BEHAVIOR-BASED CUSTOMER DEMOGRAPHY PREDICTION IN E-COMMERCE

Veronika Urbancokova
Slovak University of Technology in Bratislava
Ilkovicova 2, 842 16 Bratislava, Slovakia
Veronika.Urbancokova@stuba.sk

Michal Kompan
Slovak University of Technology in Bratislava
Ilkovicova 2, 842 16 Bratislava, Slovakia
Michal.Kompan@stuba.sk

Zuzana Trebulova
Slovak University of Technology in Bratislava
Ilkovicova 2, 842 16 Bratislava, Slovakia
Zuzana.Trebulova@stuba.sk

Maria Bielikova
Slovak University of Technology in Bratislava
Ilkovicova 2, 842 16 Bratislava, Slovakia
Maria.Bielikova@stuba.sk

ABSTRACT

The popularity of e-commerce is increasing day-by-day. In order to provide a seamless experience and tailored offer for the customers, the knowledge of their preferences and behavior is required. The demography of customers is one of the important information used for, e.g., segmentation. To provide optimal service, machine learning is often used for various tasks. However, the amount of data generated by customers and also large and changing product catalogs result in poorly performing models.

In this paper we aim at introducing behavior-based abstraction, which includes item and event abstractions also. Our method reduces the number of unique items in the e-commerce catalog and in the next step encodes the user behavior. We performed extensive evaluation over the real-world dataset. The results suggest the usefulness of proposed abstraction and also resulted in the improvement of the demography prediction performance both from model complexity and performance point of view.

Keywords: Demography prediction; Association rules mining; Topic modeling; Customer behavior

1. Introduction

The popularity of e-commerce is increasing day by day. It is expected to double the e-commerce sales by 2020[1]. In such a highly competitive environment, the knowledge of customer's needs, preferences, and demography seem to be a valuable advantage. Luckily, the activity of a customer within the e-commerce is usually recorded in the form of events or logs. These can be further used to understand and to explore the characteristics and preferences of the customers. As a result, we can improve the customer's experience by utilizing this information by generating personalized recommendations, provide specialized promotions, newsletters, or just to optimize e-commerce processes based on customer segmentation.

Demographic characteristics determine customer behavior and the process of decision making. Customer behavior, his/her taste, technology knowledge, preferred communication style, preferred advertisement varies within different demography groups (customers sharing similar demography). Various age of customers correlates with different technology knowledge (e.g., type of payment) and requires different customer service. E-commerce stores

[1] https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales

usually segment customers based on age, gender, location, or income [Oh et al. 2002] and personalize their services based on this customer segmentation.

The question of gender differences in online shopping behavior was researched in many marketing studies and research papers [Zhang et al. 2018]. Even though many differences identified in the past are not actual nowadays (e.g., men and women today share similar web knowledge, prefer similar payment methods), there are still some differences, i.e., factors in decision making, shopping frequency or dwell time. The significant differences also include the impact of price and discount – women are more often influenced by discounts than men [Pirlympou 2017]. Other factors that determine customer decision making include education level, marital status, nationality, customer religion, or net income. Customer demography is a major knowledge used for customer segmentation, advertisement campaigns, and e-commerce personalization (which is one of the major competitive advantages).

The main problem with the demography in e-commerce is its frequent absence. The customers are often not willing to provide the demography information to e-commerce. If they are required, there is also the tendency to submit fake information. Moreover, in the advent of the government protection of personal data (e.g., General Data Protection Regulation[2]), the amount of demographic information is rapidly decreasing. These are the reasons why a demography prediction task receives more attention as it can be predicted based on the users' implicit activity.

A lot of information useful for e-commerce is "hidden or encoded" in the customer footprints – actions he/she performs in e-commerce. These, however, are too noisy to be directly used. As a reason some kind of processing is needed to capture important information. Similarly, many machine learning methods used in e-commerce (e.g., segmentation, recommendation) require reducing and preprocessing of raw logs collected in e-commerce. The user behavior on the website is one of the most valuable information (often the only one), users provide. Moreover, the e-commerce domain is highly dynamic. The new products are added daily and replace the old ones. New products are added and categorized by the humans which often picks suboptimal category [Krishnan and Amarthaluri 2019]. Moreover, naturally, the category hierarchy changes over time (thanks to the products drift or simply by the sales decision). As a result, models trained over old data (often manually labeled) cannot be used. The usage of the item and behavior abstraction allows us to overcome time, user, and implementation-specific aspects while retaining the discriminative power of the data. In fact, in some extremely dynamic applications, i.e., Groupon[3]–like e-commerce, the usage of the abstraction is required.

To address these issues, we propose a behavior-based abstraction method, which introduces an abstraction on three levels – item, event, and customer behavior. We evaluate our method on the customer demography, i.e., the gender prediction task (as a supervised learning representative) on a real-world dataset. Similarly, unsupervised learning is explored through association rules mining. Proposed abstraction is however intended as an input for many other e-commerce applications, e.g., segmentation of customers or adaptation and personalized recommendation. The main contribution presented in this paper are:

- A novel customer behavior-based abstraction method consisting of item and events abstraction. The method is intended as an input for various machine learning tasks (i.e., recommendation [Ye et al. 2019]).
- Prediction of customer demography, i.e., gender – strictly on customer behavior data.
- A deep analysis of the abstraction impact on each step of the proposed method.

2. Related Work

Demography prediction can be defined as an instance of supervised learning. Usually, the prediction is based on a set of customer characteristics derived from its behavior. Customer features may cover explicit information such as frequency of purchase activity, preferred e-commerce category, and also latent features obtained from machine learning algorithms (e.g., embeddings obtained from neural networks).

One of the widely available customer information sources in e-commerce is customer behavior (his/her actions, i.e., clicks). Customer activity is stored in the form of events or logs [Kunz 1993]. Events are defined as any customer interaction recorded in an e-commerce store. They include implicit and explicit feedback generated by a customer, e.g., customer product ratings, or product views [Xu et al. 2019]. These are usually stored by e-commerce, regardless of their further usage. The events collected in their rough form are hardly ready for direct usage. As a rule, there is a necessity for high-level transformation. This high-level representation can be performed via an aggregation or behavior abstraction methods [Mannhardt and Tax 2017]. The main disadvantage of the aggregation of features is the absence of any time sequence of events (which are generated in strict order). This sequence information is most often ignored or simplified, e.g., to bigrams. As a reason usually some kind of behavior abstraction is needed.

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC

[3] https://www.groupon.com

The events abstraction is in the literature researched in two directions - item-based abstraction and behavior-based abstraction. The item-based abstraction is defined as a text mining problem aiming at discovering hidden patterns [Han et al. 2011], [Agrawal and Zhai 2012]. The usage of the abstraction in the level of items or customer behavior is necessary due to the multidimensionality of the data. The items are described by categories, tags, and natural language, which result in thousands of features for each item. Similarly, customer behavior is diverse from its nature, which produces various sizes of customer history, session lengths, etc. To aggregate this information and produce a machine-friendly data for machine learning the abstraction is usually used [Chovaňák et al. 2018]. On the contrary, the usage of the abstraction helps us overcome the data drift over time (e.g., various category hierarchy, new products in user history). Together, this helps us to train a more robust machine learning model which is robust for a longer time.

2.1. Item-based abstraction

Usually the abstraction on the level of items is realized via the clustering or some latent topic modeling approach. Clearly, these approaches (when appropriate hyperparameters are chosen) may introduce some generalization. The comparison of clustering algorithms for short-text documents was the main idea of [Rangrej et al. 2011]. Authors compared K-means, SVD, and graph-based approaches. The results showed that graph-based algorithms perform the best. The problem of short-text documents was also explored by Seifzadeh et al. [Seifzadeh et al. 2015], where the statistical semantic approach was proposed. Feature engineering was a major research question in [Aljaber et al. 2009], where citation-specific features in the context of academic text clustering were presented. This citation-based representation of documents in experiments outperforms the full-text clustering approach in two academic journal datasets. Xiong et al. in [Xiong et al. 2016] suggested an improvement of the K-means algorithm by optimization of the initialization method of cluster centers. This resulted in an improvement in the accuracy and stability of clustering. Bai and Jin in [Bai and Jin 2016] proposed a semantic graph-based structure for the text representation that has the potential to optimize similarity calculation. The results confirm the significant improvement of accuracy in the context of Chinese text clustering.

The topic modeling is often wrongly referred to as document clustering. Usually, the document clustering is based on so-called hard clustering (each document is associated with exactly one cluster). On the contrary, the topic modeling stands for the soft clustering (each document has probabilities across all clusters) [Agrawal and Zhai 2012]. The topic modeling is often used for dimensionality reduction. Wang and Blei in [Wang and Blei 2011] used statistical topic modeling as a part of collaborative topic modeling. The results of this work showed that the usage of statistical topic modeling can improve traditional recommender algorithms. The problem of user similarity computation, which is an essential part of the recommendation was discussed in [Srinivasan et al. 2017] where a method of user interest matrix creation through the topic modeling approach was suggested. The domain of micro-blocks in Twitter was examined by Hong and Davison in [Hong and Davison 2010] which focused on a problem of classification using Latent Dirichlet Allocation.

The question of LDA interpretability was discussed in [Hingmire et al. 2017]. The proposed approach consisted of two general components - finding the most relevant Wikipedia concepts for documents in a corpus and usage of Generalized Pólya Urn (GPU) - to include semantic relatedness into the process of LDA. Proposed method WikiLDA performed the best in the domains with hardly separable classes. The problem of author interest identification was a key issue of [Simon et al. 2015], where interest drift model was presented. The suggested model, thanks to the sensitivity to the ordering of words in texts, achieved better results than other state-of-the-art topic models. Chen et al. in [Chen et al. 2017] 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 a genetic algorithm was proposed.

Researchers in the last years also have been focusing on the combination of several approaches. The ensemble learning proved to be effective in many domains. Chen and Zhang in [Chen, W. and Zhang 2017] proposed a method LDA-KNN that improves the similarity computation. The proposed method includes also the semantic similarity into the neighbor search phase. The semantic similarity is computed using Latent Dirichlet Allocation. The weakness of this approach is its time efficiency. The combination of text classification and topic modeling algorithm LDA was also proposed in [Liu et al. 2017]. Their method combined LDA for the feature extraction and SVM for the text classification. The suggested method achieved better performance and reduced the training time of the classification. The integration of document clustering and topic modeling was the key idea of [Xie and Xing 2013]. The integration was designed as a 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 support the effectiveness of the proposed model.

2.2. Behavior-based abstraction

The main goal of the behavior-based abstraction is the search for a more general representation of events and customer behavior patterns with high discriminatory power. One of the first works in the field of the event abstraction was related to system debugging issues [Kunz 1993], [Kunz 1994]. The abstraction was constructed as a grouping of

low-level event sets into one high-level abstract events. Based on this, the two event set structures were proposed - complete precedence abstractions and contractions. The event abstraction was lately also used for including a time aspect as an extension of the programming language [Hooman and Roosmalen 1997], an abstraction of continuous systems [Giambiasi and Carmona 2006], sensor data integration [Llaves and Kuhn 2014], or process mining [Mannhardt and Tax 2017]. All these applications of event abstraction require to keep as much information as possible. In other words, the abstraction often reduces the discriminatory power of the information included, which results in worse machine learning models trained on such data. On the contrary, the amount of processed information is significantly lower.

The behavior abstraction is an essential part of machine learning approaches that work with the customer data represented as events. However, it is often presented as a part of feature engineering, rather than an individual research topic. Duong et al. [Duong et al. 2016] proposed session abstraction in the form of n-grams that was presented as part of feature engineering for the gender prediction task. The event abstraction is often based on the general event abstraction methods that are also applicable in other domains (e.g., the process mining or sensor data processing). However, the impact of the chosen events abstraction method on the discriminatory power in a machine learning task is not very discussed. One of the widely used behavior 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 [Tax et al. 2017] suggested a method of event abstraction based on a generation of feature vector representations using XES extensions. George et al. [George et al. 2016] presented the IL-MINER method that can discover event patterns without a priori knowledge of event abstractions.

3. The Pattern-based Customer Behavior Abstraction

The behavior abstraction is an open research problem in the e-commerce domain. Some kind of abstraction is required as a preprocessing for the machine learning algorithms. There are several levels on which the abstraction may be introduced, e.g., on the level of items (as these are often too specific train generalized model) or on the level of events or customer behavior.

In order to obtain an abstraction of customers' behavior, we propose the method for item and events abstraction which produces the behavior abstraction. Our method is designed for domains with the textual representation of items. The major step of our method is transforming customer behavior and items characteristics to the latent features, which are more appropriate for machine learning tasks (in comparison to raw events). Such features are useful for user segmentation, user modeling, personalization, or various prediction tasks (e.g., demography, personal traits, or interests). Our method consists of two major parts – items abstraction and events abstraction (Figure 1).

3.1. Item-based abstraction based on Latent Dirichlet Allocation

The Web content is usually organized into hierarchical categories. Two approaches are used for assigning a specific category to an item. The first approach is based on a domain expert assessment, however, in large catalogs this is quite an expensive way. Such an assignment is based on the subjective knowledge of the expert and his/her perception of the item description. The second approach - latent categorization - tries to construct categories automatically (based on the items description). The absence of experts, results in cheaper and faster category assignment. However, the accuracy may vary based on the quality of the items' description (in comparison with the expert assignment). Nevertheless, when used for the machine learning tasks (e.g., classification, pattern mining), categories obtained automatically can achieve comparable results with categories created by a domain expert. Given that our method of events abstraction should be generally usable, we opt for the item abstraction in the form of the topic modeling.

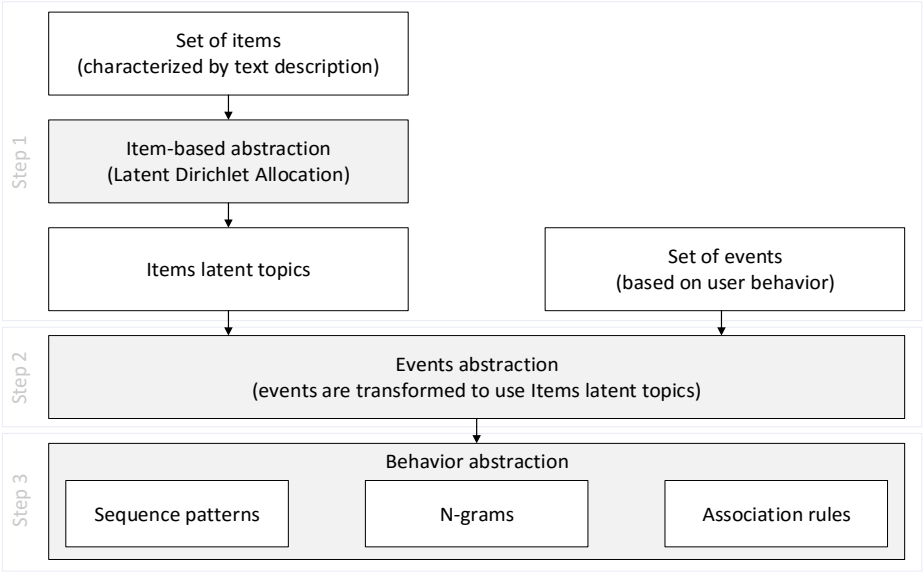

Figure 1: Behavior-based abstraction method. The input of our method consists of a set of items and events. The first component - Item-based abstraction - generates latent categories for all items (see Section 3.1). These latent categories are used as input into the Events abstraction component (see Section 3.2). The third component – Behavior abstraction processes abstract events and represents customers' behavior as sequence patterns, association rules, and n-grams (see Section 3.3).

Method of item abstraction based on Latent Dirichlet Allocation (Figure 1 – step 1) consists of 3 steps:

1. Pre-processing of items
2. Topic modeling
3. Representative topics selection

The input of LDA is a set of items, where each item corresponding to the one "text document". This text document is represented as a vector of words, i.e., tokens. Each text description of an item has to be preprocessed. Pre-processing of items includes standard text operations:

1. HTML artefacts removal
   a. removal of special characters
   b. removal of HTML elements
2. Tokenization - splitting the text into tokens, where a token represents a single unit separated by a white character
3. Lowercase transformation
4. Lemmatization - transforming tokens to the basic form
5. Stop-words removal (frequently used words, that are not important in the context of the corpus)
   a. domain-specific stop-words (e.g., in e-commerce domain: price, discount)
   b. language-specific stop-words

Topic modeling directly generates the names of clusters in the form of representative topics. It also does not require a labeled dataset. As we discussed in the related work section, the topic modeling in comparison to the document clustering 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 modeling algorithm [Bíró et al. 2009]. LDA is a three-level Bayesian model that creates a latent topic based on a document collection. Each topic is represented as a set of basic tokens probabilities [Blei et al. 2003]. For the purposes of our method we use only the first topic (the one with the highest probability). This topic represents a latent category of an item.

The last step of the item-based abstraction method is the selection of representative topics, i.e., the topic with the highest probability computed in the topic modeling step. In other words, we use the latent topic as the abstraction of the item ID.

3.2. Events abstraction

The second component of our method is the events abstraction (Figure 1 – step 2). An event characterizes a connection between a customer (who performs an action) and a specific item. Clearly, such specific information cannot be directly used in the machine learning algorithm (the event is stored on the level of the customer ID and the item

ID). That is the reason why, in this step, we enhance the events with the LDA representative topics, generated as described in Section 3.1. We propose to consider four basic types of events:

- Events associated with exactly one specific item (e.g., view of item or rating of the item) - abstraction using representative LDA topic[4]
- Events associated with multiple items (e.g., view of list of items) - abstraction using most frequent LDA representative topic
- Purchase events (one purchase usually consists of multiple items) – for each item in the purchase a new event is generated, while the abstraction uses representative LDA topic
- Events without association with specific items (e.g., view of a basket)

Events abstraction, based on item representative LDA topic, represents an effective way to achieve higher-level granularity. However, the number of such events may be still too high. That's the reason for important events selection – in the mean of important patterns. Moreover, we aim to capture customer behavior, which is still (in this step) encoded in the sequence of events.

3.3. Behavior abstraction

The last component of our method focuses on customer behavior representation (Figure 1 – step 3). This step is necessary for the selection of important behavior patterns – which can be further used in various machine learning tasks. For this task we use pattern recognition which results in several representations of customers' behavior. Hand by hand, this abstraction reduces the number of events (as only events present in important behavior patterns are considered). We propose to describe the customer behavior patterns (based on the abstract events from previous steps) by three standard approaches:

- Sequence patterns (FreeSpan algorithm [Han et al. 2000])
- N-grams
- Association rules (Apriori algorithm [Agrawal and Srikant 1994])

The actual choice of which approach should be used depends on the specific machine learning task and its performance. The choice can be performed in several ways, e.g., information gain metric, precision, or error rate. Patterns can be directly used for marketing purposes, e.g., the segmentation.

4. Evaluation

Due to the complexity of the proposed abstraction method, we evaluated our approach in two steps. In the first step, we focused on item and events abstraction based on Latent Dirichlet Allocation. Latent categories obtained from events abstraction were compared (indirectly by the pattern recognition task) with explicit categories assigned by a domain expert. These two approaches were compared in terms of pattern recognition metrics - support and confidence. In the second step, we focused on a behavior abstraction. We evaluated our three parallel pattern mining approaches (sequence patterns, n-grams, and association rules) using the customer gender prediction task.

4.1. Dataset and pre-processing

The experiments were performed over the real-traffic sample of the e-commerce portal[5]. The portal offers various discounted deals (similar to well-known Groupon). The portal is typical with short-term deals and diverse offers. Deals are generally offered during a couple of weeks or months (permanent deals in this domain do not exist). The average selling (i.e., availability) period of an offer is 2.5 weeks. Deals are organized into six basic categories - food, travel, services, goods, health, and sport. The annual sales turnover is approx. 20M €. Every day, there are some new offers (i.e., deals). The customer experience follows standard e-commerce best practices, from the visual, navigational, and logical point of view also (Figure 2). The customers' interaction with the website is stored by the server-side log in the form of several event types. The item metadata is obtained from the product catalog. Finally, customer demography is derived from the customers' profiles. For the experiments we obtain a snapshot of the website data, while we used the following events and information:

- Transaction data (approx. 3 million of purchases, 500 000 customers)
- Users activity on the website (over 18.5 million of events) (Table 1)
- Item (i.e., deals) metadata (e.g., title, description, category, location, price, discount) (Table 2)
- Customer demography information (e.g., gender, income, location) (Table 3)

[4] The representative LDA topic is considered to be the topic with highest probability.
[5] https://www.zlavadna.sk

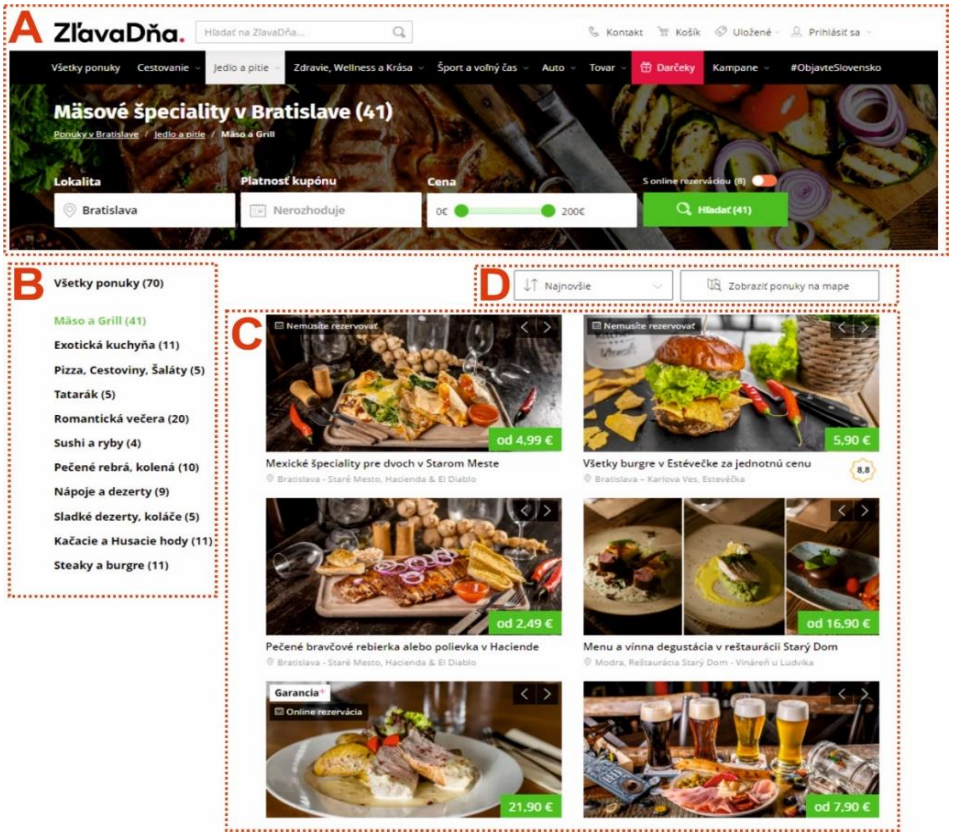

Figure 2: Example of the user interface on www.zlavadna.sk. - the top of the page (A) is reserved for the basis navigation – categories, search, account administration, chart, etc. The in-category navigation is performed by the facets (B). The main content is presented in the grid (C). The main content can be sorted according to field standards (newest, order by price, etc.) (D).

Table 1: Example of the events logged based on the user activity on the website. For every action a corresponding event is generated and stored – including event type, timestamp, user and item identifiers, etc.

user_id	event_type	event_time	query	item_id	item_type
5116495sl3d65420f083	view	12:14:11		393495	deal
3716425915bws420f09	search	12:14:16	dinner		
871d493245b65420f36	view	12:14:19		293537	deal
1217497423x25420f0d	basket	12:14:19		391545	deal

Table 2: Example of the item metadata that is stored for each item from the product catalog within the store. Each item has assigned a begin and end time when the item is available on the website.

item_id	315243
item_title	Handmade burger with fries
begin_time	00:01:00 23. February
end_time	23:59:59 20. March
item_text	Beef burger with the most popular ingredients - exactly what you can try at charming Old Town restaurant. It is served a little unconventionally, with sweet fries. The combination of these tastes will create an explosion in your mouth. Quench it with beer or the popular lemonade included.
item_city	Bratislava
main_category	Food
tags	hamburger, fries, food, dining, beef
price	9.9€

Table 3: Example of the user profile stored for the customer. Only relevant attributes are shown (for some users, some attributes may be missing).

user_name	e-mail	phone	default_city	gender	age	address	favorite
john57	john@john.qq	5621384	Springfield	male	28	Evergreen Terrace, 754	food, health, and sport

In the pre-processing phase we tried to reduce the bias and noise naturally present in the raw web-based logs. In total the four major steps were performed:

1. Elimination of incomplete events. We truncated all events without a customer identifier. Similarly, all events without connection to the domain item were truncated. As some of the items were highly unpopular, we also removed all items with less than 5 view events. As a result, approx. 17% of events were discarded.
2. Attribution of anonymous sessions. In this step, the customer metadata was assigned to every event and session (dealing with signed out customers, various customer IDs across devices etc.).
3. Outlier detection. By analyzing our dataset, we found out that the distribution of user activity is skewed. For this reason, we decided to remove events from the most active and least active users. As a result, we removed customers with less than 5 events (2.2 million cookies) and customers with more than 1000 events (240 cookies). This resulted in 18.5 million of events generated by 500 000 customers (the original dataset includes 24 million of events generated by 2.8 million of cookies).
4. Session construction. As the dataset naturally does not include the sessions, raw events have to be grouped into sessions. We used a standard approach for the session construction based on a time interval when a customer is inactive. In other words, two events are assigned to one session if the time difference between two consecutive actions is less than 25 minutes [Catledge and Pitkow 1995]. As a result, we obtained 3.3 million of sessions with an average length of 4.93 events.

4.2. Methodology

Metrics

For the comparison we used several widely used metrics. These helped us to explore various characteristics of the proposed method and each of the pattern mining approaches. The *Support* expresses the probability of the occurrence frequency of a pattern set within the whole dataset (Eq. 1). *Confidence* is an indicator of the rule truthfulness (Eq. 2). For the prediction task performance evaluation, we used the comparative metrics: *Precision* (Eq. 3), *Recall* (Eq. 4), and *F1-score* (Eq. 5).

$$\text{Support}(\{X\} \rightarrow \{Y\}) = \frac{|X \cap Y|}{|X \cup Y|} \tag{1}$$

$$\text{Confidence}(\{\text{X}\} \rightarrow \{\text{Y}\}) = \frac{|\text{X} \cap \text{Y}|}{|\text{X}|} \tag{2}$$

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

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

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

Hyperparameter tuning

One of the important steps for almost each machine learning algorithm is to define optimal parameters. As our approach utilizes several algorithms, we performed exhaustive hyperparameter tuning for each of them separately. The description of the process and best parameters further used in the experiments, see Appendix A.

4.3. RQ1: How will perform the item abstraction (without behavior-based abstraction) in comparison to standard expert-based abstraction within machine learning tasks?

As the major step of our proposed approach is the item abstraction, we aim at exploring the effectiveness of this idea in comparison to standard expert-defined categories. To explore the properties, we opt for three machine learning tasks. Two are focused on customer behavior – association rules mining, sequence pattern mining, and one on the customer gender prediction.

Association rules mining

We focus on the number of found association rules (the number of rules should be enough to cover the information contained in the dataset). In the first step, we, therefore, compared latent categories with expert categories via the number of generated rules - we assume that a bigger number of rules is better (at the same support and confidence levels) (Table 4). As we can see that latent categories are more appropriate for a smaller number of generated rules, otherwise expert categories generate a larger number of fewer rules.

The second step is the comparison of latent categories with expert categories using the top-N relevant rules. For the evaluation, the relevant pattern has to contain at least two different categories of items. The results of the top 10 relevant patterns obtained by association rules mining are shown in Table 5. As we can see, latent categories created with our method (item abstraction based on Latent Dirichlet Allocation) result in more valuable rules. As we assumed, expert categories are more detailed and therefore achieved lower values of support and confidence as rules generated via latent categories. To sum it up, latent categories generate more efficient rules with higher support and confidence. On the other hand, it generates fewer patterns that can be important for machine learning tasks (depends on a specific use-case).

Table 4: The evaluation of association rules mining step with different parameter settings – minimal support and confidence. Both latent (output of the 2nd step of our proposed approach) and expert (assigned by a domain expert) categories are compared.

Min support	Min confidence	Expert			Latent		
		Found rules	Avg support	Avg confidence	Found rules	Avg support	Avg confidence
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
0.20	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

Sequence pattern mining

Similarly to association rules mining, we compared the sequence patterns based on the expert and latent categories. As we can see in Table 6, sequence pattern mining based on latent categories, in general, generate more patterns (on the same level of minimal support). In other words, the abstraction resulted in more supported general rules. In the second step, we also compared expert categories and latent categories via the top 10 relevant patterns (Table 7). As expected, latent categories obtained by our item abstraction generate sequence patterns with higher support - these patterns are "stronger" as patterns obtained via expert categories. Moreover, these patterns are more robust, as the latent representation suppresses the details of product representation and hierarchy. On the contrary, expert categories generate more detailed patterns.

Demography prediction based on item and events abstraction

Among the unsupervised machine learning task, we explored the proposed approach also by a supervised classification task – a gender prediction (binary classification). From the available data we selected a subset containing users' demography which resulted in approx. 42k samples (15k women and 27k men). Next, we balanced the sample (by reducing the major class) to obtain similarly sized classes. The dataset was split to the train and test set (80:20 ratio).

The comparison was based on two features sets - the first feature set contains features generated by our behavior abstraction method. The second feature set consists of patterns created using expert categories. The number of features generated by our method was 5 410 while the expert category approach consists of 5 143. Because of many features,

we have applied a feature selection based on the information gain metric. The limit parameter of feature selection was the importance of a feature larger than 0.001. The feature selection reduced the original feature set to: behavior abstraction set - 124 features and expert category set 216 features.

The most important features generated via latent categories are simpler and achieved higher importance (in comparison to the most important feature generated via expert categories). Clearly, the discriminative power of the information in data is spread to more features when using the expert categories. From the quantitative point of view both approaches – expert and latent were comparable (Table 8). That is a promising result as the proposed abstraction preserved valuable information. Moreover, we used almost 50% fewer features as were required in the case of the expert version.

Table 5: The comparison of support and confidence of top 10 relevant association rules based on latent or expert categories ($support_{min} = 0.005, confidence_{min} = 0.8$).

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

Table 6: The comparison of the sequence pattern mining step with different parameter settings – minimal support and maximal pattern length.

Min support	Max pattern length	Expert		Latent	
		Found rules	Avg support	Found rules	Avg 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

To sum it up, pattern mining (represented by the association rules and the sequence pattern mining) generates stronger patterns using latent categories (obtained by our item and events abstraction method) rather than expert categories. Expert categories are more detailed and generated patterns with lower support and confidence. The comparison via the gender prediction task shown the similar discriminatory power of latent categories and expert categories. We showed the minimum drop of encoded information in comparison to expert categories. The latent categories, i.e., proposed abstraction is an alternative to manual expert categorization that is more expensive. Moreover, when applied to the machine learning task, the proposed abstraction produces more efficient and robust models (as fewer attributes are needed).

Table 7: A comparison of support of top 10 relevant sequence patterns based on latent or expert categories ($support_{min} = 0.015, length_{max} = 5$).

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

Table 8: Results of the binary classifier (Random Forest) for the gender prediction task based on the item and events abstraction (no behavior abstraction was used).

	Precision	Recall	F1
Latent	0.71	0.67	0.68
Expert	0.71	0.66	0.68

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

In previous sections we proved that the item and events abstraction (Figure 1 – Step 1 and Step 2) are beneficial for both supervised and unsupervised machine learning tasks. Analogically, to the idea of the abstraction of specific item categories, the behavior abstraction step (Figure 1 – Step 3) aims at reducing the amount of data stored for each user. Association rules mining, sequence pattern mining, and n-grams mining represent basic approaches for the behavior abstraction.

We believe that different behavior abstraction approaches will perform differently over various machine learning tasks. The N-grams and sequence patterns are generally very similar. If the complete set 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 complete subset of chosen sequence patterns. The comparison of three behavior abstraction approaches was realized in the same way as in *RQ1* by a binary classification (for detailed parameters see Appendix A).

To explore each combination (association rules, sequence pattern, N-grams) we generated seven datasets consisting of features generated by each approach (or a combination). Next, the feature selection based on the information gain metric was performed. The experimental results have shown that sequence pattern and N-grams achieved comparable results in the demography prediction task - the recall for both approaches was 67% (Table 9). Although the association rules mining generates more complex patterns, the obtained recall was 8% lower than the sequence pattern or N-grams. The best results were obtained by the combination of sequence patterns and N-grams as the behavior abstraction approach (*recall = 70%*). When comparing results of the gender prediction based on the behavior abstraction (Table 9) and the item and events abstraction (Table 8), we can observe an improvement in the case of behavior abstraction (both metrics – precision and recall). In fact, this is a positive result, which indicates the usefulness of the behavior abstraction.

As expected, the sequence pattern features and N-grams features are similar – 6/10 discovered patterns are identical. The main difference is in the importance of the top features. N-gram features, in general, achieved higher values of importance. The association rules generated more complex patterns with lower importance. The combination of pattern mining approaches often generated 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.

Table 9: Results of the binary classifier (Random Forest) for the demography prediction based on the behavior abstraction (AR - association rules, SP - sequence patterns, NG - N-grams).

Dataset	Generated features	Selected features	Precision	Recall	F1
AR	3120	244	0.63	0.59	0.60
SP	3551	167	0.71	0.67	0.68
NG	2371	140	0.71	0.67	0.68
AR + SP	5410	127	0.71	0.68	**0.69**
SP + NG	5920	125	**0.72**	**0.70**	**0.69**
AR + NG	4230	97	0.70	0.68	**0.69**
AR + NG + SP	7779	164	0.70	0.66	0.67

5. Conclusions

The customer behavior on the Web is usually modeled based on the customer logs. These need to be processed in order to apply various machine learning algorithms. Thanks to the customer increasing activity on e-commerce, both academia and business paid more attention to such information. Our proposed event and in the next step behavior abstraction method offers a straightforward approach to processing such data. The proposed method can be applied to any standard user-generated logs from e-commerce, while the items need to have a textual description. Our method consists of three parts – item, event, and the behavior abstraction.

In this paper, we explored the properties of our method on the real-world e-commerce dataset. We showed that the proposed item and events abstraction resulted in stronger top patterns than expert categories. On the contrary, expert categories are more detailed and generated patterns with lower support and confidence. We also applied items and events abstraction to the demography prediction task (focusing on gender prediction). The results suggest that the discriminatory power of item abstraction method and expert categories are similar. Moreover, to obtain comparable classification performance fewer features were required when using the event abstraction. We showed that the behavior abstraction retains important information which can be utilized for gender prediction. We believe that other behavior-dependent demography characteristics would also benefit from proposed abstraction (i.e., the income, age).

Next, we explored the behavior abstraction based on the 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 the demography prediction task. However, the combination of these approaches increased the recall, and thus improved the performance of the prediction task. Moreover, the complexity of the prediction model is lower, when using our approach, which helps the portability and training costs.

The usage of some kind of abstraction is required for almost any machine learning algorithm. The e-commerce domain is highly dynamic from all perspectives – new products are added daily, the customer base is evolving. Even, extremely popular neural networks need to include some kind of embedding to produce meaningful results. Our experiments showed that, the item abstraction helps us for unsupervised machine learning tasks – pattern and association rules mining. More robust patterns have been produced. On the contrary, on the supervised machine learning task, comparable results were obtained with fewer features used.

The behavior abstraction is required in order to aggregate users' actions made within a website. We found out that the usage of the behavior abstraction improves the gender prediction task. One of the important aspects of the proposed approach is the robustness of the learned model. Thanks to the abstraction, some domain drift (new items added, website structure changes etc.) can be handled without the need for model re-train. This is an important benefit in a highly dynamic environment as e-commerce.

One of the important aspects, when user generated data are processed, is privacy. The proposed abstraction method helps to reduce the amount of information required to store about a user. Moreover, it increases the robustness of trained models and helps to reduce some bias present in the data (e.g., miss information provided by users). The item-based abstraction helps to hide specific user's purchase history (as it abstracts from the product ID to latent topic ID). The event and behavior abstraction further abstracts from specific user's actions and results to store only general behavior patterns shared across clusters of users. Thanks to the efforts of introducing the rules of user data usage (e.g., GDPR), users can explicitly allow or reject the usage of the data, which reduces the problems on both sides. Without this, any of the modern Web-based service or e-commerce cannot provide precise, attractive, and enjoyable content for its customers.

The application of any machine learning task should result in the improvement of the end-user. In case of the e-commerce these are the customers of the store. Our proposed abstraction approach can improve many of these tasks and thus indirectly improves the customer experience in many ways. Improved customer segmentation helps to understand customer segments and their needs. Hierarchy abstraction helps to process new products and discovers

hidden connections across various categories. The usage of the proposed approach improves the association rules discovery (or market basket analysis), which are used to design more attractive campaigns. The prediction of customer behavior is intended to reduce user effort in addressing his/her information needs. Moreover, for the business it can be used in the strategic decision, e.g., in the customer churn prediction task [Kompan et al. 2019].

To sum it up, proposed behavior-based abstraction is beneficial for many applications of machine learning task in e-commerce namely:

- Customer behavior analysis and predictions. The output of our abstraction method is the sequence patterns and association rules, which are, in fact, extremely valuable for any business. These are used to explore and to understand the behavior, trends, and preferences of customers. Alternatively, they are often used in the prediction of customers' future behavior, which, again, helps in optimizing e-commerce processes (stock levels, prices, campaigns etc.) [Berger and Kompan 2019].
- Customer demography predictions. As we show in our experiments, the usage of proposed abstraction improves the demography prediction (e.g., gender prediction) in the mean of the precision and complexity. The demography is often used in customer segmentation.
- New product and customer – model robustness. A less activity generated by a new customer is required, as the abstract representation is used instead of raw data. This holds for new products too – as hidden connections across the product catalog are explored. As a result, the new customer receives an optimized user experience immediately (including newly added products).
- Privacy. Thanks to the three-level abstraction customer's actions are aggregated and transformed into the patterns which abstract from specific product IDs. This also reduces the amount of information stored about each customer.

There are also some limitations of the proposed approach and the evaluation. It would be interesting to explore the effect of the behavior abstraction on other demography prediction tasks. Moreover, other tasks, i.e., churn rate or personalized recommendations may benefit from the proposed abstractions. As we pointed, model robustness is one of the benefits of the proposed approach. Detailed experiments over various time periods need to be performed in order to explore this aspect. Another interesting direction is the research on data amount requirements. For instance, how many users providing fake information (or withdrawing their data) will affect any of the application tasks (e.g., gender prediction). This can be used in interpreting the results in domains where users are not willing to store their actions (due to privacy issues). Such experiments can simulate the performance of the model reflecting various user privacy-related types. We believe, that proposed abstraction is also valuable for business or marketing. The user segmentation is a field standard nowadays. As we showed the usage of the abstraction was beneficial for the unsupervised machine learning tasks (which is represented by the user segmentation) and thus we expect similarly positive results.

Acknowledgment

This work was partially supported by the Slovak Research and Development Agency under contract No. APVV-15-0508, the Scientific Grant Agency of the Slovak Republic, grants No. VG 1/0667/18 and VG 1/0725/19 and is the partial result of the Research & Development Operational Programme for the project Research of methods for acquisition, analysis and personalized conveying of information and knowledge, ITMS 26240220039, co-funded by the ERDF.

REFERENCES

Agrawal, R. and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," *In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487-499, 1994.

Agrawal, C., and Ch. Zhai, Mining Text Data, Springer-Verlag New York, 2012.

Aljaber, B., N. Stokes, J. R. Bailey, and J. Pei, "Document clustering of scientific texts using citation contexts," *Information Retrieval*, vol 13, 101-131, 2009.

Bai, Q., and C. Jin, "Text Clustering Algorithm Based on Semantic Graph Structure," *In proceedings of the 9th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, 312-316, 2016.

Berger, P., and M. Kompan, "User Modelling for Churn Prediction in E-commerce", *In Intelligent Systems Journal*, IEEE, Vol. 34, 2:44-52, 2019.

Bíró, I., J. Szabó, W. Buntine, M. Grobelnik, D. Mladenić, and J. Shawe-Taylor, "Latent Dirichlet Allocation for Automatic Document Categorization," *In Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer Berlin Heidelberg, 430-441, 2009.

Blei, D.M., A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, Vol. 3, 993-1022, 2003.

Chen, L., J. Li, and L. Zhang, "A method of text categorization based on genetic algorithm and LDA," *In Proceedings of the 36th Chinese Control Conference (CCC)*, Dalian, 10866-10870, 2017.

Chen, W., and X. Zhang, "Research on text categorization model based on LDA-KNN," *In Proceedings of the IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Chongqing, 2719-2726, 2017.

Chovanak, T., O. Kassak, M. Kompan, and M. Bielikova, "Fast Streaming Behavioural Pattern Mining," *In New Generation Computing*, Springer, Vol.17, No. 4:365-391, ISSN 0288-3635, 2018.

Duong, D., H. Tan, and S. Pham, "Customer gender prediction based on E-commerce data," *In Proceedings of the 8th International Conference on Knowledge and Systems Engineering (KSE)*, Hanoi, 91-95, 2016.

George, L., B. Cadonna, and M. Weidlich, "IL-miner: instance-level discovery of complex event patterns," *In Proceedings of the VLDB Endow*, Vol. 10, 1:25-36, 2016.

Giambiasi, N., and J. C. Carmona, "Generalized discrete event abstraction of continuous systems: GDEVS formalism," *Simulation Modelling Practice and Theory*, Elsevier, Vol. 14, 1:47-70, 2006.

Han, J., M. Kamber and J. Pei, Data Mining: Concepts and Techniques, 3rd Edition, Morgan Kaufman, 2011.

Han, J., J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M-Ch. Hsu, "FreeSpan: frequent pattern-projected sequential pattern mining," *In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '00)*, ACM, New York, 355-359, 2000.

Hingmire, S., S. Chakraborti, G. Palshikar, and A. Sodani, "WikiLDA: Towards More Effective Knowledge Acquisition in Topic Models using Wikipedia," *In Proceedings of the Knowledge Capture Conference (K-CAP 2017)*, ACM, New York, NY, USA, Article 37, 4 pages, 2017.

Hooman, J., and O. van Roosmalen, "Timed-event abstraction and timing constraints in distributed real-time programming," *In Proceedings of the Third International Workshop on Object-Oriented Real-Time Dependable Systems*, Newport Beach, CA, USA, 153-160, 1997.

Hong, L., and B. D. Davison, "Empirical study of topic modeling in Twitter," *In Proceedings of the First Workshop on Social Media Analytics (SOMA '10)*, ACM, New York, NY, USA, 80-88, 2010.

Kompan, M, O. Kassak, and M. Bielikova, "The Short-term User Modeling for Predictive Applications" *In Journal on Data Semantics*, Springer, Vol.8, 1: 21-37, 2019.

Krishnan, A., and A. Amarthaluri, "Large Scale Product Categorization using Structured and Unstructured Attributes" *In arXiv*, 1903.04254, 9 pages, 2019.

Kunz, T., "Event Abstraction: Some Definitions and Theorems", Technische Universität Darmstadt, 1-37, 1993.

Kunz, T., "Reverse engineering distributed applications: An event abstraction tool," *International Journal of Software Engineering and Knowledge Engineering*, World Scientific Publishing Co., Vol. 04, 03:303-323, 1994.

Liu, H., J. Li, Y. Fan, and Z. Song, "The Research of Web Text Classification Based on Wechat Article," *In Proceedings of the 6th International Conference on Information Engineering (ICIE '17)*, ACM, New York, NY, USA, Article 2, 5 pages, 2017

Llaves, A., and W. Kuhn, "An event abstraction layer for the integration of geosensor data," *International Journal of Geographical Information Science*, Vol. 28, 5:1085-1106, 2014.

Mannhardt, F. and N. Tax, "Unsupervised Event Abstraction using Pattern Abstraction and Local Process Models", *In Proceedings of Enabling Business Transformation by Business Process Modeling, Development, and Support Working Conference*, 1-9, 2017.

Oh, H., S.C. Parks, and F.J. Demicco, "Age- and Gender-Based Market Segmentation," *International Journal of Hospitality and Tourism Administration*, Vol. 3, No. 1:1-20, 2002.

Pirlympou, Z., "A Critical Study: How Gender Determines Consumer Preferences," *East-West Journal of Economics and Business*, No. 2:29-37, 2017.

Rangrej, A., S. Kulkarni, and A. V. Tendulkar, "Comparative study of clustering techniques for short text documents," *In Proceedings of the 20th international conference companion on World wide web*, ACM, 111-112, 2011.

Seifzadeh, S., A. K. Farahat, M. S. Kamel, and F. Karray, "Short-Text Clustering using Statistical Semantics," *In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion)*, ACM, New York, NY, USA, 805-810, 2015.

Simon, A-R., R. Bois, G. Gravier, P. Sébillot, E. Morin, and S. Moens, "Hierarchical Topic Models for Language-based Video Hyperlinking," *In Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia (SLAM '15)*, ACM, New York, NY, USA, 31-34, 2015.

Srinivasan, R., M. Senthilraja, and S. Iniyan, "Pattern recognition of Twitter users using semantic topic modelling," *In Proceedings of the International Conference on IoT and Application (ICIOT)*, Nagapattinam, 1-4, 2017.

Tax, N., N. Sidorova, R. Haakma, W. M. P. van der Aalst, Y. Bi, S. Kapoor, and R. Bhatia, "Event Abstraction for Process Mining Using Supervised Learning Techniques," *In Proceedings of SAI Intelligent Systems Conference (IntelliSys)*, Springer International Publishing, Cham, 251-269, 2017.

Wang, Ch., and D. 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, 448-456, 2011.

Xie P., and E. P. Xing, "Integrating document clustering and topic modeling," *In Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI'13)*, Ann Nicholson and Padhraic Smyth (Eds.), AUAI Press, Arlington, Virginia, United States, 694-703, 2013.

Xiong, C., Z. Hua, K. Lv, and X. Li, "An Improved K-means Text Clustering Algorithm by Optimizing Initial Cluster Centers," *In Proceedings of the 7th International Conference on Cloud Computing and Big Data (CCBD)*, Macau, 265-268, 2016.

Xu, D., H. Hong, Q. Ye, and D. Xu, "Regional Economic Status and Online Rating Behavior," Journal of Electronic Commerce Research, Vol 20., 3:184-198, 2019.

Ye, B. K, Y. J. Tu, and T.P. Liang, "A Hybrid System for Personalized Content Recommendation," *Journal of Electronic Commerce Research*, Vol 20., 2:91-104, 2019.

Zhang, Z., H. Li, F. Meng, and S Qiao, "Gender Difference in Restaurant Online Booking Timing and the Moderating Effect of Sell-out Risk and Information Type," *Journal of Electronic Commerce Research*, Vol 19., 3:266-279, 2018.

Appendix A – Hyperparameter tuning

As the proposed approach uses several algorithms, we performed hyperparameter tuning to guarantee the optimal results of each computation step.

Latent Dirichlet Allocation hyperparameters

The Latent Dirichlet Allocation (LDA) is the basis for the item-based abstraction. Based on the input documents (i.e., deals) it generates a set of topics with corresponding distributions. We use the topic with the highest probability as the representative topics represent the latent category that we use as the item-based abstraction.

The most important parameter of LDA is the number of topics to be discovered. The number of topics was selected by iterative testing of the topics' number from 50 to 300 with a step of 20. The evaluation was performed based on the following constraints:

- The cluster generated by the LDA should contain as few real metadata categories as possible. In other words, we want deals with the same LDA topics to be ideally from the same categories (assigned by the domain expert).
- In contrary to previous requirements, we also want the cluster of the same metadata category to contain as many LDA topics as possible. The reason is simple – we want to capture lower granularity of detail by the LDA topics.

In order to quantify abovementioned constraints, for each parameter iteration, we computed the utility function as:

$$\text{Score} = \frac{\frac{Avg(LDA\ topics)}{|Topics|} + 1 - \frac{Avg(Metadata\ categories)}{|Categories|}}{2} \quad (6)$$

where the *LDA topics* represent the number of unique LDA topics that are associated with deals sharing the same metadata category. Similarly, *Metadata categories* represent the number of unique metadata categories that are associated with deals sharing the same LDA topic. As a result, the best score was obtained with parameters set as follows: $passes = 20, probability = 0.01, topics = 70$. In other words, the dataset item-based abstraction is described through 70 topics (which is clearly less as assigned by an expert).

FP-growth hyperparameters

The behavior abstraction method uses the association rules as one of the approaches for the abstraction. The mining of association rules was performed via the FP-growth algorithm.

The FP-growth algorithm considers a set of parameters: *minimum support* and *minimum confidence*. As a rule, larger datasets require smaller values of these parameters and vice versa. The choice of optimal parameters was based on the number of generated rules (we require at least several hundreds or thousands of generated rules).

As we can see (Table 4), lower parameters of minimum support, and minimum confidence increased the number of generated rules. However, the average support and average confidence of found rules decreased also. A compromise between a large number of generated rules and expected average support and confidence is used as a result: $support_{min} = 0.005, confidence_{min} \in \langle 0.5, 0.8 \rangle$.

Parallel PrefixSpan algorithm hyperparameters

The sequence pattern mining is a part of the behavior abstraction along with the N-grams and association rules. The sequence pattern mining was performed via the Parallel PrefixSpan algorithm. We focused on two major parameters: *minimum support* and *maximal pattern length*, depending on the size of the dataset. 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 (Table 6). The maximal pattern length in general influences the uniqueness of found patterns and therefore should be set to lower values (shorter patterns are more general - they are more useful for following machine learning tasks).

Random forest hyperparameters

The gender prediction was used as an evaluation task in our both research questions. The prediction was realized by a Random Forest algorithm with these parameters: the *number of estimators*, *criterion*, *maximal depth*, *max. features* and *minimum sample split*.

The number of estimators corresponds to the number of trees that are used in the forest. The higher values of trees, in general, increase the performance of the algorithm but make it also slower. Usually, the increase of maximum features in general also increases the performance of Random Forest. However, it is not always true - the increase in

max. features decrease the diversity of individual trees. The max. depth limits the depth, i.e., the number of levels in an individual tree. We used a grid search method (focusing on *Recall macro* metric) and 10-fold cross-validation. (Table 10).

Table 10: The hyperparameter tuning of Random Forest. The best values are presented for AR - association rules, SP - sequence patterns, N - N-grams, T - patterns based on LDA, EC - patterns based on expert categories.

Features	Parameters				
	No. of estimators	Criterion	Max depth	Max features	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