perform the evaluation in two steps - evaluation on the lowest level of characteristic types (e.g. extraversion, gender or purchases as a binary attribute in basic categories) and evaluation on levels presented in 4.1 (e.g. personality, demography or purchases). On the lowest level of characteristic types also the combinations within one characteristic type where considered (e.g. extraversion & openness or net income & gender).

The result of the evaluation is presented in Table 37. As can be seen, the majority of user characteristics made the result of recommendation worse. This caused because of increasing of matrix sparsity. However, there also some types of characteristics, that improve recommendation - preferences expressed via purchases on higher-level characteristic types and on lower-level characteristics types followed characteristics that express preferences via (we presented the top 5 characteristics - in general, 24 characteristics and its combinations outperform the result of recommendation baseline):

1. purchases in expert categories as binary attribute
   MAP=6.59%, nDCG=7.62%

2. percentage ratio of purchases in expert categories
   MAP=6.52%, nDCG=7.64%

3. percentage ratio of purchases in latent categories
   MAP=6.27%, nDCG=7.42%

4. purchases in latent categories as binary attribute
   MAP=6.23%, nDCG=7.27%

5. normalized count of purchases in latent categories
   MAP=6.18%, nDCG=7.17%

For finding the best performing characteristic combination is, therefore, necessary also look at the combination between different characteristics types (e.g. combination between personality characteristics and user preferences expressed via purchases).

As can be seen in Table 37 only one type of triple Traits, Preferences and Patterns in behaviour outperform the recommendation baseline - it was user preferences, precisely user categorical preferences. However, the combination of Traits and Pattern with Preference characteristics improve the recommendation too.

4.5 RQ2: What is the impact of personality-based characteristics in combination with other characteristics types on the personalized recommendation?

Personality traits as a part of the user model and personality-based recommendation were discussed in several research thesis. In our second research question, we, therefore, focused more closely on its combination with different characteristics' types that create our user model.

As can be seen in Table 38 personality traits and also their combination do not improve our baseline recommendation, where MAP=5.51% and nDCG=6.4%. However, using personality traits in combination with other characteristics types (user category preferences) can improve the baseline recommendation (see Table 39). Combination of preferences expressed as purchases in expert categories and user extraversion improves the baseline recommendation by 1.14% in MAP and 1.3% in nDCG metric. Personality characteristics itself mostly improve the result of demography characteristics and user temporal preferences - however, this combination does not perform as well as combination with user category preferences presented in Table 39.

4.6 Summary of user model for recommendation in e-commerce

User modelling is one of the ways, how the results of recommendation can be improved. In this chapter of our work, we have presented a created user model based on e-commerce click-stream data

Table 37: The impact of different characteristic types on the result of recommendation. The followed characteristics were considered: Personality, Demography, Purchases, Ratings, Views, Domain characteristics, Technical preferences, Temporal and Regional preferences. BP represents the best performing characteristic or best performing characteristic combination from a given characteristic type.

Characteristics	MAP	nDCG
BP in Pesonality: Extraversion	4.74%	5.85%
BP in Pesonality: Openess	4.72%	5.81%
Personality	**3.84%**	**3.84%**
BP in Demography: Gender	5.04%	6.03%
BP in Demography: Child in family	4.99%	5.99%
Demography	**4.15%**	**4.25%**
BP in Purchases: Expert categories, binary attribute	6.23%	7.62%
BP in Purchases: Latent categories, percentage ratio	6.27%	7.42%
Purchases in categories	**5.8%**	**6.88%**
BP in Ratings: Expert categories, binary attribute	5.62%	6.83%
BP in Ratings: Latent categories, normalized quantity	5.3%	6.31%
Ratings in categories	**4.6%**	**5.72%**
BP in Views: Expert categories, percentage ratio	5.61%	6.7%
BP in Views: Basic categories, normalized quantity	5.45%	6.36%
Views in categories	**4.76%**	**5.89%**
BP in Domain: User activity	5.47%	6.31%
Domain characteristics	**5.32%**	**6.19%**
BP in Patterns: Patterns in expert categories	4.44%	5.36%
BP in Patterns: N-grams in latent categories	4.24%	5.19%
Patterns in behaviour	**2%**	**2.87%**
Technical preferences	**3.49%**	**4.38%**
BP in Temporal: Weekend	5.16%	6.04%
BP in Temporal: Day in week	5.03%	5.98%
Temporal preferences	**4.55%**	**5.49%**
BP in Regional preferences: Normalized quantity	5.4%	6.57%
Regional preferences	**4.91%**	**6.22%**

Table 38: The impact of personality traits on the result of recommendation.

Personality characteristics	**MAP**	**nDCG**
Agreeableness	4.67%	5.75%
Openess	4.72%	5.81%
Conscientiousness	4.65%	5.73%
Extraversion	4.74%	5.85%
Neuroticism	4.65%	5.69%
Agreeableness & Openess	4.42%	5.44%
Agreeableness & Conscientiousness	4.28%	5.33%
Agreeableness & Extraversion	4.39%	5.45%
Agreeableness & Neuroticism	3.88%	4.96%
Openess & Conscientiousness	4.25%	5.31%
Openess & Extraversion	4.58%	5.67%
Openess & Neuroticism	4.01%	5.04%
Conscientiousness & Extraversion	4.45%	5.5%
Conscientiousness & Neuroticism	4.02%	5.07%
Extraversion & Neuroticism	4.17%	5.27%
overall	3.84%	4.79%

Table 39: The combinations of personality-based characteristics with user category preferences, that improve the recommendation in better way (PUR represents purchases, RAT rating, EC expert categories, LC latent categories, B representation as binary attribute, C representation as normalized count, R representation as percentage ratio, E user extraversion and 0 user openness to experiences - in this way the PUR-LC_R & E represents the combination of purchases in latent categories represented as percentage ratio and user extraversion).

Characteristics	**Original values**		**Values with personality**	
	MAP	**nDCG**	**MAP**	**nDCG**
PUR-LC_R & E	6.27%	7.42%	6.31%	7.45%
PUR-EC_R & E	6.52%	7.64%	6.53%	7.57%
PUR-EC_C & E	5.95%	6.95%	5.96%	6.7%
PUR-EC_B & E	6.59%	7.62%	6.65%	7.7%
PUR-EC_R & O	6.52%	7.64%	6.56%	7.59%
PUR-EC_B & O	6.59%	7.62%	6.63%	7.68%
RAT-EC-B & O	5.62%	6.83%	5.75%	6.93%

and user profiles. The suggested user model is defined as triplet that consists of user preferences, user traits and patterns in user behaviour.

Our user model, its parts, was evaluated using e-commerce dataset. Based on the comparison of recommendation based on user purchases with the recommendation that uses the user latent features we can see that in the context of e-commerce with short-time deals, user category preferences mostly improve the results of recommendation. Combination of preferences expressed as purchases in expert categories and user extraversion improves the baseline recommendation by 1.14% in MAP and 1.3% in nDCG metric. Personality characteristics itself mostly improve the result of recommendation based on demography characteristics and user temporal preferences.

5 Conclusion and future work

At the time of the society informatization, a personalized recommendation has become an integral part of the Web. Efforts to improve the personalized recommendations are still up to date. One of the means to recommendation improvement is user modelling. In our work, we therefore focused on user model suggestion as a way to increase the success of recommendation.

Because of the user modelling complexity (where we have besides user model suggestion also focused on user model characteristics - its acquisition in the form of predictions, abstraction methods and selection of most representative characteristics), we had to study relatively many papers from different research areas. Our suggestion was based on analysis of 167 research papers and books (click-stream abstraction: 76 sources, demography prediction: 42, user modelling: 26, recommendation: 17, psychology: 6).

The main result of our work is three independent thesis chapters that are the basis for the publication of the results in the form of three research papers.

Event abstraction. In 2 we have focused on event abstraction as an essential part of many machine learning algorithms. We have suggested the method of pattern-based users' events abstraction for domains with textual representation of items. The suggested method consists of item abstraction based on Latent Dirichlet Allocation and event abstraction based on pattern recognition.

The evaluation via 3 machine learning tasks (association rules mining, sequence pattern mining and gender prediction) showed that latent categories, suggested as a part of item abstraction, overcome the expert categories in the pattern mining tasks and achieved comparable results in gender prediction task. In this way, they represent an alternative to manual expert categorization. The comparison of three pattern mining approaches - association rules mining, sequence pattern mining and N-grams mining have shown that sequence patterns and N-grams achieved comparable results in gender prediction task -the recall for both approaches was 67%. However, the combination of this approaches increased the recall by 3%, what suggests that the combination of these approaches can improve the performance of machine learning tasks.

Item abstraction part of suggested method generates the space for comparison with other item-based abstraction types (e.g. methods of clustering or classification - however, classification requires the labelled dataset that is not available). In the case of event abstraction, there are no more common pattern mining approaches are available. Our full method - Pattern-based users' events abstraction can be also compared to approaches presented in the field of process mining - however, because the implementation of this methods is not available, this comparison is not a trivial task.

Demography prediction. The 3 is dedicated to demography prediction as basic attributes that influence a shopping behaviour. To create prediction with good performance, the user modelling consisting of 6 feature types (transaction-based, temporal-based, rating-based, domain-based,

technical-based and activity-based features.) was suggested.

The evaluation via 4 prediction tasks (prediction of net income, prediction of partner relationship, prediction of a child in family and prediction of user residence region) showed that temporal-based features improved the net income prediction task by 7.7% and activity-based features by 7.2% in Recall Macro (in comparison with prediction based on transaction-based features, that were our reference features). The improvement by 4% in Recall Macro was recorded using a combination of activity-based features and transaction-based features in the region of user residence prediction task. As experimental results showed, domain-based features mostly influenced net income prediction - improvement by 7.3%, partner relationship prediction - improvement by 4.1% and prediction of a child in the family - improvement by 1.8%.

The comparison with 2 state-of-the-art models, which have been placed in the top 10 in the PAKDD 2015 competition, showed the robustness of our model. In general, our model outperforms these models in each prediction task. In net income prediction task, our model reached 60.6% in Recall Macro which was 6.7% more than Duongs' model and 7.1% more than Lus' model. The final score of our model in partner relationship prediction was 64.6% - Duongs' model lagged by 11.2% and Lus' model by 12.1%. In a child in family prediction, our model obtained score 64.5%, Duongs' model 53.9% (10.6% less than our model) and Lus' model 53.7%. The Recall Macro of our model in the region of user residence prediction was 28.4% which was 8.4% more than the Duongs' model and 7.5% more than Lus' model.

Based on these results, we can conclude that in the context of e-commerce with less user shopping frequency models created on activity-based data are not as effective as models created on a combination of transaction-based data and activity-based data. Based on this results, there is a space for comparison of the suggested model with also other models suggested for the domain of e-commerce - models that work on a combination of transaction-based data and activity-based data. However, in our analysis, we see that many models are focused only on one data source especially in the domain of e-commerce.

User modelling. Chapter 4 focused on user modelling in the domain of e-commerce. In order to analyze the effect of different user characteristics on recommendation, the user model based on e-commerce click-stream data and user profiles were suggested. The suggested user model was defined as a triplet (user preferences, user traits, patterns in behaviour).

Based on the comparison of recommendation based on user purchases with the recommendation that uses the user latent features the evaluation showed that in the context of e-commerce with short-time deals, user category preferences improve the results of recommendation at the most. Combination of preferences expressed as purchases in expert categories and user extraversion improves the baseline recommendation by 1.14% in MAP and 1.3% in nDCG metric.

Based on the results of the evaluation, where a lot of features caused that the result of recommendation was getting worse, the process of finding the optimal feature combination is very important. In the future work is, therefore, space for a suggestion of an intelligent identification of best performing combination (currently we work with brute force method).

Besides three research papers that copy the content of our master thesis, other two papers were already published. The first one focus on net income prediction and was presented in Workshop on Intelligent and Knowledge oriented Technologies 2017 (WIKT 2017) and the second one is dedicated to the event abstraction and was presented on student research conference IIT.SRC. 2018.

While research on real data (where the so-called dirty work is necessary), the output of our work also includes 3 datasets: dataset of click-stream user activity that was cleaned and preprocessed, transaction dataset complemented by demography and domain characteristics and dataset created as the result of cooperation in the project HIBER. Click-stream dataset preprocessing included: dataset cleaning, cookies - user ID mapping and session identification. Dataset as the output of project HIBER was created thanks to the cooperation of several teams. However in accordance with our research, the author of this thesis was actively working on all stages of its formation and subsequent application.

The biggest challenges of our work were followed: click-stream dataset of user activity that was not preprocessed, absence of demography and personality information, implementation of state-of-the-art prediction models. The click-stream dataset represents the real-time e-commerce logs which were not preprocessed yet. The preprocessing required a lot of time to dataset analysis and dataset cleaning. Another challenge that had to be solved was logged-in and not-logged-in user mapping. This mapping was made in many iterations because we did not know the company processes very well at the beginning (e.g. deals can be ordered by company employees on request of customers).

When the demography prediction was made, the lack of labelled samples was detected. As the solution, the process of demography data acquisition from other data sources was suggested. The first approach was a questionnaire made by e-commerce ZlavaDna, FIIT STU and FiF UK. Because of interdisciplinary cooperation, questionnaire realization was quite difficult. The second approach was social network scraping, where the 7 797 unique data inputs had to be labelled manually (it took 22 hours). The last challenge was the implementation of state-of-the-art prediction models, where the individual features were quite poorly described.

To conclude, in our work, we focused on means with a potential to improve the results of recommendation. We dedicated to user modelling and its features. We suggested the method of pattern-based event abstraction, the model for demography prediction and user model for the recommendation in e-commerce. Individual methods were compared with either expert assessment or state-of-the-art approaches. The results of our work can be used as a base for next research (e.g. exploration of an effective way to best-performing features identification).

6 Resumé

Web je zdrojom obrovského množstva informácií a služieb, ktoré sú vzájomne poprepájané. Jednotlivé služby sú prevádzkované sídlami, ktoré tieto služby poskytujú. Konzumenti služieb ich využívajú s cieľom uspokojiť svoje potreby. Maslow vo svojej štúdii [103] rozdeľuje potreby na: základné telesné a fyziologické potreby; potreby bezpečia; potreby lásky, potreby spolupatričnosti a prijatia; potreby uznania a úcty a potreby sebarealizácie. Web predstavuje prostriedok na uspokojenie rôznych typov potrieb, pričom s každým jedným uspokojením prináša pocit satisfakcie. Satisfakcia v psychológii predstavuje mieru spokojnosti s vykonanou činnosťou, v kontexte webu mieru potešenia zo získaných informácií, resp. zo získaných služieb. Na mieru uspokojenia do veľkej miery vplýva výber vhodného zdroja informácií.

Nakoľko web poskytuje obrovské množstvo informácií, používateľ nemá priestor na preskúmanie všetkých alternatív, aby našiel tú najvhodnejšiu. Táto nemožnosť vyplýva z obmedzeného času, ktorý má používateľ na vyhľadanie určitých informácií a taktiež z predpokladu, že samotný používateľ nemusí vedieť, čo je pre neho najvhodnejšie - nemá dostatok informácií a ani nemusí vedieť o vhodnejšej alternatíve. V súčasnej dobe personalizácie webu je snahou poskytovateľov služieb zvýšiť pocit satisfakcie ich zákazníkov na najvyššiu možnú hranicu. Jeden z prostriedkov na zvyšovanie miery uspokojenia používateľa je odporúčanie [146].

Odporúčacie systémy sa zameriavajú na jednotlivcov, ktorí nemajú dostatočné skúsenosti na vyhodnotenie množstva alternatív, ktoré sú im poskytované [122]. Vo všeobecnosti sa odporúčanie využíva v rôznych rozhodovacích procesoch, napr. pri výbere filmu, hudby, alebo produktov v e-obchode. Odporúčacie systémy sú často personalizované t.j. zameriavajú na jednotlivcov. V niektorých situáciách je však vhodné využiť tzv. nepersonalizované odporúčanie, ktoré je jednoduchšie a je založené na top hodnotených položkách. Odporúčanie vo všeobecnosti predstavuje zoznam ohodnotených položiek, ktorý je vytvorený na základe predikcie vzhľadom na preferencie používateľa [124])

V praxi sa využíva niekoľko typov odporúčaní, ktoré sú založené na dôvere ľudí v skupinový názor, názor blízkych osôb alebo v porovnanie s odskúšanými produktmi [100]. Medzi najznámejšie typy odporúčaní patria kolaboratívne odporúčanie a odporúčania založené na obsahu. Kombináciou týchto odporúčacích techník vzniká tzv. hybridné odporúčanie. Burke vo svojej práci [29] definoval 6 základných prístupov odporúčania, ktoré zahŕňajú: kolaboratívne filtrovanie, odporúčanie založené na obsahu, odporúčanie založené na znalostiach, odporúčanie založené na demografii, odporúčanie založené na komunite a odporúčanie založené na užitočnosti.

Prvotnou myšlienkou kolaboratívneho filtrovania podľa Schafer a kol. [125] bolo odporúčanie položiek aktívnym používateľom na základe položiek, ktoré sa v minulosti páčili používateľom s rovnakým vkusom. Podobnosť bola určená na základe historických hodnotení používateľov. V praxi sa stretávame s dvomi základnými typmi kolaboratívneho filtrovania - kolaboratívne filtrovanie založené na pamäti a kolaboratívne filtrovanie založené na modeli [129]. Kolaboratívne filtrovanie založené na pamäti je založené na štatistických metódach, vďaka ktorým hľadá skupinu najpodobnejších používateľov, tzv. susedov, resp. najpodobnejších položiek. Vo všeobecnosti ho môžeme

rozdeliť na kolaboratívne založené na používateľovi (tento prístup sa snaží odhadnúť záujem používateľa na položku na základe hodnotení iných používateľov) a kolaboratívne filtrovanie založené na položke (v tomto prístupe je snaha odhadnúť hodnotenie používateľa na položku na základe hodnotenie tohto používateľa na iné podobné položky - položky sú navzájom podobné v prípade, keď ich viacero používateľov ohodnotí podobným spôsobom). Kolaboratívne filtrovanie založené na modeli je založené na vytvorení modelu hodnotení používateľa. Snahou je predikcia hodnotenia používateľa na položku na základe jeho hodnotení na iné položky. Tvorba modelu je často krát založená na prístupoch strojového učenia.

Základnou myšlienkou odporúčania založeného na obsahu je odporúčanie položiek, ktoré zodpovedajú preferenciám používateľa [116], kde podobnosť položiek je vypočítaná na základe špecifických charakteristík jednotlivých položiek (napr. žáner, kľúčové slová, podobnosť textu a pod.). Zoznam odporúčaných položiek je následne vytvorený ako list top N najpodobnejších položiek k položkám, ktoré používateľ kladne ohodnotil v minulosti. Odporúčanie založené na znalostiach odporúča položky, ktorých črty najlepšie zodpovedajú preferenciám používateľa na základe konkrétnych doménových znalostí [147]. Odporúčanie vychádza z poznatkov, ktoré získava zo správania používateľa v prostredí webovej služby, ako aj zo samotného obsahu. Ďalším zdrojom znalostí sú informácie získané explicitne od experta v danej doméne.

Demografické odporúčanie je založené na výpočte podobnosti používateľov na základe vybranej demografickej charakteristiky. Okrem tradičných demografických údajov, demografické odporúčanie uvažuje informácie ako je prítomnosť zvieraťa v rodine [81]. Odporúčanie založené na komunite vychádza z preferencií priateľov používateľa [19]. Predpokladom tohto odporúčania je väčšia dôvera v priateľov než v iných, síce podobných, ale neznámych používateľov [139]. Odporúčanie založené na užitočnosti odporúča položky na základe funkcie užitočnosti, ktorá je špecifická pre danú doménu. V niektorých prípadoch dosahuje vyššiu presnosť než iné typy odporúčaní [67].

Existuje množstvo štúdií, ktoré za zaoberajú zlepšovaním odporúčačov. Od porovnania jednotlivých odporúčačov v rôznych doménach, cez nastavenia jednotlivých odporúčačov, dolovanie v texte, dolovanie vo webe, modelovania používateľa až po vytváranie profilu používateľa prinášajú mnohé výsledky, ktoré môžu byť v praxi využité. Kopeinik a kol. sa v [80] zameriavajú na porovnávanie množiny odporúčacích algoritmov v doméne vzdelávania. Po porovnaní 6 algoritmov na 6 datasetoch bolo zistené, že najlepšie výsledky dosahuje hybridný odporúčač, ktorý sa skladá z kognitívneho odporúčača a odporúčača založeného na popularite.

Podobné štúdie boli vykonané aj v oblasti e-obchodu. Paraschakis a kol. sa v [112] zamerali na porovnanie niekoľkých typov odporúčačov v kontexte e-obchodu. Ich výskumy dokazujú, že lepšie výsledky dosahujú tzv. tradičné odporúčače. Všeobecne najlepšie výsledky dosahovalo kNN kolaboratívne filtrovanie a dolovanie založené na asociačných pravidlách (opísané v [119]). Zložité odporúčacie algoritmy takúto úspešnosť nedosiahli. Porovnanie dvoch odporúčacích prístupov - Najfrekventovanejšia položka a Odporúčanie založené na asociačných pravidlách v kontexte e-obchodov bolo taktiež vykonané Sarwar a kol. v [129]. Najlepšie výsledky v rámci tohto výskumu dosiahlo odporúčanie založené na asociačných pravidlách. Huang a kol. sa vo svojej práci [69] venovali porovnaniu metód kolaboratívneho filtrovania. Na evaluáciu využili tri datasety (jedlo, móda

a filmy). Najlepšie výsledky na prvých dvoch datasetoch dosiahla metóda Link-analysis a pre tretí dataset metóda založená na položkách. Metóda Link-analysis je typ kolaboratívneho filtrovania, ktorý je založený na výbere tzv. hypertextových indukovaných nadpisoch [70].

Ďalším prístupom, ktorý zvyšuje presnosť odporúčania je štúdium výberu metód dolovania v dátach, ktoré dokážu do veľkej miery ovplyvniť úspešnosť odporúčania. Gibert a kol. sa v [52] zamerali na opis výberu najvhodnejšej metódy dolovania v dátach spomedzi 27 opísaných metód. Prípadovou štúdiou vytvorili akýsi typ odporúčania, ktorý odporúča jednotlivé metódy dolovania v dátach vzhľadom na kontext využitia. Výsledky odporúčania boli vyhodnotené expertne. Segrega a Moreno sa v [132] zaoberali otázkou klasifikátorov a ich využitím v odporúčaní. Vo svojej štúdii analyzovali tri základné typy klasifikátorov: Bayesovský klasifikátor, klasifikátor založený na algoritme najbližších susedov a klasifikátor založený na rozhodovacom strome. Okrem individuálnych klasifikátorov sa zamerali aj na vytváranie tzv. multiklasifikátorov, ktoré vznikajú na základe baggingu, boostingu alebo stackingu. Spomedzi uvedených spôsobov vytvárania multiklasifikátorov bol najlepšie vyhodnotený stacking, ktorý je založený na hybridnej klasifikácii prostredníctvom rôznych učiacich algoritmov.

Ďalším zo spôsobov, ktorým sa dá zvyšovať presnosť odporúčania je doplnenie vstupných dát na základe analýzy správania používateľov. Tento prístup môže byť realizovaný v rôznych rovinách. Môže sa jednať o predikciu správania, hľadanie vzorov v správaní používateľa, alebo tiež modelovanie používateľa. Jednotlivé prístupy môžu byť vzájomne prepojené alebo na seba môžu nadväzovať. Grbovic a kol. sa v [56] zaoberali problematikou predikcie najbližšieho nákupu zákazníka na základe vzorov v nákupnom správaní. Na túto prácu nadviazala práca Kooti a kol. [79], ktorá doplnila predchádzajúcu štúdiu o odhad času a ceny najbližšieho nákupu. V porovnaní s náhodnou metódou, navrhnutá metóda dosiahla signifikantne lepšie výsledky. V otázke ceny položky dosiahla RMSE 0.3806 a v čase nákupu 0.4272.

Medzi práce, ktoré sa zaoberajú štúdiom správania používateľa so zameraním na problematiku nakupovania patrí napr. práca Wang J. a kol. [149], ktorí sa snažili rozlišovať medzi tzv. miestnymi nakupujúcimi a návštevníkmi. Z experimentálnej štúdie im vyplynulo, že miestni sa zameriavajú viac na zľavu než návštevníci a naopak návštevníci sa viacej orientujú na výslednú cenu, dobu využitia a kategóriu, teda typ produktu. Štúdiou vplyvu demografických charakteristík na proces nakupovania sa zaoberali aj Ning a Zhang v [108], ktorí sa snažili usporiadať pojmy: cena, zľava, kategória výrobku a trvanie v závislosti od pohlavia kupujúceho. Z výskumu vyplynulo, že dôležitosť jednotlivých atribútov u žien je: cena, zľava, kategória výrobku a trvanie, zatiaľ čo u mužov: cena, kategória, trvanie a zľava. Okrem demografických čŕt používateľa má na jeho správanie výrazný vplyv jeho motivácia. Tejto problematike sa venovali Fronimos a Kourouthanassis v [48], ktorí vytvorili štúdiu motivácie zákazníka e-obchodu. Cieľom tejto štúdie bolo vytvorenie typov zákazníkov na základe ich motivácie nakupovať. Z výsledkov analýzy 45 vedeckých prác bolo vytvorených 5 základných typov používateľov (na základe motivácie): apatetický, pohodlný, tradičný, entuziastický a hedonický nakupujúci.

Vzhľadom na vyššie uvedené, úspešnosť odporúčania je ovplyvnená niekoľkými faktormi, ktoré zahŕňajú dáta, ktoré sú pri tvorbe odporúčania k dispozícii (dáta o používateľoch a sledovanom

prostredí), taktiež metódy využité na interpretáciu dát ako aj vybratý odporúčací prístup. V našej práci sa venujeme spracovaniu dát pre odporúčanie s cieľom priniesť nové informácie, ktoré dokážu zvýšiť úspešnosť odporúčania a tým zvýšiť mieru satisfakcie zákazníka. Tieto informácie získavame metódami strojového učenia, ktoré sú využité ako prostriedok pre analýzu správania sa používateľa. Venujeme sa modelovaniu používateľa. Podľa psychologickej teórie [34] môžeme osobu vnímať ako súhrn charakteristík, rysov a čŕt daného ľudského jedinca. Táto myšlienka sa využíva v doméne modelovania používateľa v on-line prostredí, kedy sa model používateľa najčastejšie vytvára ako súbor charakteristík konkrétnej osoby, t.j. konkrétneho používateľa [41]. Charakteristiky človeka zároveň predstavujú základný faktor ovplyvňujúci správanie jedinca [12]. Tento fakt sa využíva pri predikcii vlastností jedinca na základe informačného správania.

V našom návrhu vytvárame model používateľa, s cieľom zvýšiť presnosť odporúčania, na základe dvoch pohľadov: samotných charakteristík používateľa a správania sa používateľa v prostredí webovej služby. Medzi charakteristiky používateľa môžeme zaradiť: demografické charakteristiky, doménové charakteristiky, webovú gramotnosť, osobnostné charakteristiky používateľa, jeho ciele, záujmy alebo motiváciu. Jeden zo spôsobov ako sa tieto charakteristiky môžu predikovať je prostredníctvom správania sa používateľa v sledovanom systéme. My sa venujeme najmä predikcii demografických charakteristík a modelovaniu používateľa, ktorý je založený na kombinácii správania sa používateľa a jeho dlhodobých charakteristík ako je napr. osobnosť alebo demografia.

Naša práca je zameraná na modelovanie používateľa v systémoch poskytujúcich služby. Primárne sa zameriavame na poskytovateľov ako sú e-obchody alebo e-banky. Analyzujeme reálne dáta, získané z komerčných projektov, ktoré zachytávajú stopy používateľov na webe. Návrh sme realizovali nad datasetom zo Zľavy Dňa[11], ktorý je bohatý na rôzne typy dát. Tento dataset obsahuje - transakčné dáta používateľov; aktivitu používateľov na webe; demografické dáta používateľov; texty jednotlivých ponúk, ako aj komunikáciu používateľov so zákazníckym centrom za obdobie 5 rokov. Ďalší zber dát v podobe dotazníkov a testovaní sme realizovali priamo na stránke Zľavy Dňa, čo nám umožnilo získať ďalšie dodatočné informácie o používateľoch.

Časť našej práce je priamo prepojená s projektom HIBER, ktorého cieľom je lepšie interpretovanie stôp, ktoré za sebou používatelia zanechávajú v digitálnom priestore [21, 22]. Projekt APVV-15-0508, HIBER – Human Information Behavior in the Digital Space (Informačné správanie sa človeka v digitálnom priestore) je založený na interdisciplinárnej spolupráci medzi FIIT STU a FiF UK. Vďaka tejto spolupráci vznikol osobnostný dotazník zameraný na rôzne ľudské vlastnosti, ktorý bol zverejnený na stránke ZľavyDňa a vďaka ktorému sa nám podarilo získať 4 443 kompletných vyplnení tohto dotazníka. Formuláciu osobnostného dotazníka a jeho vyhodnotenie zastrešila Katedra psychológie, FiF UK. Samotnú prezentáciu na stránke ZľavyDňa sme zastrešili my (Ústav informatiky, informačných systémov a softvérového inžinierstva (FIIT)) v spolupráci s tímom marketingu a vývoja na strane ZľavyDňa. Podporu dotazníka zabezpečoval marketing ZľavyDňa a technickú realizáciu my v spolupráci s ďalšími členmi projektu HIBER na strane FIIT STU.

[11] https://www.zlavadna.sk/

6.1 Otvorené výskumné otázky a ciele

Ako sme už spomenuli, úspešnosť odporúčania je ovplyvnená niekoľkými faktormi, medzi ktoré patria dáta, ktoré máme k dispozícii a s ktorými môžeme pri odporúčaní pracovať. Avšak, dostupný dataset nemusí niesť dostatočnú informáciu na efektívne odporúčanie - toto je priestor pre našu prácu. Hlavným cieľom našej práce je odpovedanie na nasledovnú otázku:

RQ Z: Ako je vhodné predspracovať dáta z e-obchodu na získanie dodatočných informácií, ktoré môžeme využiť pri odporúčaní?

Nakoľko táto otázka je pomerne všeobecná a rozsiahla, pozreli sme sa na ňu z viacerých perspektív.

V kontexte datasetov poskytovaných e-obchodmi môžeme pracovať s tromi typmi dát: transakčné dáta, aktivita používateľov alebo profily používateľov a položiek (napr. opis ponuky). Transakčné dáta a aktivita používateľov na webe je zvyčajne zaznamenávaná v podobe prúdu dát. V prvom kroku našej práce sme sa z tohto dôvodu zamerali na spracovanie prúdu dát, nakoľko ten je v podobe transakčných dát dostupný v každom e-obchode (transakčné dáta môžu byť taktiež doplnené o aktivitu používateľov na webe). Našim cieľom bolo nájdenie vhodnej metódy abstrakcie prúdu dát do podoby, ktorá je vhodná pre úlohy strojového učenia. Našou základnou výskumnou otázkou bolo:

RQA: Aká je najvhodnejšia metóda abstrakcie eventov pre spracovanie prúdu klikov ako vstupu pre úlohy strojového učenia?

Ako ukázali mnohé výskumy, demografické charakteristiky sú jeden z faktorov, ktorý do veľkej miery ovplyvňuje ľudské správanie a ľudské rozhodovanie [108]. Ľudské rozhodovanie pri odporúčaní zodpovedá za výsledky samotného odporúčania. Zároveň jeden zo základných typov odporúčanie je založený na demografických charakteristikách [12, 81]. Toto boli dôvody, prečo sme sa v druhom kroku zamerali na demografické charakteristiky. Naša druhá výskumná otázka bola nasledovná:

RQB: Ako vieme využiť metódy inžinierstva zameraného na črty *(z angl. feature engineering)* pre efektívne predikovanie demografických charakteristík?

V praxi sa stretávame s viacerými typmi odporúčania. Väčšina odporúčacích techník je založená na modeli používateľa [27, 160]. Z tohto dôvodu je modelovanie používateľov jedna z najdiskutovanejších tém v oblasti výskumu v odporúčaní. Posledná časť nášho výskumu je zameraná na modelovanie používateľa a preskúmanie charakteristík modelu používateľa. Naša tretia výskumná otázka je nasledovná:

RQC: Ako komponenty komplexného modelu používateľa ovplyvňujú výsledky odporúčania?

6.2 Štruktúra práce

Naša práca je štruktúrovaná do troch kapitol a zhrnutia. Jednotlivé kapitoly boli písané ako nezávislé časti a z tohto dôvodu majú niektoré časti spoločné (napr. opis datasetu, opis metód, ktoré sú využité vo viacerých častiach). Dôvodom pre písanie jednotlivých kapitol ako nezávislých častí

bol náš cieľ využiť tieto kapitoly ako základ pre publikáciu výsledkov v podobe troch výskumných článkov.

Hlavná časť našej práce je napísaná v anglickom jazyku. Vzhľadom na fakt, že predchádzajúce verzie našej práce (Diplomový projekt I a diplomový projekt II) boli napísané v slovenskom jazyku, niektoré prílohy sú v taktiež napísané v slovenskom jazyku.

Jednotlivé kapitoly našej práce sú nasledovné:

Abstrakcia eventov pre predikciu demografie. Táto kapitola sa zameriava na problém abstrakcie eventov, ktoré majú potenciál stať sa efektívnym zdrojom informácií pre úlohy strojového učenia. V tejto kapitole sme navrhli metódu abstrakcie eventov založenú na hľadaní vzorov. Táto metóda je určená pre domény s textovou reprezentáciou položiek (napr. e-obchody, e-banky alebo časopisy). Navrhnutá metóda abstrakcie sa skladá z dvoch základných častí - abstrakcie položiek a abstrakcie eventov. Abstrakcia položiek je vnímaná ako problém modelovania nadpisov *(z angl. topic modelling)* a je spojená s úlohou predspracovania textov, ktorá je pre ňu nevyhnutná. Abstrakcia eventov je založená na metóde rozpoznávania vzorov, ktorá zahŕňa tri typy hľadania vzorov - hľadanie asociačných pravidiel, hľadanie sekvenčných vzorov a hľadanie N-gramov. Naša metóda bola vyhodnotená využitím datasetu z e-obchodu. Evaluácia bola vykonaná v dvoch krokoch - evaluácia abstrakcie položiek (porovnanie s expertnými kategóriami zadanými doménovým expertov prostredníctvom troch úloh strojového učenia - dolovanie asociačných pravidiel, dolovanie sekvenčných vzorov a predikcia pohlavia) a evaluácia abstrakcie eventov (porovnanie troch metód hľadania vzorov v úlohe predikcie pohlavia).

Koncept abstrakcie položiek a koncept evaluácie prostredníctvom dolovania asociačných pravidiel a dolovania sekvenčných vzorov vznikli ako výstup v rámci predmetu Objavovania znalostí v spolupráci s mojou spolužiačkou Zuzanou Bobotovou na základe konzultácii s Ing. Michal Kompan, PhD. Evaluácia ako taká, samotné experimentovanie, ako aj text kapitoly sú výsledkom našej samostatnej práce.

Predikcia demografie v e-obchode. Táto kapitola diskutuje problém predikcie demografie v kontexte e-obchodov. V tejto kapitole sme navrhli model používateľa ako vstupu do úlohy predikcie demografie. Navrhnutý model používateľa bol vytvorený vzhľadom na dostupnosť dát v e-obchode a skladá zo 6 typov charakteristík - transakčné charakteristiky, časové charakteristiky, charakteristiky založené na hodnotení, doménové charakteristiky, technické charakteristiky a charakteristiky založené na aktivite používateľov. Navrhnutý model bol overený prostredníctvom 4 predikčných úloh – predikcii čistého mesačného príjmu, predikcii partnerského vzťahu, predikcii dieťaťa v rodine a predikcii kraja, v ktorom používateľ žije. Evaluácia bola vykonaná v dvoch krokoch - evaluácia príspevku jednotlivých typov charakteristík vzhľadom na predikciu demografie (porovnanie transakčných charakteristík s kombináciami transakčných charakteristík a časových charakteristík/charakteristík založených na aktivite používateľov/doménových charakteristík prostredníctvom 4 predikčných úloh) a evaluácia nami navrhnutého modelu v porovnaní s inými sofistikovanými modelmi navrhnutými pre predikciu demografie v doméne e-obchodov.

Modelovanie používateľa pre odporúčanie v e-obchode. Posledná kapitola sa zameriava na problém modelovania používateľa v doméne e-obchodov. V tejto kapitole je prezentovaný návrh modelu používateľa, ktorý je založený na prúde klikov používateľov a profiloch používateľov, ktoré sú bežne dostupné v doméne e-obchodov. Navrhnutý model používateľa je definovaný ako trojica (preferencie používateľa, črty používateľa, vzory v správaní používateľa na webe). Model používateľa bol vyhodnotený prostredníctvom úlohy odporúčania položiek v e-obchode. Evaluácia bola vykonaná v dvoch krokoch - evaluácia jednotlivých častí modelu používateľa (vplyv týchto charakteristík na výsledky odporúčania) a vyhodnotenie vplyvu personálnych charakteristík, na ktoré sme sa sústredili podrobnejšie (vplyv personálnych charakteristík v kombinácii s inými typmi charakteristík na výsledky odporúčania).

6.3 Zhrnutie a ďalšia práca

V dobe informatizácie spoločnosti sa personalizované odporúčanie stalo neoddeliteľnou súčasťou webu. Snaha o zlepšovanie výsledkov personalizovaného odporúčania je stále aktuálna. Jeden zo spôsobov zlepšovania odporúčania je modelovanie používateľa. V našej práci sa venujeme návrhu modelu používateľa ako prostriedku pre zvýšenie presnosti odporúčania.

Vzhľadom na komplexnosť modelovania používateľa (kde sme sa okrem návrhu samotného modelu používateľa taktiež zamerali na charakteristiky tvoriace model používateľa - ich získavanie v podobe predikcií, metódy abstrakcie, ako aj výber reprezentatívnych charakteristík), naša práca vyžadovala štúdium viacerých výskumných prác z rôznych výskumných oblastí. Náš návrh je založený na analýze 167 výskumných prác (abstrakcia udalostí: 76 zdrojov, predikcia demografie: 42 zdrojov, modelovanie používateľa: 26 zdrojov, odporúčanie: 17 zdrojov a 6 zdrojov z oblasti psychológie).

Hlavným výsledkom našej práce sú tri nezávislé kapitoly, ktoré sú základom pre publikáciu výsledkov v podobe troch výskumných článkov.

Abstrakcia eventov. V 2 sme sa zamerali na abstrakciu eventov ako na podstatný vstup pre úlohy strojového učenia. My sme navrhli metódu abstrakcie eventov založenú na hľadaní vzorov. Táto metóda bola vytvorená pre domény s textovou reprezentáciou položiek. Navrhnutá metóda sa skladá z abstrakcie položiek založenej na algoritme Latent Dirichlet Allocation a abstrakcie eventov založenej na rozpoznávaní vzorov.

Evaluácia prostredníctvom 3 úloh strojového učenia (dolovanie asociačných pravidiel, dolovanie sekvenčných vzorov a predikcia pohlavia) ukázala, že latentné kategórie navrhnuté ako súčasť abstrakcie položiek, prekonali expertné kategórie v úlohách dolovania vzorov a zároveň dosiahli porovnateľné výsledky v úlohe predikcie pohlavia. Vzhľadom na tieto výsledky, latentné kategórie predstavujú alternatívu k manuálnej expertnej kategorizácií položiek. Porovnanie troch úloh dolovania vzorov - dolovania asociačných pravidiel, dolovania sekvenčných vzorov, dolovania N-gramov, ukázalo, že sekvenčné vzory a N-gramy dosahujú porovnateľné výsledky v úlohe predikcie demografie (pokrytie pre obidva prístupy bolo 67%). Avšak, kombinácia týchto prístupov dokázala zvýšiť

pokrytie o ďalšie 3%, čo naznačuje, že kombinácia viacerých prístupov môže zvýšiť úspešnosť úloh strojového učenia.

Časť abstrakcia položiek navrhnutej metódy generuje priestor pre porovnanie s ďalšími metódami abstrakcie položiek (napr. metódami zhlukovania alebo klasifikácie - avšak klasifikácia vyžaduje označkovaný dataset, ktorý nemáme k dispozícii). V prípade abstrakcie eventov nemáme k dispozícii iné používané metódy abstrakcie eventov. Avšak naša celková metóda môže byť taktiež porovnaná voči metódam dolovania v procesoch - nakoľko však implementácie týchto metód nie sú priamo k dispozícii, táto úloha by nebola triviálna.

Predikcia demografie. Kapitola 3 je zameraná na predikciu demografie ako základného atribútu, ktorý ovplyvňuje nákupné správanie ľudí. Za týmto účelom sme navrhli model používateľa, ktorý sa skladal zo 6 typov charakteristík (transakčné charakteristiky, časové charakteristiky, charakteristiky založené na hodnoteniach, doménové charakteristiky, technické charakteristiky a charakteristiky založené na aktivite používateľov).

Evaluácia prostredníctvom 4 predikčných úloh (predikcia čistého mesačného príjmu, predikcia partnerského vzťahu, predikcia dieťaťa v rodine, predikcia kraja v ktorom používateľ žije) ukázala že časové charakteristiky zlepšia predikciu čistého mesačného príjmu o 7.7% a charakteristiky založené na aktivite o 7.2% v miere Pokrytie Makro (v porovnaní s transakčnými charakteristikami, ktoré predstavujú náš referenčný model). Zlepšenie o 4% v miere Pokrytie Makro bolo zaznamenané využitím kombinácie transakčných charakteristík a charakteristík založených na aktivite používateľa v úlohe predikcie kraja, kde používateľ žije. Ako ukazujú experimentálne výsledky, doménové charakteristiky najviac ovplyvňujú predikciu čistého mesačného príjmu - zlepšenie o 7.3%, predikciu partnerského vzťahu - zlepšenie o 4.1% a predikciu dieťaťa v rodine - zlepšenie o 1.8%.

Porovnanie s dvomi modelmi, ktoré sa umiestnili na top 10 pozíciách v rámci súťaže PAKDD 2015, ukázalo robustnosť nášho modelu. Vo všeobecnosti náš model prekonal obidva modely vo všetkých predikčných úlohách. V úlohe predikcie čistého mesačného príjmu náš model dosiahol 60.6% v miere Pokrytie Makro, čo bolo o 6.7% viac ako model vytvorený Duong a 7.1% viac ako model vytvorený Lu. Výsledné skóre nášho modelu v úlohe predikcie partnerského vzťahu bolo 64.6% - model vytvorený Duong zaostal o 11.2% a model vytvorený Lu o 12.1%. V úlohe predikcie dieťaťa náš model dosiahol skóre 64.5%, model vytvorený Duong 53.9% (10.6% menej než náš model) a model vytvorený Lu 53.7%. Pokrytie Makro nášho modelu v úlohe predikcie kraja, kde používateľ žije bola 28.4%, čo je o 8.4% viac ako model vytvorený Duong a o 7.5% viac ako model vytvorený Lu.

Vzhľadom na tieto výsledky môžeme konštatovať, že v kontexte e-obchodov z nižšou frekvenciou nákupov, modely vytvorené na základe aktivity používateľov nie sú natoľko efektívne ako modely vytvorené na kombinácií transakčných dát a aktivity používateľov. Vzhľadom na tieto výsledky sa otvára priestor na porovnanie nášho modelu s modelmi vytvorenými v doméne e-obchodov na predikciu demografie, ktoré sú zároveň založené na kombinácii transakčných charakteristík a aktivity používateľa. Avšak, ako sme v analýze zistili, väčšina modelov vytvorených v rámci e-obchodov vychádza výlučne z transakčných dát resp. výlučne z aktivity používateľa na webe.

Modelovanie používateľa. Kapitola 4 sa zameriava na modelovanie používateľov v doméne e-obchodov. S cieľom analyzovať vplyv charakteristík používateľa na odporúčanie sme navrhli model používateľa, ktorý bol založený na prúdoch klikov a používateľských profiloch. Navrhnutý model používateľa bol definovaný ako trojica (preferencie používateľa, črty používateľa, vzory v správaní používateľa na webe.)

Na základe porovnania odporúčania založeného na nákupoch s odporúčaním, ktoré využívalo latentné charakteristiky používateľov, evaluácia ukázala, že v kontexte e-obchodov s krátkodobými ponukami, kategorické preferencie používateľa najviac ovplyvňujú výsledky odporúčania. Kombinácia preferencií vyjadrených ako nákupy v expertných kategóriách a extroverziou používateľa dokázala zlepšiť referenčné odporúčanie o 1.14% v metrike MAP a 1.3% v metrike nDCG.

Vzhľadom na výsledky evaluácie, kde sa ukázalo, že väčšina charakteristík referenčné odporúčanie zhoršuje (nakoľko ich pridaním sa zvýši riedkosť matice), proces hľadania optimálnej kombinácie charakteristík je veľmi aktuálny. V budúcej práci je preto priestor na preskúmanie efektívneho hľadania optimálnych kombinácií charakteristík (aktuálne sme hľadanie najvhodnejších kombinácií riešili prístupom hrubej sily, čo nie je efektívne).

Okrem troch výskumných článkov, ktoré kopírujú obsah našej diplomovej práce sme taktiež napísali ďalšie dva vedecké články, prezentované na slovenskom fóre. Prvý článok sa zameriava na predikciu čistého mesačného príjmu a bol prezentovaný na konferencii WIKT 2017. Druhý článok sa zameriava na abstrakciu eventov a bol prezentovaný na študentskej vedeckej konferencii IIT.SRC. 2018.

Nakoľko výskum na reálnych dátach je spojený s veľkým množstvom tzv. špinavej práce, súčasťou výstupu našej práce sú 3 dasety: dataset udalostí, ktorý prešiel predspracovaním, dataset transakcií doplnený o doménové a demografické údaje a dataset vytvorený v rámci projektu HIBER. Predspracovanie datasetu udalosti pozostávalo z odstraňovania chybných udalostí, mapovania používateľov a dopĺňania kategórií. Dataset vytvorený v rámci projektu HIBER bol vytvorený vďaka spolupráce viacerých tímov, pričom v rámci našej práce sa autorka tejto práce aktívne podieľala na všetkých fázach jeho tvorby a následnej aplikácie.

Najväčšie výzvy našej práce boli nasledovné: dataset eventov, ktorý zachytáva aktivitu používateľov a doposiaľ nebol žiadnym spôsobom predspracovaný, absencia demografických údajov a osobnostných údajov, implementácia nových predikčných modelov prezentovaných na PAKDD 2015 konferencii. Samotné predspracovanie datasetu udalostí vyžadovalo veľké úsilie na jeho podrobnú analýzu a následné vyčistenie. Ďalšia výzva, ktorú sme museli vyriešiť bolo mapovanie prihlásených a neprihlásených používateľov. Toto mapovanie bolo realizované na niekoľko pokusov, nakoľko sme sa až postupne dozvedali, že niektoré udalosti môžu vykonávať v mene používateľa zamestnanci.

Ďalšia výzva nastala pri snahe o predikciu demografických charakteristík, kedy sme zistili, že početnosť označkovaných záznamov je pomerne nízka. Za týmto účelom sme navrhli proces dopĺňania týchto údajov z iných zdrojov než je samotný dataset. Prvým pomerne náročným prístupom bol dotazník realizovaný v rámci projektu HIBER. Nakoľko tento dotazník vyžadoval spoluprácu niekoľkých strán (ZľavaDňa, FIIT STU, Katedra psychológie UK), jeho realizácia bola pomerne obtiažna.

Druhým veľmi prácnym prístupom bolo získavanie údajov zo sociálnej sieti, kde sme 7 797 unikátnych viet museli označkovať manuálne (celkovo nám to trvalo 22 hodín čistého času). Poslednou výzvou bola implementácia predikčných modelov navrhnutých v rámci PAKDD 2015, ktoré neboli príliš podrobne opísané, čo sťažovalo ich implementáciu.

Pre zhrnutie, naša práca sa zameriava na prostriedky, ktoré majú potenciál zlepšiť výsledky odporúčania. My sme sa zamerali na modelovanie používateľa a jeho charakteristiky. V rámci našej práci sme navrhli metódu abstrakcie eventov, model pre predikciu demografie ako aj model používateľa pre odporúčanie v e-obchodoch. Jednotlivé metódy boli porovnané voči expertnej metóde resp. voči aktuálnym výskumným metódam. Výsledky našej práce môžu byť využité ako základ pre ďalší výskum (napr. preskúmanie efektívneho hľadania optimálnej kombinácie charakteristík v modeli používateľa).

References

[1] K Chaitanya, Durvasula VLN Somayajulu a P Radha Krishna. "A Novel Approach for Classification of E-Commerce Data". In: *Proceedings of the 8th Annual ACM India Conference*. ACM. 2015, p. 129–134.

[2] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal a Prabhakar Raghavan. "Using taxonomy, discriminants, and signatures for navigating in text databases". In: *VLDB*. Zv. 97. 1997, p. 446–455.

[3] Guibin Chen, Deheng Ye, Zhenchang Xing, Jieshan Chen a Erik Cambria. "Ensemble application of convolutional and recurrent neural networks for multi-label text categorization". In: *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE. 2017, p. 2377–2383.

[4] Jianle Chen, Tianqi Xiao, Jie Sheng a Ankur Teredesai. "Gender prediction on a real life blog data set using LSI and KNN". In: *Computing and Communication Workshop and Conference (CCWC), 2017 IEEE 7th Annual*. IEEE. 2017, p. 1–6.

[5] Lei Chen, Jun Li a Li Zhang. "A method of text categorization based on genetic algorithm and LDA". In: *Control Conference (CCC), 2017 36th Chinese*. IEEE. 2017, p. 10866–10870.

[6] Weihua Chen a Xian Zhang. "Research on text categorization model based on LDA x2014; KNN". In: *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. Mar. 2017, p. 2719–2726. DOI: 10.1109/IAEAC.2017.8054520.

[7] Na Cheng, Xiaoling Chen, Rajarathnam Chandramouli a KP Subbalakshmi. "Gender identification from e-mails". In: *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on*. IEEE. 2009, p. 154–158.

[8] Yoon Ho Cho, Jae Kyeong Kim a Soung Hie Kim. "A personalized recommender system based on web usage mining and decision tree induction". In: *Expert systems with Applications* Vol. 23. No. 3 (2002), p. 329–342.

[9] Jaromır Šavelka a Kevin D Ashley. "Transfer of predictive models for classification of statutory texts in multi-jurisdictional settings". In: *Proceedings of the 15th International Conference on Artificial Intelligence and Law*. ACM. 2015, p. 216–220.

[10] Khaled Abdalgader. "Soft Short-Text Clustering using PageRank as a Centrality Measure". In: *Proceedings of the 9th International Conference on Machine Learning and Computing*. ACM. 2017, p. 452–455.

[11] Charu C Aggarwal a ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012.

[12] Icek Ajzen. "Attitudes, traits, and actions: Dispositional prediction of behavior in personality and social psychology". In: *Advances in experimental social psychology*. Zv. 20. Elsevier, 1987, p. 1–63.

[13] Bader Aljaber, Nicola Stokes, James Bailey a Jian Pei. "Document clustering of scientific texts using citation contexts". In: *Information Retrieval* Vol. 13. No. 2 (2010), p. 101–131.

[14] Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu a Michael Zhu. "A practical algorithm for topic modeling with provable guarantees". In: *International Conference on Machine Learning*. 2013, p. 280–288.

[15] Pelin Atahan. *Learning profiles from user interactions and personalizing recommendations based on learnt profiles*. The University of Texas at Dallas, 2009.

[16] Qiuchan Bai a Chunxia Jin. "Text Clustering Algorithm Based on Semantic Graph Structure". In: *Computational Intelligence and Design (ISCID), 2016 9th International Symposium on*. Zv. 2. IEEE. 2016, p. 312–316.

[17] L Douglas Baker a Andrew Kachites McCallum. "Distributional clustering of words for text classification". In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1998, p. 96–103.

[18] Sugato Basu, Mikhail Bilenko a Raymond J Mooney. "A probabilistic framework for semi-supervised clustering". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2004, p. 59–68.

[19] David Ben-Shimon, Alexander Tsikinovsky, Lior Rokach, Amnon Meisles, Guy Shani a Lihi Naamani. "Recommender system from personal social networks". In: *Advances in Intelligent Web Mastering*. Springer, 2007, p. 47–55.

[20] Bin Bi, Milad Shokouhi, Michal Kosinski a Thore Graepel. "Inferring the demographics of search users: Social data meets search queries". In: *Proceedings of the 22nd international conference on World Wide Web*. ACM. 2013, p. 131–140.

[21] Mária Bieliková, Pavol Návrat, Jakub Šimko, Jozef Tvarožek, Michal Barla, Róbert Móro, Eduard Kuric, Martin Labaj a Martin Konôpka. "Projekt HIBER: hlbšie poznávanie správania sa človeka v digitálnom priestore". In: *11th Workshop on Intelligent and Knowledge Oriented Technologies 35th Conference on Data and Knowledge*. [Online; 30-April-2018]. 2016, p. 141–144. URL: `https://wikt-daz2016.fiit.stuba.sk/wp-content/uploads/2016/11/WIKT-DaZ-2016_Proceedings.pdf`.

[22] Mária Bieliková et al. "Rozsiahla dátová vzorka prepájajúca správanie používateľov e-obchodu s explicitne zistenými osobnostnými charakteristikami". In: *12th Workshop on Intelligent and Knowledge Oriented Technologies*. [Online; 30-April-2018]. 2017. URL: `http://web.tuke.sk/fei-cit/wikt2017/zbornik/s01_paper01x.pdf`.

[23] David M Blei. "Probabilistic topic models". In: *Communications of the ACM* Vol. 55. No. 4 (2012), p. 77–84.

[24] David M Blei, Andrew Y Ng a Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* Vol. 3. No. Jan (2003), p. 993–1022.

[25] Ciro Bologna, Anna Chiara De Rosa, Alfonso De Vivo, Matteo Gaeta, Giuseppe Sansonetti a Valeria Viserta. "Personality-Based Recommendation in E-Commerce." In: *UMAP Workshops*. Citeseer. 2013.

[26] Nesrine Bouadjenek, Hassiba Nemmour a Youcef Chibani. "Age, gender and handedness prediction from handwriting using gradient features". In: *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE. 2015, p. 1116–1120.

[27] Peter Brusilovsky. "Methods and techniques of adaptive hypermedia". In: *Adaptive hypertext and hypermedia*. Springer, 1998, p. 1–43.

[28] John D Burger, John Henderson, George Kim a Guido Zarrella. "Discriminating gender on Twitter". In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics. 2011, p. 1301–1309.

[29] Robin Burke. "Hybrid web recommender systems". In: *The adaptive web*. Springer, 2007, p. 377–408.

[30] William M Campbell, Joseph P Campbell, Douglas A Reynolds, Elliot Singer a Pedro A Torres-Carrasquillo. "Support vector machines for speaker and language recognition". In: *Computer Speech & Language* Vol. 20. No. 2-3 (2006), p. 210–229.

[31] Vitor R Carvalho a William W Cohen. "On the collective classification of email speech acts". In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2005, p. 345–352.

[32] Ling Cen a Dymitr Ruta. "A Map-Based Gender Prediction Model for Big E-Commerce Data". In: *Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2017 IEEE International Conference on*. IEEE. 2017, p. 1025–1029.

[33] William W Cohen et al. "Learning rules that classify e-mail". In: *AAAI spring symposium on machine learning in information access*. Zv. 18. California. 1996, p. 25.

[34] PT Costa a RR McCrae. "NEO five-factor inventory (NEO-FFI)". In: *Odessa, FL: Psychological Assessment Resources* (1989).

[35] Aron Culotta, Nirmal Kumar Ravi a Jennifer Cutler. "Predicting the demographics of twitter users from website traffic data". In: *29th AAAI Conference on Artificial Intelligence, AAAI 2015 and the 27th Innovative Applications of Artificial Intelligence Conference, IAAI 2015*. AI Access Foundation. 2015.

[36] Douglass R Cutting, David R Karger, Jan O Pedersen a John W Tukey. "Scatter/gather: A cluster-based approach to browsing large document collections". In: *ACM SIGIR Forum*. Zv. 51. 2. ACM. 2017, p. 148–159.

[37] Yae Dai, HongWu Ye a SongJie Gong. "Personalized recommendation algorithm using user demography information". In: *Knowledge Discovery and Data Mining, 2009. WKDD 2009. Second International Workshop on*. IEEE. 2009, p. 100–103.

[38] KoenW De Bock a Dirk Van den Poel. "Predicting website audience demographics forweb advertising targeting using multi-website clickstream data". In: *Fundamenta Informaticae* Vol. 98. No. 1 (2010), p. 49–70.

[39] William Deitrick, Zachary Miller, Benjamin Valyou, Brian Dickinson, Timothy Munson a Wei Hu. "Author gender prediction in an email stream using neural networks". In: *Journal of Intelligent Learning Systems and Applications* Vol. 4. No. 03 (2012), p. 169.

[40] Chris Ding, Tao Li a Wei Peng. "Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence chi-square statistic, and a hybrid method". In: *AAAI*. Zv. 42. 2006, p. 137–143.

[41] Peter Dolog a Wolfgang Nejdl. "Challenges and benefits of the semantic web for user modelling". In: *Proceedings of the Workshop on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2003) at 12th International World Wide Web Conference, Budapest*. 2003.

[42] R Douglass. “A cluster-based approach to browsing large document collections”. In: *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1992, p. 318–329.

[43] Duc Duong, Hanh Tan a Son Pham. “Customer gender prediction based on E-commerce data”. In: *Knowledge and Systems Engineering (KSE), 2016 Eighth International Conference on*. IEEE. 2016, p. 91–95.

[44] Mehdi Elahi, Matthias Braunhofer, Francesco Ricci a Marko Tkalcic. “Personality-based active learning for collaborative filtering recommender systems”. In: *Congress of the Italian Association for Artificial Intelligence*. Springer. 2013, p. 360–371.

[45] Deborah Fallows. “How women and men use the Internet”. In: *Pew Internet & American Life Project* Vol. 28 (2005), p. 1–45.

[46] Kostas Fragos a Christos Skourlas. “Ranking tokens with class label frequencies for medical article classification”. In: *Proceedings of the 19th Panhellenic Conference on Informatics*. ACM. 2015, p. 359–360.

[47] Kostas Fragos a Christos Skourlas. “Smoothing Class Frequencies for KNN Medical Article Classification”. In: *Proceedings of the 20th Pan-Hellenic Conference on Informatics*. ACM. 2016, p. 79.

[48] Dimitris Fronimos a Panos Kourouthanassis. “In search for online shopping mission types based on social network analysis”. In: *Proceedings of the 19th Panhellenic Conference on Informatics*. ACM. 2015, p. 289–294.

[49] Nikesh Garera a David Yarowsky. “Modeling latent biographic attributes in conversational genres”. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics. 2009, p. 710–718.

[50] Lars George, Bruno Cadonna a Matthias Weidlich. “IL-miner: instance-level discovery of complex event patterns”. In: *Proceedings of the VLDB Endowment* Vol. 10. No. 1 (2016), p. 25–36.

[51] Norbert Giambiasi a Jean Claude Carmona. “Generalized discrete event abstraction of continuous systems: GDEVS formalism”. In: *Simulation Modelling Practice and Theory* Vol. 14. No. 1 (2006), p. 47–70.

[52] Karina Gibert, Miquel Sànchez-Marrè a Vıctor Codina. “Choosing the right data mining technique: classification of methods and intelligent recommendation”. In: (2010).

[53] Dan Gillick. “Can conversational word usage be used to predict speaker demographics?” In: *Eleventh Annual Conference of the International Speech Communication Association*. 2010.

[54] Sharad Goel, Jake M Hofman a M Irmak Sirer. “Who Does What on the Web: A Large-Scale Study of Browsing Behavior.” In: *ICWSM*. 2012.

[55] Michael Grahame, Jason Laberge a Charles T Scialfa. “Age differences in search of Web pages: The effects of link size, link number, and clutter”. In: *Human Factors* Vol. 46. No. 3 (2004), p. 385–398.

[56] Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, Jaikit Savla, Varun Bhagwan a Doug Sharp. “E-commerce in your inbox: Product recommendations at

scale". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2015, p. 1809–1818.

[57] Nan Guo, Yuan He, ChunGang Yan, Lu Liu a Cheng Wang. "Multi-level topical text categorization with wikipedia". In: *Proceedings of the 9th International Conference on Utility and Cloud Computing*. ACM. 2016, p. 343–352.

[58] Meryeme Hadni a Mounir Gouiouez. "Graph Based Representation for Arabic Text Categorization". In: *Proceedings of the 2nd international Conference on Big Data, Cloud and Applications*. ACM. 2017, p. 75.

[59] Jiawei Han, Jian Pei a Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[60] Susan C Herring a John C Paolillo. "Gender and genre variation in weblogs". In: *Journal of Sociolinguistics* Vol. 10. No. 4 (2006), p. 439–459.

[61] Susan C Herring, Lois Ann Scheidt, Sabrina Bonus a Elijah Wright. "Bridging the gap: A genre analysis of weblogs". In: *System sciences, 2004. proceedings of the 37th annual Hawaii international conference on*. IEEE. 2004, p. 11–pp.

[62] Swapnil Hingmire, Sutanu Chakraborti, Girish Palshikar a Abhay Sodani. "WikiLDA: Towards More Effective Knowledge Acquisition in Topic Models using Wikipedia". In: *Proceedings of the Knowledge Capture Conference*. ACM. 2017, p. 37.

[63] Liangjie Hong a Brian D Davison. "Empirical study of topic modeling in twitter". In: *Proceedings of the first workshop on social media analytics*. ACM. 2010, p. 80–88.

[64] Jozef Hooman a Onno van Roosmalen. "Timed-event abstraction and timing constraints in distributed real-time programming". In: *Object-Oriented Real-Time Dependable Systems, 1997. Proceedings., Third International Workshop on*. IEEE. 1997, p. 153–160.

[65] Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu a Zheng Chen. "Demographic prediction based on user's browsing behavior". In: *Proceedings of the 16th international conference on World Wide Web*. ACM. 2007, p. 151–160.

[66] Rong Hu a Pearl Pu. "Enhancing collaborative filtering systems with personality information". In: *Proceedings of the fifth ACM conference on Recommender systems*. ACM. 2011, p. 197–204.

[67] Shiu-Li Huang. "Designing utility-based recommender systems for e-commerce: Evaluation of preference-elicitation methods". In: *Electronic Commerce Research and Applications* Vol. 10. No. 4 (2011), p. 398–407.

[68] Zan Huang, Wingyan Chung a Hsinchun Chen. "A graph model for E-commerce recommender systems". In: *Journal of the Association for Information Science and Technology* Vol. 55. No. 3 (2004), p. 259–274.

[69] Zan Huang, Daniel Zeng a Hsinchun Chen. "A comparison of collaborative-filtering recommendation algorithms for e-commerce". In: *IEEE Intelligent Systems* Vol. 22. No. 5 (2007).

[70] Zan Huang, Daniel Zeng a Hsinchun Chen. "A link analysis approach to recommendation under sparse data". In: *AMCIS 2004 Proceedings* (2004), p. 239.

[71] Eleonora Ivanova et al. "Predicting website audience demographics based on browsing history". In: (2013).

[72] Chun-Xia Jin a Qiu-Chan Bai. "Text Clustering Algorithm Based on the Graph Structures of Semantic Word Co-occurrence". In: *Information System and Artificial Intelligence (ISAI), 2016 International Conference on*. IEEE. 2016, p. 497–502.

[73] Thorsten Joachims. "Text categorization with support vector machines: Learning with many relevant features". In: *European conference on machine learning*. Springer. 1998, p. 137–142.

[74] Santosh Kabbur, Eui-Hong Han a George Karypis. "Content-based methods for predicting web-site demographic attributes". In: *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE. 2010, p. 863–868.

[75] Misha Kakkar a Divya Upadhyay. "Web Browsing Behaviors Based Age Detection". In: *International Journal of Soft Computing and Engineering (IJSCE) ISSN* (2013), p. 2231–2307.

[76] Ondrej Kassak, Michal Kompan a Maria Bielikova. "User preference modeling by global and individual weights for personalized recommendation". In: *Acta Polytechnica Hungarica* Vol. 12. No. 8 (2015), p. 27–41.

[77] Iljoo Kim. *Predicting Audience Demographics of Web Sites Using Local Cues*. ERIC, 2011.

[78] Michal Kompan a Maria Bielikova. "Group recommendations: survey and perspectives". In: *Computing and Informatics* Vol. 33. No. 2 (2014), p. 446–476.

[79] Farshad Kooti, Kristina Lerman, Luca Maria Aiello, Mihajlo Grbovic, Nemanja Djuric a Vladan Radosavljevic. "Portrait of an online shopper: Understanding and predicting consumer behavior". In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM. 2016, p. 205–214.

[80] Simone Kopeinik, Dominik Kowald a Elisabeth Lex. "Which algorithms suit which learning environments? A comparative study of recommender systems in tel". In: *European Conference on Technology Enhanced Learning*. Springer. 2016, p. 124–138.

[81] Bruce Krulwich. "Lifestyle finder: Intelligent user profiling using large-scale demographic data". In: *AI magazine* Vol. 18. No. 2 (1997), p. 37.

[82] Tayfun Kucukyilmaz, B Barla Cambazoglu, Cevdet Aykanat a Fazli Can. "Chat mining for gender prediction". In: *International Conference on Advances in Information Systems*. Springer. 2006, p. 274–283.

[83] Maciej Kula. "Metadata embeddings for user and item cold-start recommendations". In: *arXiv preprint arXiv:1507.08439* (2015).

[84] Pradnya Kumbhar, Manisha Mali a Mohammad Atique. "A Genetic-Fuzzy Approach for Automatic Text Categorization". In: *Advance Computing Conference (IACC), 2017 IEEE 7th International*. IEEE. 2017, p. 572–578.

[85] Thomas Kunz. *Event Abstraction: Some Definitions and Theorems*. Citeseer, 1993.

[86] Thomas Kunz. "Reverse engineering distributed applications: An event abstraction tool". In: *International Journal of Software Engineering and Knowledge Engineering* Vol. 4. No. 03 (1994), p. 303–323.

[87] Fang-Fei Kuo, Meng-Fen Chiang, Man-Kwan Shan a Suh-Yin Lee. "Emotion-based music recommendation by association discovery from film music". In: *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM. 2005, p. 507–510.

[88] Yasmine Lamari a Saıd Chah Slaoui. “Parallel document clustering using iterative mapreduce”. In: *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*. ACM. 2016, p. 37.

[89] Ken Lang. “Newsweeder: Learning to filter netnews”. In: *Machine Learning Proceedings 1995*. Elsevier, 1995, p. 331–339.

[90] Yong-Bae Lee a Sung Hyon Myaeng. “Text genre classification with genre-revealing and subject-revealing features”. In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2002, p. 145–150.

[91] David D Lewis a Kimberly A Knowles. “Threading electronic mail: A preliminary study”. In: *Information processing & management* Vol. 33. No. 2 (1997), p. 209–217.

[92] Baoli Li. “Selecting Features with Class Based and Importance Weighted Document Frequency in Text Classification”. In: *Proceedings of the 2016 ACM Symposium on Document Engineering*. ACM. 2016, p. 139–142.

[93] Hairong Li, Cheng Kuo a Maratha G Rusell. “The impact of perceived channel utilities, shopping orientations, and demographics on the consumer’s online buying behavior”. In: *Journal of Computer-Mediated Communication* Vol. 5. No. 2 (1999), p. 0–0.

[94] Shangsong Liang, Emine Yilmaz a Evangelos Kanoulas. “Dynamic clustering of streaming short documents”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, p. 995–1004.

[95] Tatiana Litvinova, Pavel Seredin, Olga Litvinova, Olga Zagorovskaya, Aleksandr Sboev, Dmitry Gudovskih, Ivan Moloshnikov a Roman Rybka. “Gender Prediction for Authors of Russian Texts Using Regression And Classification Techniques.” In: *CDUD@ CLA*. 2016, p. 44–53.

[96] Huan Liu, Jun Li, Yaqin Fan a Zekun Song. “The Research of Web Text Classification Based on Wechat Article”. In: *Proceedings of the 6th International Conference on Information Engineering*. ACM. 2017, p. 2.

[97] Alejandro Llaves a Werner Kuhn. “An event abstraction layer for the integration of geosensor data”. In: *International Journal of Geographical Information Science* Vol. 28. No. 5 (2014), p. 1085–1106.

[98] Siyu Lu, Meng Zhao, Hui Zhang, Chen Zhang, Wei Wang a Hao Wang. “Genderpredictor: a method to predict gender of customers from e-commerce website”. In: *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on*. Zv. 3. IEEE. 2015, p. 13–16.

[99] Nizar R Mabroukeh a Christie I Ezeife. “A taxonomy of sequential pattern mining algorithms”. In: *ACM Computing Surveys (CSUR)* Vol. 43. No. 1 (2010), p. 3.

[100] Tariq Mahmood a Francesco Ricci. “Improving recommender systems with adaptive conversational strategies”. In: *Proceedings of the 20th ACM conference on Hypertext and hypermedia*. ACM. 2009, p. 73–82.

[101] Fragkiskos D Malliaros a Konstantinos Skianis. “Graph-based term weighting for text categorization”. In: *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*. IEEE. 2015, p. 1473–1479.

[102] Felix Mannhardt a Niek Tax. "Unsupervised event abstraction using pattern abstraction and local process models". In: *arXiv preprint arXiv:1704.03520* (2017).

[103] Abraham H Maslow. "A theory of human motivation." In: *Psychological review* Vol. 50. No. 4 (1943), p. 370.

[104] Raymond J Mooney a Loriene Roy. "Content-based book recommending using learning for text categorization". In: *Proceedings of the fifth ACM conference on Digital libraries*. ACM. 2000, p. 195–204.

[105] Dan Murray a Kevan Durrell. "Inferring demographic attributes of anonymous internet users". In: *International Workshop on Web Usage Analysis and User Profiling*. Springer. 1999, p. 7–20.

[106] Eric WT Ngai, Angappa Gunasekaran, Samuel Fosso Wamba, Shahriar Akter a Rameshwar Dubey. "Big data analytics in electronic markets". In: *Electronic Markets* Vol. 27. No. 3 (2017), p. 243–245.

[107] Kamal Nigam, Andrew McCallum, Sebastian Thrun, Tom Mitchell et al. "Learning to classify text from labeled and unlabeled documents". In: *AAAI/IAAI* Vol. 792 (1998).

[108] Lian-ju NING a Ying-ying ZHANG. "An Empirical Study on Group-buying Consumers' Preferences - Illustrated by Catering Group-buying [J]". In: *Journal of Northeastern University (Social Science)* Vol. 5 (2011), p. 007.

[109] Seyednaser Nourashrafeddin, Ehsan Sherkat, Rosane Minghim a Evangelos E Milios. "A Visual Approach for Interactive Keyterm-Based Clustering". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* Vol. 8. No. 1 (2018), p. 6.

[110] Maria Augusta SN Nunes a Rong Hu. "Personality-based recommender systems: an overview". In: *Proceedings of the sixth ACM conference on Recommender systems*. ACM. 2012, p. 5–6.

[111] Fábio AP Paiva, José AF Costa a Cláudio RM Silva. "A Personality-Based Recommender System for Semantic Searches in Vehicles Sales Portals". In: *International Conference on Hybrid Artificial Intelligence Systems*. Springer. 2017, p. 600–612.

[112] Dimitris Paraschakis, Bengt J Nilsson a John Holländer. "Comparative evaluation of top-n recommenders in e-commerce: An industrial perspective". In: *Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on*. IEEE. 2015, p. 1024–1031.

[113] Moon-Hee Park, Jin-Hyuk Hong a Sung-Bae Cho. "Location-based recommendation system using bayesian user's preference model in mobile devices". In: *International Conference on Ubiquitous Intelligence and Computing*. Springer. 2007, p. 1130–1139.

[114] Pratiksha Y Pawar a SH Gawande. "A comparative study on different types of approaches to text categorization". In: *International Journal of Machine Learning and Computing* Vol. 2. No. 4 (2012), p. 423.

[115] Michael J Pazzani. "A framework for collaborative, content-based and demographic filtering". In: *Artificial intelligence review* Vol. 13. No. 5-6 (1999), p. 393–408.

[116] Michael J Pazzani a Daniel Billsus. "Content-based recommendation systems". In: *The adaptive web*. Springer, 2007, p. 325–341.

[117] Marco Pennacchiotti a Ana-Maria Popescu. "A Machine Learning Approach to Twitter User Classification." In: *Icwsm* Vol. 11. No. 1 (2011), p. 281–288.

[118] Tu Minh Phuong et al. “Gender prediction using browsing history”. In: *Knowledge and Systems Engineering*. Springer, 2014, p. 271–283.

[119] Bruno Pradel, Savaneary Sean, Julien Delporte, Sébastien Guérif, Céline Rouveirol, Nicolas Usunier, Françoise Fogelman-Soulié a Frédéric Dufau-Joel. “A case study in a recommender system based on purchase data”. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2011, p. 377–385.

[120] Sahil Raj a Dilpreet Singh. “Impact of demographic factors on online purchase frequency—A decision tree approach”. In: *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on*. IEEE. 2016, p. 3789–3793.

[121] Aniket Rangrej, Sayali Kulkarni a Ashish V Tendulkar. “Comparative study of clustering techniques for short text documents”. In: *Proceedings of the 20th international conference companion on World wide web*. ACM. 2011, p. 111–112.

[122] Paul Resnick a Hal R Varian. “Recommender systems”. In: *Communications of the ACM* Vol. 40. No. 3 (1997), p. 56–58.

[123] Dahiya Richa. “Impact of demographic factors of consumers on online shopping behaviour: A study of consumers in India”. In: *International Journal of Engineering and Management Sciences* Vol. 3. No. 1 (2012), p. 43–52.

[124] Francesco Ricci, Lior Rokach a Bracha Shapira. “Introduction to recommender systems handbook”. In: *Recommender systems handbook*. Springer, 2011, p. 1–35.

[125] J Ben Schafer, Joseph A Konstan a John Riedl. “E-commerce recommendation applications”. In: *Data mining and knowledge discovery* Vol. 5. No. 1-2 (2001), p. 115–153.

[126] Hinrich Schütze a Craig Silverstein. “Projections for efficient document clustering”. In: *ACM SIGIR Forum*. Zv. 31. SI. ACM. 1997, p. 74–81.

[127] Farzad Sanati. “Multilevel life-event abstraction framework for e-government service integration”. In: *European Conference on e-Government*. Academic Conferences International (ACI). 2009.

[128] Badrul Sarwar, George Karypis, Joseph Konstan a John Riedl. “Analysis of recommendation algorithms for e-commerce”. In: *Proceedings of the 2nd ACM conference on Electronic commerce*. ACM. 2000, p. 158–167.

[129] Badrul Sarwar, George Karypis, Joseph Konstan a John Riedl. “Item-based collaborative filtering recommendation algorithms”. In: *Proceedings of the 10th international conference on World Wide Web*. ACM. 2001, p. 285–295.

[130] Abhaysingh Saste, Mangesh Bedekar a Pranali Kosamkar. “Predicting demographic attributes from web usage: Purpose and methodologies”. In: *I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2017 International Conference on*. IEEE. 2017, p. 381–386.

[131] Fabrizio Sebastiani. “Machine learning in automated text categorization”. In: *ACM computing surveys (CSUR)* Vol. 34. No. 1 (2002), p. 1–47.

[132] Saddys Segrera a Marıa N Moreno. “An experimental comparative study of web mining methods for recommender systems”. In: *Proceedings of the 6th WSEAS International Conference on Distance Learning and Web Engineering*. World Scientific, Engineering Academy a Society (WSEAS). 2006, p. 56–61.

[133] Sepideh Seifzadeh, Ahmed K Farahat, Mohamed S Kamel a Fakhri Karray. "Short-text clustering using statistical semantics". In: *Proceedings of the 24th International Conference on World Wide Web*. ACM. 2015, p. 805–810.

[134] Christophe Senot, Dimitre Kostadinov, Makram Bouzid, Jérôme Picault, Armen Aghasaryan a Cédric Bernier. "Analysis of strategies for building group profiles". In: *International Conference on User Modeling, Adaptation, and Personalization*. Springer. 2010, p. 40–51.

[135] Mugdha Sharma a Jasmeen Kaur. "A novel data mining approach for detecting spam emails using robust chi-square features". In: *Proceedings of the Third International Symposium on Women in Computing and Informatics*. ACM. 2015, p. 49–53.

[136] Lindsay H Shaw a Larry M Gant. "Users divided? Exploring the gender gap in Internet use". In: *CyberPsychology & Behavior* Vol. 5. No. 6 (2002), p. 517–527.

[137] Dan Shen, Jean-David Ruvini a Badrul Sarwar. "Large-scale item categorization for e-commerce". In: *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM. 2012, p. 595–604.

[138] Anca-Roxana Simon, Rémi Bois, Guillaume Gravier, Pascale Sébillot, Emmanuel Morin a Sien Moens. "Hierarchical topic models for language-based video hyperlinking". In: *Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia*. ACM. 2015, p. 31–34.

[139] Rashmi R Sinha, Kirsten Swearingen et al. "Comparing recommendations made by online systems and friends." In: *DELOS workshop: personalisation and recommender systems in digital libraries*. Zv. 106. 2001.

[140] Sara Speltdoorn a Dirk Van den Poel. "Predicting demographic characteristics of web users using semi-supervised classification techniques". Diz. pr. Master's dissertation, Ghent University, 2009-2010.

[141] R Srinivasan, M Senthilraja a S Iniyan. "Pattern recognition of Twitter users using semantic topic modelling". In: *IoT and Application (ICIOT), 2017 International Conference on*. IEEE. 2017, p. 1–4.

[142] Efstathios Stamatatos, Nikos Fakotakis a George Kokkinakis. "Automatic text categorization in terms of genre and author". In: *Computational linguistics* Vol. 26. No. 4 (2000), p. 471–495.

[143] Michael Steinbach, George Karypis, Vipin Kumar et al. "A comparison of document clustering techniques". In: *KDD workshop on text mining*. Zv. 400. 1. Boston. 2000, p. 525–526.

[144] Hidayet Takçı a Tunga Güngör. "A high performance centroid-based classification approach for language identification". In: *Pattern Recognition Letters* Vol. 33. No. 16 (2012), p. 2077–2084.

[145] Niek Tax, Natalia Sidorova, Reinder Haakma a Wil MP van der Aalst. "Event abstraction for process mining using supervised learning techniques". In: *Proceedings of SAI Intelligent Systems Conference*. Springer. 2016, p. 251–269.

[146] Nava Tintarev a Judith Masthoff. "Designing and evaluating explanations for recommender systems". In: *Recommender systems handbook*. Springer, 2011, p. 479–510.

[147] Shari Trewin. “Knowledge-based recommender systems”. In: *Encyclopedia of library and information science* Vol. 69. No. Supplement 32 (2000), p. 180.

[148] Chong Wang a David M Blei. “Collaborative topic modeling for recommending scientific articles”. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2011, p. 448–456.

[149] Jiyuan Wang, Jiayin Qi, Seongmin Jeon a Xiangling Fu. “Are you a local or a visitor?: an exploratory study on consumer behavior in online group buying commerce”. In: *Proceedings of the 18th Annual International Conference on Electronic Commerce: e-Commerce in Smart connected World*. ACM. 2016, p. 25.

[150] Licai Wang, Xiangwu Meng, Yujie Zhang a Yancui Shi. “New approaches to mood-based hybrid collaborative filtering”. In: *Proceedings of the workshop on context-aware movie recommendation*. ACM. 2010, p. 28–33.

[151] Zhiyi Wang, Liang Li, Chunjie Zhang a Qingming Huang. “Image-regulated graph topic model for cross-media topic detection”. In: *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*. ACM. 2015, p. 76.

[152] Ingmar Weber a Carlos Castillo. “The demographics of web search”. In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2010, p. 523–530.

[153] Ingmar Weber a Alejandro Jaimes. “Who uses web search for what: and how”. In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM. 2011, p. 15–24.

[154] Jason Weston, Samy Bengio a Nicolas Usunier. “Wsabie: Scaling up to large vocabulary image annotation”. In: *IJCAI*. Zv. 11. 2011, p. 2764–2770.

[155] Natalie Widmann a Suzan Verberne. “Graph-based semi-supervised learning for text classification”. In: *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. ACM. 2017, p. 59–66.

[156] Wen Wu a Li Chen. “Implicit acquisition of user personality for augmenting movie recommendations”. In: *International Conference on User Modeling, Adaptation, and Personalization*. Springer. 2015, p. 302–314.

[157] Liang Xiang, Quan Yuan, Shiwan Zhao, Li Chen, Xiatian Zhang, Qing Yang a Jimeng Sun. “Temporal recommendation on graphs via long-and short-term preference fusion”. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2010, p. 723–732.

[158] Pengtao Xie a Eric P Xing. “Integrating document clustering and topic modeling”. In: *arXiv preprint arXiv:1309.6874* (2013).

[159] Caiquan Xiong, Zhen Hua, Ke Lv a Xuan Li. “An Improved K-means text clustering algorithm By Optimizing initial cluster centers”. In: *Cloud Computing and Big Data (CCBD), 2016 7th International Conference on*. IEEE. 2016, p. 265–268.

[160] Guandong Xu, Yanchun Zhang a Xun Yi. “Modelling user behaviour for web recommendation using lda model”. In: *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT’08. IEEE/WIC/ACM International Conference on*. Zv. 3. IEEE. 2008, p. 529–532.

[161] Min Yang, Jincheng Mei, Fei Xu, Wenting Tu a Ziyu Lu. "Discovering author interest evolution in topic modeling". In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM. 2016, p. 801–804.

[162] Wenchuan Yang, Qiwei Wu a Zishuai Cheng. "Research on distributed text clustering based on frequent itemset". In: *Control Conference (CCC), 2017 36th Chinese*. IEEE. 2017, p. 5700–5705.

[163] Xiwang Yang, Yang Guo a Yong Liu. "Bayesian-inference-based recommendation in online social networks". In: *IEEE Transactions on Parallel and Distributed Systems* Vol. 24. No. 4 (2013), p. 642–651.

[164] Yi Yang, Shimei Pan, Jie Lu, Mercan Topkara a Yangqiu Song. "The stability and usability of statistical topic models". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* Vol. 6. No. 2 (2016), p. 14.

[165] Yiming Yang. "An evaluation of statistical approaches to text categorization". In: *Information retrieval* Vol. 1. No. 1-2 (1999), p. 69–90.

[166] Yiming Yang a Xin Liu. "A re-examination of text categorization methods". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1999, p. 42–49.

[167] Hsiang-Fu Yu, Cho-Jui Hsieh, Hyokun Yun, SVN Vishwanathan a Inderjit S Dhillon. "A scalable asynchronous distributed algorithm for topic modeling". In: *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2015, p. 1340–1350.

[168] Zhiwen Yu, Xingshe Zhou, Yanbin Hao a Jianhua Gu. "TV program recommendation for multiple viewers based on user profile merging". In: *User modeling and user-adapted interaction* Vol. 16. No. 1 (2006), p. 63–82.

[169] Lei Zhu, Guijun Wang a Xianchun Zou. "Improved information gain feature selection method for Chinese text classification based on word embedding". In: *Proceedings of the 6th International Conference on Software and Computer Applications*. ACM. 2017, p. 72–76.