Taylor & Francis
Taylor & Francis Group

Journal of Organizational Computing and Electronic Commerce

ISSN: 1091-9392 (Print) 1532-7744 (Online) Journal homepage: http://www.tandfonline.com/loi/hoce20

Application of the Information Bottleneck method to discover user profiles in a Web store

Jacek Iwański, Grażyna Suchacka & Grzegorz Chodak

To cite this article: Jacek Iwański, Grażyna Suchacka & Grzegorz Chodak (2018) Application of the Information Bottleneck method to discover user profiles in a Web store, Journal of Organizational Computing and Electronic Commerce, 28:2, 98-121, DOI: 10.1080/10919392.2018.1444340

To link to this article: https://doi.org/10.1080/10919392.2018.1444340

Published online: 26 Mar 2018.

JOURNAL OF ORGANIZATIONAL COMPUTING AND ELECTRONIC COMMERCE
2018, VOL. 28, NO. 2, 98–121
https://doi.org/10.1080/10919392.2018.1444340

Taylor & Francis
Taylor & Francis Group

Application of the Information Bottleneck method to discover user profiles in a Web store

Jacek Iwański[a], Grażyna Suchacka[a], and Grzegorz Chodak[b]

[a]Institute of Mathematics and Informatics, University of Opole, Opole, Poland; [b]Department of Operations Research, Wroclaw University of Science and Technology, Wroclaw, Poland

ABSTRACT

The paper deals with the problem of discovering groups of Web users with similar behavioral patterns on an e-commerce site. We introduce a novel approach to the unsupervised classification of user sessions, based on session attributes related to the user click-stream behavior, to gain insight into characteristics of various user profiles. The approach uses the agglomerative Information Bottleneck (IB) algorithm. Based on log data for a real online store, efficiency of the approach in terms of its ability to differentiate between buying and non-buying sessions was validated, indicating some possible practical applications of the our method. Experiments performed for a number of session samples showed that the method is capable of separating both types of sessions to a large extent. A detailed analysis was performed for the number of clusters ranging from two to seven, and the results were compared to those achieved by applying the most common clustering algorithm, *k*-means. Increasing the number of clusters generally leads to better results for both algorithms. However, IB demonstrated much higher average efficiency than *k*-means for the corresponding number of clusters, and this superiority was especially clear for lower number of clusters. The IB-based division of user sessions into seven clusters gives the mean entropy value of 0.28, which means the 95% separation of sessions of both types. Furthermore, a big advantage of our approach is that it gives a possibility to analyze the probability distribution of session attributes in individual clusters, which allows one to discover hidden knowledge about common characteristics of various user profiles and use this knowledge to support managerial decisions.

KEYWORDS

Agglomerative Information Bottleneck; unsupervised classification; clustering; machine learning; data mining; e-commerce; Web store; user profile; customer profile

Introduction

Over the last few years, we have been witnessing the rapid development of the electronic commerce. With the widespread use of Internet-connected mobile devices and the development of advanced e-commerce support Web-based services, like safe online payment tools, multi-store price comparers, or product search engines, the pace of e-commerce development is now faster than ever.

The electronic environment makes it possible to collect a variety of data on Web users' behavior during their visits to an online store and to analyze it in detail. Results may be used to improve and personalize service offered to users. Advanced data mining techniques, like collaborative filtering (CF) (Kumar and Bala 2017; Wang et al. 2015), make it possible to predict interests and needs of a user based on information collected from many other users, considered similar to the given user. From an e-customer's point of view, such techniques make it easier for the customer to find products meeting his or her needs. From an online retailer's perspective, they provide an opportunity to offer better quality of service, gain loyal customers, and increase sales.

CONTACT Grażyna Suchacka ✉ gsuchacka@uni.opole.pl Institute of Mathematics and Informatics, University of Opole, ul. Oleska 48, Opole 45-052, Poland.

This paper deals with the problem of discovering groups of similar Web users based only on some observed features of their click-stream behavior on an e-commerce website. We introduce a new approach to the unsupervised classification of user sessions in a Web store to discover various e-customer profiles and to investigate differences in characteristics of several distinct types of buying and non-buying sessions. The approach uses the Information Bottleneck (IB) clustering algorithm. Its efficiency is validated by using real data on e-commerce traffic obtained from HTTP-level Web server log files. The obtained results indicate some possible practical applications of the approach, which may be developed in the future, for example, in an automated process of customers' classification (to one of the previously determined clusters), integrated with a data source (a stream of HTTP requests).

The main contribution of our study includes:

(1) a proposal of a set of user session features significant in terms of differentiating between various behavioral patterns on an e-commerce site
(2) a new approach to clustering Web user sessions using the Information Bottleneck method
(3) a case study verifying the efficiency of the approach using real e-commerce data

The rest of the paper is organized as follows. First, the Information Bottleneck method is presented. Then, related work is overviewed with a focus on two aspects: 1) applications of IB in the Internet environment and 2) methods applied to discover e-customer profiles using unsupervised machine learning algorithms. Then we discuss our research methodology and experimental results. The last section summarizes our findings and outlines directions of future work.

Information Bottleneck method

The Information Bottleneck method was introduced by Tishby, Pereira, and Bialek (1999), who generalized their previous findings on clustering nouns according to their presence in phrases with certain verbs (Pereira, Tishby, and Lee 1993). The idea of the IB method is to find a compact representation (clusters) of a random variable A that preserves the maximum information about a related random variable C, given some joint probability distribution for these two variables. For example, having a random variable A, which relates to multiple user sessions in an online store, and a random variable C, which relates to some features (attributes) describing the sessions, we aim at constructing an assignment of A into a set of clusters B that will preserve as much as possible original information about C.

A key quantity examined in the IB method is the mutual information $I(A;B)$ of two random variables A and B, which in a discrete case is given by the formula

$$I(A;B) = \sum_{i=1}^{|A|} \sum_{j=1}^{|B|} p(a_i, b_j) \log \frac{p(a_i, b_j)}{p(a_i)p(b_j)}, \tag{1}$$

where $p(a_i, b_j)$ is the joint probability distribution for A and B, $p(a_i)$ is the marginal probability distribution for A, and $p(b_j)$ is the marginal probability distribution for B.

We search for a compact representation B of a random variable A that preserves the maximum information $I(B;C)$ about a related random variable C. Formally, we minimize $I(A;B)$ under the constraint $I(B;C) \geq D$, where an arbitrary constant D is a threshold. This can be expressed by the formula

$$R(D) = \min_{\{p(b_j|a_i)\,:\,I(B;C) \geq D\}} I(A;B). \tag{2}$$

The optimization is performed over all normalized conditional distributions $p(b_i|a_i)$, for which the constraint $I(B;C) \geq D$ is satisfied. $R(D)$ is called the relevance-compression function. One can define the well-known relations (3) and (4) between probability distributions, assuming the following Markovian independence of the random variables A, B, C: $B \leftrightarrow A \leftrightarrow C$. It comes from the fact that B, as a compact representation of A, should depend only on A.

$$p(b_j) = \sum_{i,k} p(a_i, c_k, b_j) = \sum_i p(a_i) p(b_j|a_i), \tag{3}$$

$$p(c_k|b_j) = \frac{1}{p(b_j)} \sum_i p(a_i, c_k, b_j) = \frac{1}{p(b_j)} \sum_i p(a_i, c_k) p(b_j|a_i). \tag{4}$$

This is a standard optimization problem with constraints, which is solved using the Lagrange multipliers method. Let us define the Lagrange functional

$$L[p(b_j|a_i\ )] = I(A;B) - \beta I(B;C), \tag{5}$$

where β is a Lagrange multiplier. The compactness of B depends on the value of β. If $\beta \to 0$, and then the compactness prevails and $I(A;B) = 0$, which means that we have just one cluster. On the contrary, if $\beta \to \infty$, then $I(B;C) = I(A;B)$ and B is a copy of A. A detailed discussion on the solution can be found in (Tishby, Pereira, and Bialek 1999); here we just present the result. The conditional probability distribution $p(b_j|a_i)$ is a stationary point of the functional (5), if and only if for each b_j and a_i the following formula is satisfied

$$p(b_j|a_i) = \frac{p(b_j)}{Z(a_i, \beta)} e^{-\beta D_{KL}[p(c_k|a_i)||p(c_k|b_j)]}, \tag{6}$$

where:

$$Z(a_i, \beta) = \sum_j p(b_j)\ e^{-\beta D_{KL}[p(c_k|a_i)||p(c_k|b_j)]}, \tag{7}$$

where the normalization function $Z(a_i, \beta)$ is called the partition function. $D_{KL}(p_1||p_2)$ is a Kullback-Leibler divergence of a probability distribution, given by

$$D_{KL}(p_1||p_2) = \sum_{i=1}^{|A|} p_1(a_i) \log \frac{p_1(a_i)}{p_2(a_i)}. \tag{8}$$

Unfortunately, to find the solution for a given problem, one has to apply numerical iterative methods because $p(b_j)$ and $p(c_k|b_j)$ depend on $p(b_j|a_i)$ through Equations (3) and (4).

A number of iterative algorithms have been proposed to solve this problem. Two basic algorithms were introduced in (Pereira, Tishby, and Lee 1993) and (Tishby, Pereira, and Bialek 1999). Modified algorithms have been discussed in (Slonim and Tishby 1999) and (Slonim, Friedman, and Tishby 2002). Depending on the chosen algorithm, one can perform soft or hard clustering. In the soft clustering, each value a_i of random variable A belongs to every cluster b_j with certain probability, whereas in the hard clustering, each value a_i belongs to a single cluster b_j so the conditional probabilities $p(b_j|a_i)$ are equal 0 or 1.

In this study, we apply the agglomerative Information Bottleneck (aIB) algorithm, introduced in (Slonim and Tishby 1999) to perform the hard clustering. This greedy algorithm finds a hierarchical bottom-up clustering tree. The algorithm maximizes the functional

$$L_{max} = I(B;C) - \beta^{-1} I(B;A). \tag{9}$$

It starts clustering at $B = A$ (every value of A is in a single cluster), and then it reduces the number of clusters by merging two selected clusters b_m and b_n into one cluster b^*. We assign probabilities $p(b^*|a_i)$, $p(b^*)$, and $p(c_k|b^*)$ to the newly created cluster b^* according to the following formulas

$$p(b^*|a_i) = p(b_m|a_i) + p(b_n|a_i), \tag{10}$$

$$p(b^*) = p(b_m) + p(b_n), \tag{11}$$

$$p(c_k|b^*) = \pi_m * p(c_k|b_m) + \pi_n * p(c_k|b_n), \tag{12}$$

$$\pi = \{\pi_m,\ \pi_n\} = \left\{\frac{p(b_m)}{p(b^*)},\ \frac{p(b_n)}{p(b^*)}\right\}. \tag{13}$$

The key point of the algorithm is to choose clusters to merge. Because we maximize (9), we choose these clusters that minimize $\Delta L_{max}(b_m,\ b_n)$ given by

$$\Delta L_{max}(b_m,\ b_n) = L_{max}^{(before)} - L_{max}^{(after)}. \tag{14}$$

Merging is performed until the required number of clusters is reached or until a single cluster containing all values of A remains.

The time complexity of the aIB algorithm is $O(|A|^3|C|)$. For a detailed discussion on the Information Bottleneck theory and algorithms, we refer to (Slonim 2002).

Related work

Application of the IB method in the Internet environment

Early implementations of the Information Bottleneck method concerned document classification with respect to the word frequency distribution in the documents (Slonim and Tishby 2000). Since then, the method has been applied to clustering problems in multiple areas. Example applications include unsupervised image classification (Gordon, Greenspan, and Goldberger 2003), extraction of relevant speech features (Hecht and Tishby 2005), analysis of proteins structures (Ofran and Margalit 2006), clustering neural codes of a fly visual system (Schneidman et al. 2002), or classification of galaxy spectra (Slonim et al. 2001).

Regarding the Internet environment, applications of the IB method have focused on the analysis of the Web text contents. In (Du and Tan 2009) and (Ganu, Kakodkar, and Marian 2013), IB was applied to the sentiment analysis through opinion mining. Du and Tan (2009) proposed a framework based on the improved IB algorithm to discover the hidden sentiment association in online reviews between a review feature set and an opinion word set. Based on categories of the review features and groups of opinion words extracted from the review corpus, the proposed mechanism clusters product features and opinion words simultaneously. Ganu, Kakodkar, and Marian (2013) proposed an approach to derive ratings from the review contents and make fine-grained predictions of user sentiments toward the individual topics covered in reviews. Based on the textual structure and sentiment extracted from the reviews, an iterative IB algorithm clusters like-minded users for personalized recommendations with high accuracy.

Another area of IB application to the Web contents has been Web search support (Ikeda et al. 2009; Ling et al. 2008). Ling et al. (2008) proposed an approach for classifying unlabeled Web pages in Chinese using labeled Web pages in English as training documents. By using the IB algorithm, the common part in the two languages could be extracted and used for classification. The novel approach achieved the best average performance results compared to traditional supervised classification algorithms, supporting cross-language Web search. Ikeda et al. (2009) proposed an algorithm for disambiguating person names in Web search results to deal with the "same name" problem in Web searches. A document clustering algorithm uses named entities, compound keywords, and URLs as features for computing the similarity of documents.

Aside from the "typical" application of Information Bottleneck clustering to text documents, there have also been successful approaches to using this technique for community detection in "social networks" (Ziv, Middendorf, and Wiggins 2005) and for intrusion detection based on network traffic anomalies (Panda and Patra 2009). Ziv, Middendorf, and Wiggins (2005) applied the IB algorithm to discover modules in a collaboration network of paper coauthors. Results showed that the scientific collaboration is determined to a large extent by the institutional affiliation and geography. Panda and Patra (2009) used a sequential IB clustering in the network intrusion detection system. Network traffic features used in the clustering algorithm were based on TCP/IP-level connection data. The proposed approach proved to be effective in terms of high intrusion detection accuracy and low false positive rates.

Methods applied in a click-stream analysis on e-commerce sites

Analysis and characterization of user sessions on websites have been typically performed using data recorded in Web server logs. HTTP data from logs makes it possible to reconstruct user online behavior and perform a click-stream analysis. Some studies have addressed problems connected with log data preprocessing and cleaning, including identification and elimination of Web bot traffic (Chen, Fu, and Tong 2003; Huiying and Wei 2004). Other studies have been devoted to the analysis of user navigation paths and discovery of sequential patterns in Web stores (Adnan et al. 2011; Shim, Choi, and Suh 2012). This has resulted in developing user session models for e-commerce sites — for all customers (Jenamani, Mohapatra, and Ghose 2003; Kwan, Fong, and Wong 2005) and for multiple customer groups (Chang, Hung, and Ho 2007). Session models have been used in synthetic Web traffic generators and Web benchmarks in evaluating Web server system performance under various request service mechanisms (Suchacka and Borzemski 2013; Zhou, Wei, and Xu 2006). However, the main motivation to monitor and analyze users' online behavior have come from multiple real-life examples, showing that behavioral data may be successfully implemented to personalize online service strategies, thus allowing online retailers to increase their return on investment (ROI). Especially successful technique has been collaborative filtering (CF), a method of making automatic predictions about needs or interests of a user based on information on preferences collected from other users, considered "similar" to the given user. CF has revealed a great potential in the e-commerce environment, especially in online recommender systems (Adomavicius and Tuzhilin 2005; Jiang, Song, and Feng 2006).

Advanced Web usage mining methods make use of artificial intelligence and machine learning techniques. In the context of our study, two main classification approaches may be distinguished: supervised and unsupervised classification. Supervised classification consists in using a training dataset of observations, each described with an input object (typically represented as a vector of features) and the corresponding output value (a class label). Based on the training dataset a classifier is built, able to infer classes for future, unknown observations. Examples of supervised classification techniques successfully applied to e-customers data are Naïve Bayes classifier (Poggi et al. 2007), *k*-Nearest Neighbors (Cho and Kim 2004; Soiraya, Mingkhwan, and Haruechaiyasak 2008; Suchacka, Skolimowska-Kulig, and Potempa 2015a), decision trees (Poggi et al. 2007), decision forest (Hop 2013), Support Vector Machine (Chen, Fan, and Sun 2012; Soliman et al. 2012; Suchacka, Skolimowska-Kulig, and Potempa 2015b), and artificial neural networks (Chou et al. 2010; Suchacka and Stemplewski 2017).

In this paper, we focus on the latter classification approach. Unsupervised classification, also called clustering, aims at dividing a set of observations into groups (clusters) of objects that are similar in some sense to one another and dissimilar from objects in other clusters. Instead of relying on a training dataset of labeled observations, it deals with finding hidden relationships between unlabeled data. Example clustering algorithms include *k*-means, fuzzy *c*-means, hierarchical clustering, and mixture of Gaussian.

The majority of approaches to unsupervised classification of Web user sessions have been based on the *k*-means algorithm, the oldest and the most popular clustering technique. One of the first such studies was proposed in (Menascé et al. 1999). A user session in an online store was represented as a state transition graph, called a CBMG (*Customer Behavior Model Graph*). Sessions were described in terms of session states, corresponding to types of operations typical for an e-commerce site, performed by a user during the session. *K*-means clustering resulted in a set of CBMGs, characterized by different probabilities of making a purchase in the online store.

Another example of using the *k*-means algorithm is clustering of customer values, expressed as *recency, frequency, and monetary* (RFM) values (Cheng and Chen 2009). The clustering result was a set of clusters characterized by various levels of customer loyalty. Then, using historical transaction data, classification rules were extracted by rough sets (the LEM2 algorithm), allowing one to find some customers' characteristics and to use them to improve *Customer Relationship Management* (CRM).

A number of approaches combined *k*-means clustering with association rule mining (Carmona et al. 2012; Chang, Hung, and Ho 2007; Mohammadnezhad and Mahdavi 2012; Nenava and Choudhary 2013). In a recommender system proposed in (Nenava and Choudhary 2013), a modified *k*-means algorithm was combined with distributed association rules to discover group profiles of m-customers. Profiles were created based on users' searches and then used to generate online recommendations. Another recommender system was proposed for a mobile e-tourism site in (Mohammadnezhad and Mahdavi 2012). First, Web users were clustered using the *k*-means algorithm. Then, tours ordered by the users in individual clusters were analyzed, and association rules were discovered using A-priori algorithm. Based on their ordering history, users visiting the site were assigned to one of the clusters and received the customized tour recommendations.

In (Chang, Hung, and Ho 2007) *k*-means clustering was combined with association rule mining to predict purchasing behavior of potential customers. The proposed anticipation model uses loyal customers' clusters, which are built based on past purchasing behavior of loyal customers. From loyal customers' profiles, so-called past purchasing pendency values are derived. Using association rules, potential customers are identified and provided with the personalized products recommendations.

Similar data mining techniques were applied in (Carmona et al. 2012). User sessions were described with such features as the type of Web browser used, the referrer, session duration, the number of clicks, or keywords used. Sessions were clustered using *k*-means, and association rules were discovered for each cluster using A-priori algorithm. Then, subgroups were discovered using the evolutionary fuzzy algorithm, and fuzzy rules were mined for the subgroups. The goal was to improve a design of the e-commerce website to increase the site usability and user satisfaction.

Other unsupervised classification approaches, aimed at discovering Web user profiles, combined vector analysis and fuzzy clustering. They utilized information about the site organization (Joshi, Joshi, and Krishnapuram 2000; Song and Shepperd 2006). An approach proposed in (Joshi, Joshi, and Krishnapuram 2000) is based on robust fuzzy clustering. It requires computation of the relation matrix, describing dissimilarities between all session pairs. Dissimilarity measures incorporate both the site structure and the URLs visited in the sessions. Two fuzzy clustering algorithms were applied: FCMdd (*Fuzzy c Medoids*) and FCTMdd (*Fuzzy c Trimmed Medoids*). As a result, profiles of users with various information goals were identified.

In (Song and Shepperd 2006), vector analysis and fuzzy set theory-based methods were used to cluster both Web users and Web pages. A website topology was represented as a directed graph, and users visiting the site were characterized based on URLs of the visited pages. Based on discovered Web page clusters, frequent access paths were identified, taking into account the underlying website structure.

A rough leader clustering algorithm for e-commerce sessions was developed in (Su and Chen 2015). Session features were defined taking products' categories into consideration and included: a visiting sequence, frequency of visits, and time spent on each category. The authors modified the leader clustering algorithm, which is able to discover a set of leaders (i.e., the cluster representatives) in only a single pass through the dataset. A rough set theory was integrated to this algorithm to reflect rough characteristics of a user's interest in products available in the online store.

Other clustering approaches used a Kohonen neural network. For example, in (Zhang, Edwards, and Harding 2007), a user session was considered as the user's browsing activity in retrieving a series of Web pages. Thus, each session was represented as an *n*-dimensional vector over the space of all query strings and all time intervals between each two consecutive searches. Then, a Kohonen neural network was applied to discover user profiles. The resulting model was able to create clusters of queries related to user sessions. Its goal was to predict Web links or products that an active user may be interested in.

To the best of our knowledge, the Information Bottleneck algorithm has not yet been applied to discovering Web user profiles or customer profiles. As regards the problem under consideration in our study, that is, the classification of buying and non-buying user sessions on e-commerce sites, some supervised learning techniques have been successfully applied, as discussed in this section. However, a drawback of these methods is that they require indicating the desired output value (the supervisory signal), so their ability to discover hidden relationships in the data is limited. In contrast, the Information Bottleneck method belongs to unsupervised learning techniques. Moreover, it makes it possible to analyze the probability distributions of object features for the generated clusters and, thus, to explore hidden knowledge underlying the resulting cluster split. A big advantage of IB is its capability of dealing with high-dimensional data, which is the case of session-level Web usage data.

Research methodology

Problem formulation

The main task of unsupervised classification of user sessions on an e-commerce site is to discover differentiated user profiles characterizing online behavior of buyers and non-buyers. Our previous studies (Suchacka and Chodak 2016; Suchacka, Skolimowska-Kulig, and Potempa 2015a, 2015b; Suchacka and Stemplewski 2017) showed the high efficiency of supervised classification of user sessions in an online bookstore as regards the fact of making a purchase in session or not. Thus, we can expect that sessions ending with a purchase (which we call *buying sessions*) have quite different profiles than sessions without a purchase (called *browsing sessions* or *non-buying sessions*). Our hypothesis is that based on session features, determined from a user click-stream behavior on a website, it is possible to distinguish several various user session profiles, each of which is characteristic for either buying sessions or browsing ones.

We describe each session with a number of features (attributes), which are used in the clustering process for the given number of clusters. Furthermore, each session has a label assigned that corresponds to the session class ("*buying*" or "*browsing*"). Class labels are used to evaluate clustering efficiency in terms of "purity" of the generated clusters as regards the session class. We also investigate characteristics of discovered user profiles, determined by the clusters.

Reformulation of a user session for use in the aIB algorithm

Raw data used in the analysis is typical HTTP-level data extracted from Web server log files. The data corresponds to the traffic in a Web bookstore registered from April to September 2014 (the store identity is not revealed in the paper due to a non-disclosure agreement). The bookstore offers mainly traditional books, audiobooks, and computer games; besides, an e-commerce module is integrated with some entertainment interactive contents, providing users with access to short movies, quizzes, online mini games, and so on.

Visits of Web users to the analyzed website are represented as user sessions, reconstructed from logs. Each session is a sequence of HTTP requests with the same IP address and user agent string, assuming that a gap between two consecutive sessions of the same user is not shorter than half an hour. After excluding sessions containing only one page and sessions identified as the ones performed by Web bots - according to the approach proposed in (Suchacka 2014), the session dataset contained 33,354 sessions: 873 buying sessions and 32,481 browsing ones.

Applying the aIB algorithm requires the selection of two discrete random variables: A (classification variable) and C (relevant variable), and calculation of the joint probability distribution $p(a_i, c_j)$ for them. Because our aim is to cluster user sessions, a natural choice of A is a session identifier (session number). A choice of C is not so obvious. Our data describing each user session is a result of the click-stream analysis, and thus, it includes mainly numbers of user clicks on certain types of pages and some aggregated statistics determined for the session. We distinguish 20 user session features (attributes), which will be transformed to values of C. The features are summarized in Table 1 (note that in the remainder of the paper, we will refer to the features' numbers).

Values of each session attribute were normalized to fit in the range (0, 1). Furthermore, regarding the IB clustering, we additionally normalize all values of the session attributes to map them to joint probabilities so that the sum of all normalized attribute values for all the sessions is 1. As regards the click-related attributes, our interpretation is intuitive: the more times a user enters a page assigned to a certain type, the higher the probability of spending time on pages of that type by a user is. Analogically, one can interpret the time-related attributes. During this normalization, we assume that the marginal probability distribution of the random variable C is uniform and the original value of an attribute is proportional to the probability assigned. To sum up, we chose the session numbers for the random variable A and session attributes' numbers for the random variable C.

The aIB algorithm has a relatively high computational complexity, $O(|A|^3|C|)$. It means that it is possible to investigate a sample of the size of the order of 10^4 sessions but in long time, whereas in real-time applications, clustering results should be generated in time of the order of milliseconds or seconds. The approach advanced in this paper operates offline, but having in mind its possible applications in real-time Web server systems in the future, we decided to validate its efficiency for the reduced samples of sessions. The size of each working dataset is 250 sessions, including 125 buying sessions and 125 browsing ones, randomly drawn from the 873 buying session set and 32,481 browsing session set, respectively. Because of the random character of such prepared data, we generated 100 different subsets (samples) and investigated possible statistical differences between them.

The agglomerative IB algorithm is applied for a given number of clusters, k. We experimented with multiple values of k, varying from 2 to 249. For each k, the analysis was performed for the same 100 samples — subsets of randomly drawn 125 buying and 125 browsing sessions. To show the

Table 1. Attributes of a user session used in the clustering process.

No.	Description
1	Number of page views, corresponding to user clicks
2	Number of HTTP requests in session
3	Total volume of data sent in session [KB]
4	Duration of the session [s]
5	Mean time per page [s] (excluding time spent by the user on the last page in session)
6	Number of views of the home page of the site
7	Number of registration trials, i.e., user's attempts to create an individual account
8	Number of successful registration operations
9	Number of successful logging on operations
10	Number of logging off operations
11	Number of views of the page containing description of shipping terms and conditions
12	Number of clicks corresponding to reading information about the total charge (including prices of products in the cart and the shipping fee) after starting the checkout process
13	Number of searches via the internal website search engine
14	Number of page views via following the internal site's hyperlinks, excluding pages used to determine other session features
15	Number of views of product description pages
16	Number of clicks corresponding to adding a product to the shopping cart
17	Number of attempts to start the checkout process
18	Number of views of pages containing information about the bookstore and the trading company,
19	Number of views of pages providing entertainment contents
20	Number of clicks resulting in untypical operations (e.g., in error messages)

significance of the choice of k for the clustering results, we start at discussing results for $k = 2$ (the lowest reasonable number of clusters) up to $k = 7$ (i.e., the most suitable number of clusters determined in Experimental results Section, leading to the most explanatory and potentially useful results for our scenario).

Classification-oriented measures of cluster validity

To assess the ability of our clustering approach to differentiate between buying and non-buying sessions, we measure the degree of correspondence between the cluster labels and the sessions' class labels ("*buying*" or "*browsing*"). We use a classification-oriented measure of cluster validity, entropy of the generated clusters (Tan, Steinbach, and Kumar 2006).

Let $p_{i,j}$ be the probability that a session in cluster i belongs to class j:

$$p_{i,j} = \frac{l_{i,j}}{l_i}, \tag{15}$$

where $l_{i,j}$ is the number of sessions of class j in cluster i and l_i is the number of all sessions in cluster i, $i = 1, 2, \ldots, k$, and $j = 1, 2$. Using the class distribution of sessions in cluster i, the entropy of the cluster may be determined. The entropy of cluster i is given by the formula

$$H_i = -\sum_{j=1}^{2} p_{i,j}\log_2 p_{i,j}. \tag{16}$$

The total entropy of a set of clusters is a sum of the cluster entropies weighted by the cluster sizes:

$$H = \sum_{i=1}^{k} \frac{l_i}{l} H_i, \tag{17}$$

where k is the number of clusters, l_i is the number of sessions in cluster i, l is the total number of sessions, and H_i is the entropy of cluster i.

Entropy is a measure of the degree to which individual clusters contain sessions of a single class. The minimum possible value of entropy is 0 (which means the perfect separation of buying and browsing sessions). Because the probabilities of sessions of both classes are the same (the numbers of buying and browsing sessions in each sample are equal), the maximum possible entropy value is 1 (which means that the session attributes do not allow us to recognize the session type).

We also calculate a percentage separation of sessions regarding their class labels, as the probability for the binary random variable corresponding to the calculated entropy. For example, an entropy lower than 0.47 means a degree of session separation higher than 90%.

Implementation of k-means as a reference algorithm

To assess the efficiency of IB in clustering e-customer sessions, we implemented k-means, the most common unsupervised classification technique. K-means is a hard partitioning algorithm, so each observation (data point) is assigned to only one cluster. Its goal is to divide data in groups characterized by relatively small intracluster distances and relatively large intercluster distances (Abbot 2014). The most popular distance metric in k-means has been Euclidean distance. The algorithm starts from choosing cluster centers for a given number of clusters, usually in a random way. Then, the distance between each cluster center and each data point is computed, and each point is marked with a label indicating its nearest cluster center. For each newly created cluster, mean attribute values are computed that lead to creating the new cluster center. This procedure is repeated iteratively until cluster membership does not change. The resulting clusters are described by clusters' centroids, that is, collections of attribute values.

To implement k-means clustering of e-customer sessions we used a professional, advanced analytics software packet Statistica. The analysis was performed for the number of clusters ranging from 2 to 10, for the same data samples as were used in the IB clustering. Values of each session attribute were normalized in the same way as for the IB clustering, to fit in the range (0, 1). Initial cluster centers were chosen by sorting distances and taking observations at constant intervals, with Euclidean distance as a distance measure. The quality of the k-means clustering was assessed with the entropy measure, and the results were compared to those of the IB clustering.

Experimental results

Experimental determination of aIB parameter values

The agglomerative Information Bottleneck algorithm requires choosing the value of the Lagrange multiplier, β. There are no strict rules for this choice but, as we discussed earlier, the value of the mutual information between discrete random variables B and C, $I(B;C)$, for a given number of clusters depends on it. Lower values of β are related to the increased contribution of $I(B;A)$ at the expense of $I(B;C)$. This usually results in generating a large cluster with a_i values containing little information about C. This may be useful for filtering out less distinctive sessions, but usually higher values of β give better results (while still allowing the not balanced clustering).

We investigated experimentally the dependence of $I(B;C)$ on the number of clusters for various β. Figure 1 presents the results for one sample of all 250 sessions (results for other samples lead to the same conclusions). Because the graph for the full range of possible k in Figure 1 does not allow us to observe changes of the curve shapes for lower k, results for k in the range of 1 to 20 are additionally presented in Figure 2. Experimental results show that β in the range of 7–15 is a good choice for our problem, and so we chose $\beta = 10$.

Figures 1 and 2 can also suggest some estimates about the suitable number of clusters separating different classes of e-customer click patterns. The more clusters there are, the more mutual information between B and C is preserved and vice versa. Thus, the goal is to find such a point on the graph for which there is as much mutual information between B and C as possible and the number of clusters is relatively low. The choice of the most suitable value or range of k may be made based on a visual inspection of curve shapes. Moving from the highest possible number of clusters

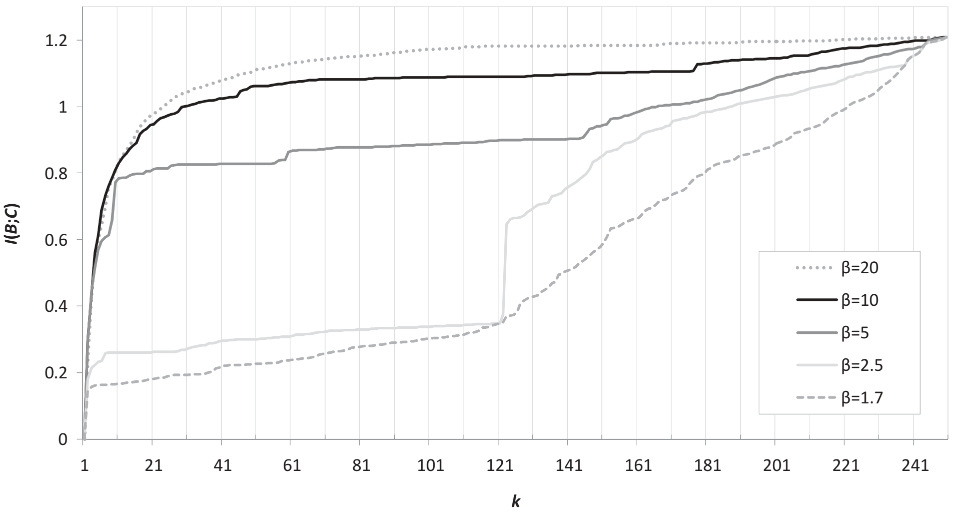

Figure 1. Mutual information between B and C as a function of the number of clusters for different β. (for the full range of k).

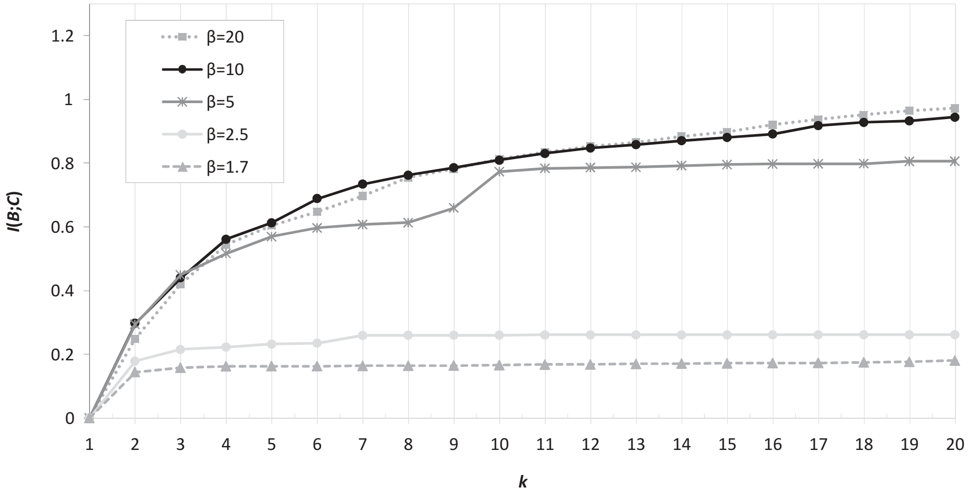

Figure 2. Mutual information between *B* and *C* as a function of the number of clusters for different *β* (for lower values of *k*).

toward lower values, we should consider ranges of *k* for which *I*(*B*;*C*) starts decreasing more rapidly. We can see in Figure 2 that for $\beta = 10$ an estimate for *k* will be approximately seven clusters.

Assessment of cluster validity with classification-oriented measures for various numbers of clusters

To evaluate the adequacy of IB clustering to differentiate between the two session classes, the entropy was determined according to (17) for all sample session sets for different numbers of clusters. Figure 3 shows basic entropy statistics for *k* ranging from 2 to 20. The three-plot series corresponds to the minimum, mean, and maximum entropy values over all 100 samples of 250 sessions each. For mean values the standard deviation is plotted as well.

It can be seen that for *k* between two and seven, the entropy decreases rapidly as the number of clusters increases. When *k* exceeds seven, the decrease in entropy with the increase in *k* is much

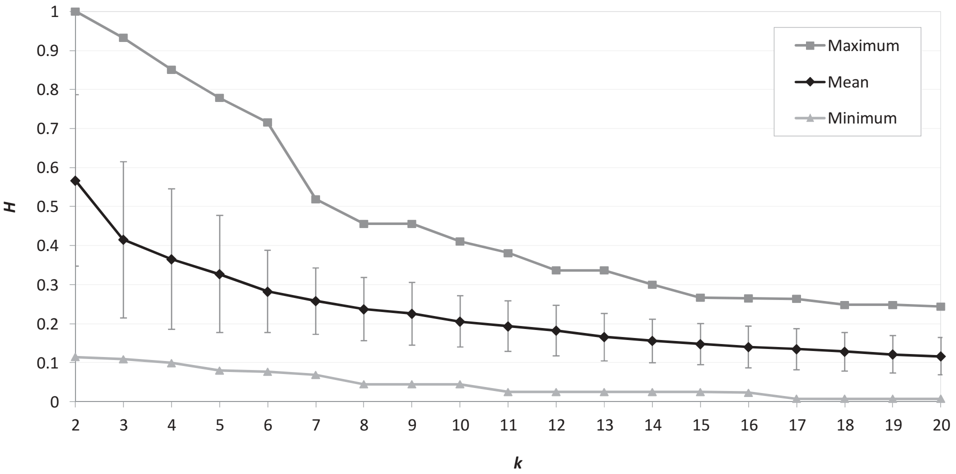

Figure 3. Entropy as a function of the number of clusters.

slower. This observation coincides with our previous findings regarding $I(B;C)$ in the previous Subsection.

It is interesting to observe in Figure 3 that even for a very coarse division of sessions into two clusters, the mean entropy value is less than 0.6. This is a very good result as it means that more than 85% of all sessions are well separated in terms of the session class. However, for $k = 2$, one can observe cases of both a very good session separation at the level of 98% ($H = 0.1$ for the best case) and a uniform session distribution between the clusters ($H = 1$ for the worst case), depending on a sample (cf. minimum and maximum H values for $k = 2$ in Figure 3). This observation leads to the conclusion that several distinct behavioral patterns may be distinguished for buying and browsing sessions in our dataset, not just one pattern for each session class.

Results for more clusters confirm this conclusion. The division of sessions into seven clusters gives the mean entropy value of 0.28, which means the 95% separation of sessions of both classes; for the best case $H = 0.07$ (more than 99% of session separation), and for the worst case $H = 0.5$ (89% of session separation). It is worth noting that for $k > 10$ for the best cases almost all sessions are perfectly separated into distinct clusters.

In-depth analysis of cluster characteristics

Because the IB method makes it possible to compute conditional probability distribution $p(c_i|b_j)$ of the random variable C for each cluster, we can analyze characteristics of individual attributes for the generated clusters in detail. Thus, we can discover hidden knowledge about common features of various profiles of buying and browsing sessions. In this section, we discuss results of the in-depth analysis of cluster characteristics for $k = 2$, 3, 5, and 7. We do not limit our discussion to the results obtained for the most suitable number of clusters ($k = 7$), determined in the previous part of this Section, to illustrate the impact of k on clustering results and to show a high potential of the method even for a very coarse-grained clustering (for $k = 2$).

We focus mainly on presenting results obtained for the best and worst case scenarios out of all 100 samples' results. Such an approach clearly illustrates drawbacks of a very coarse-grained division of data into few (2, 3, or 5) clusters, when results obtained for the best and the worst cases differ significantly. On the other hand, it can be seen that for more fine-grained divisions, particularly into seven clusters, differences between the best and worst cases are statistically insignificant. Thus, our approach to statistically analyze the best and worst cases for each number of clusters is helpful in illustrating the process of improving the results across the consecutive experiments, as we approach the most suitable number of clusters.

Furthermore, we decided to present a discussion on resulting probability distributions of session attributes in the clusters for one example sample (no. 25) for all k values under consideration. The goal is to illustrate the agglomerative way of the algorithm operation and to illustrate the merging patterns.

Division into two clusters

We start our analysis from the division of user sessions into two clusters. We discuss results for the best and worst case (in terms of the resulting entropy value) out of all 100 samples. The distribution of browsing and buying sessions among both clusters is presented in Figure 4. For the best case (Figure 4—left), one can observe a very clear separation of both kinds of sessions between the clusters: cluster 1 contains the vast majority (98%) of all browsing sessions, and cluster 2 contains most of all buying sessions. Good separation is reflected by the very low entropy value, equal to 0.11. On the other hand, for the worst case (Figure 4—right), sessions of both classes are almost randomly distributed among both clusters, which is confirmed by the entropy of 1.

Figures 5 and 6 illustrate the probability distributions of session attributes in both clusters for the best and worst case, respectively. It can be seen that in the best case, many browsing sessions classified to cluster 1 are characterized by much longer mean time per page (attribute 5) than buying

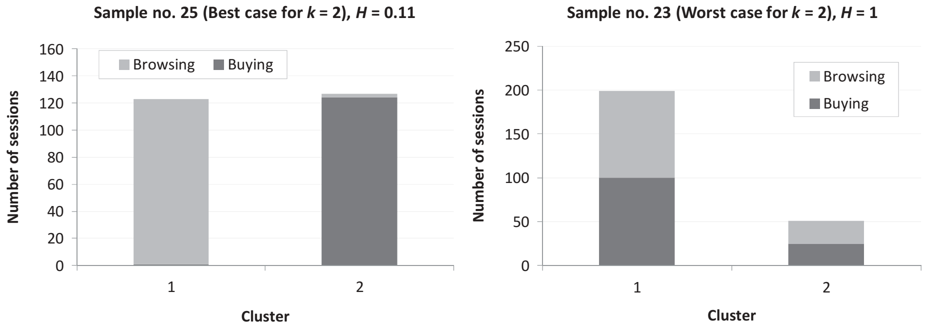

Figure 4. Distribution of browsing and buying sessions for two clusters, for the best case (left) and for the worst case (right).

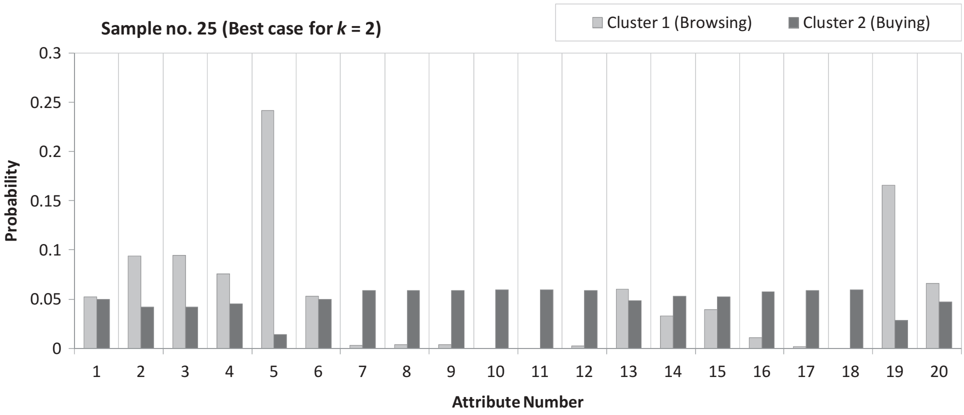

Figure 5. Probability distributions of session attributes in the clusters for the best case, $k = 2$.

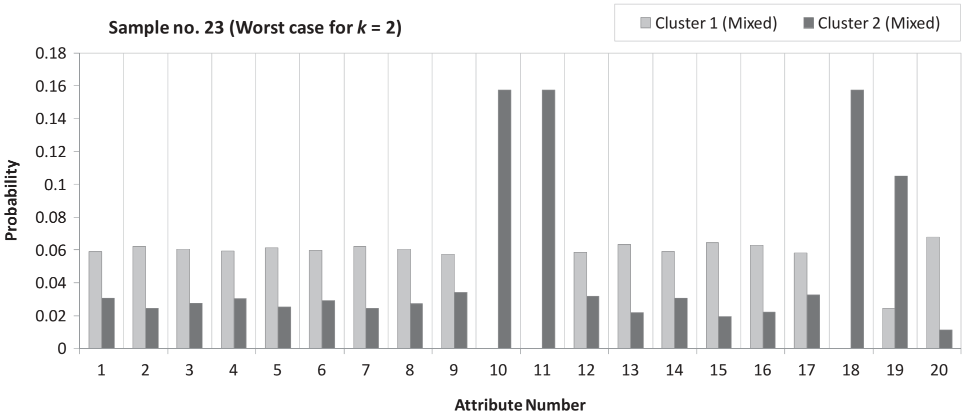

Figure 6. Probability distributions of session attributes in the clusters for the worst case, $k = 2$.

sessions. They are also described by the high probability of visiting pages with entertainment contents (attribute 19), which suggests that non-buyers spend a lot of time getting acquainted with the online mini games and multimedia contents available on the website. Furthermore, as opposed to users classified to cluster 2, users in cluster 1 hardly ever open pages related to registration (attributes 7, 8), logging on (9), logging off (10), starting the checkout process (17), as well as reading information about shipping terms (11, 12) and the store (18).

The results observed for the best case confirm the intuition that users who decide to make a purchase in a current session (cluster 2) spend relatively little time on individual pages (very low probability of attribute 5). Thus, they seem to be just acquainted with the store offer. If it were always the case, clustering results would be perfect in terms of session class differentiation. However, the analysis of the worst case scenario (with two mixed clusters) reveals that some browsing sessions have shorter mean time per page than some buying sessions. This suggests that probably some first-time buyers spend more time browsing the store offer, and thus, the difference in the attribute 5 is less pronounced. This is confirmed by the observation in Figure 6 that some visitors relatively often read information about the store (attribute 18) and shipping terms (attribute 11) and visit pages with multimedia contents related to the offered products (attribute 19). These users are probably very cautious customers. It is confirmed by the high probability of logging off (attribute 10). The worst case scenario is not a representative one, of course. Some customers visiting the store may exhibit unusual or abnormal behavior, and it is very likely that some samples contain relatively many such "outstanding" sessions. Nevertheless, the above observations lead to the conclusion that more than two clusters are needed due to the heterogeneity of profiles of buyers and non-buyers.

Division into three clusters

The distribution of browsing and buying sessions among three clusters for the best and worst cases is illustrated in Figure 7. The probability distributions of session attributes in the three clusters are presented for the best and worst case in Figures 8 and 9, respectively.

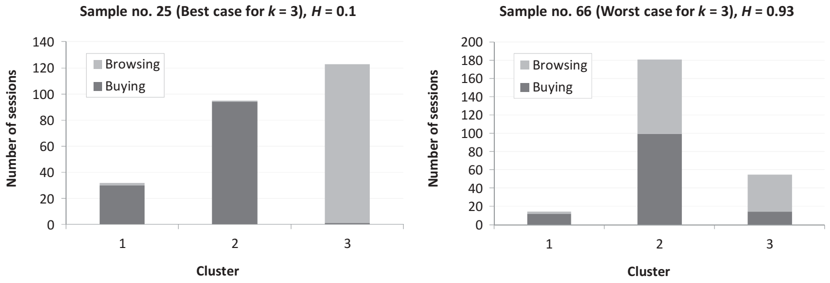

Figure 7. Distribution of browsing and buying sessions for three clusters, for the best case (left) and for the worst case (right).

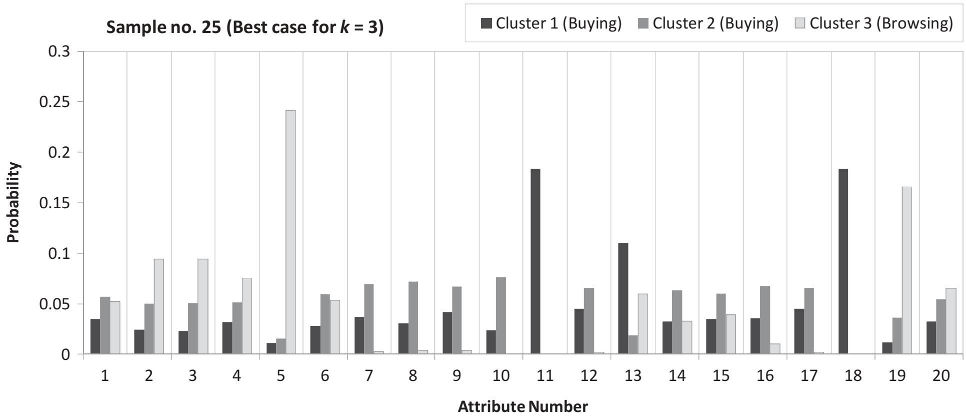

Figure 8. Probability distributions of session attributes in the clusters for the best case, k = 3.

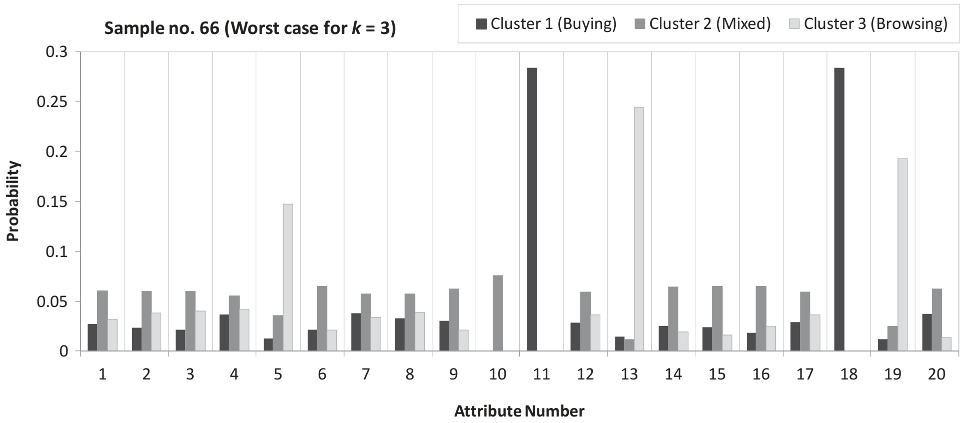

Figure 9. Probability distributions of session attributes in the clusters for the worst case, $k = 3$.

One can notice that in the best case there is a clear split into two different kinds of buying sessions and one kind of browsing sessions (Figure 7—left). As it can be seen in Figure 8, the first “buying” cluster (cluster 1) contains sessions characterized by distinguishable peaks of probability distribution for attributes 11 (pages informing about shipping terms and conditions), 18 (pages informing about the bookstore and the retailer), and 13 (searching for products). Thus, we can conclude that the aIB algorithm identified a group of first-time buyers in the analyzed Web store as a separate cluster. The second “buying” cluster (cluster 2) contains sessions of buyers who evidently visited the store before, as it reveals minima of probability distribution for the attributes 11, 18, and 13. Probability distribution of the attributes of the “browsing” cluster (no. 3) for the best case reveals maxima for the attributes 5 (mean time per page) and 19 (visits to pages with the multimedia entertainment contents), like for the case with two clusters — which is not surprising.

Considering the worst case for the three-cluster scenario (Figure 7—right), we can see that there are two quite clearly separated clusters (the “buying” cluster 1 and the “browsing” cluster 3), and the third one is mixed. This observation leads to the conclusion that splitting the sessions into three clusters is not optimal yet.

Division into five clusters

Because of a limited amount of space, we omit the discussion on the experiment results for four clusters and proceed to the in-depth analysis of the five-cluster case, which gave more interesting results, suggesting some practical implications for the online retailer.

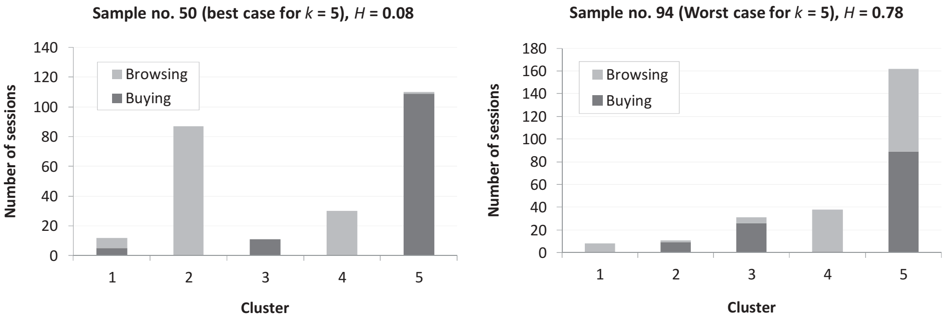

Figure 10. Distribution of browsing and buying sessions for five clusters, for the best case (left) and for the worst case (right).

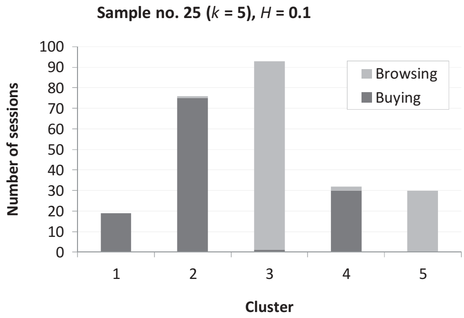

Figure 11. Distribution of browsing and buying sessions for five clusters for sample no. 25, k = 5.

Figure 10 shows distribution of sessions of both classes among five clusters for the best and worst cases. For the best case, results are summarized by a very low entropy value, equal to 0.08. Clustering results for this case (Figure 10—left) show a clear division of sessions into two "buying" clusters (no. 3 and 5), two "browsing" ones (no. 2 and 4); however, there is also one small mixed cluster (cluster no. 1). As we proceed toward samples with increasing entropy values, this mixed cluster grows. The worst case sample (Figure 10—right) resulted in one big mixed cluster (no. 5) and four smaller but clearly separated clusters. This suggests the need to fine-tune the clustering algorithm into more clusters for some rarely occurred samples containing untypical sessions.

Figure 11 shows distribution of sessions of both classes among five clusters for the sample no. 25 (the 3rd best sample in this case). Because it was the best case sample for k = 2 and k = 3, we analyze this case here in detail for comparative purposes. Clustering results for this sample show a clear division of sessions into three "buying" clusters (no. 1, 2, and 4) and two "browsing" ones (no. 3 and 5). Very good results are confirmed by a very low entropy value, equal to 0.1.

The in-depth analysis of probability distributions for "buying" clusters 1 and 2 for the sample no. 25 allows us to notice that these clusters have very similar characteristics in general (Figure 12). Buyers classified to these clusters seem to be well acquainted with the store offer and conditions — they probably visited the website before, completing most of their browsing and searching operations then. Their sessions reveal minima of probability distribution for the attributes corresponding to the following operations: reading pages with shipping terms and conditions (attribute 11), reading pages with information on the store (attribute 18), and searching for products (attribute 13). The main difference between these two groups is that users in cluster 1 perform the operation of logging off (attribute 10), as opposed to users in cluster 2.

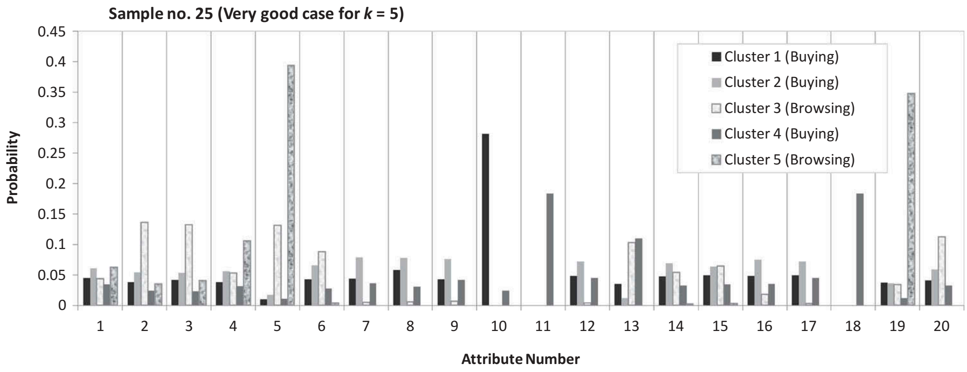

Figure 12. Probability distributions of session attributes in the clusters for sample no. 25, k = 5.

To conclude, users assigned to cluster 1 and 2 reveal very similar behavioral patterns, and from a practical point of view, their division into two separate target groups is not crucial. However, it may be taken into consideration in some marketing activities in which customer personality characteristics are important because the customers who intentionally log out are usually detail oriented and cautious.

Users classified to the third "buying" cluster (cluster 4) differ from buyers in clusters 1 and 2 significantly in that their sessions reveal much higher probability for the attributes 11, 13, and 18. This cluster may represent two types of buyers. The first type of buyers are customers who care about delivery time. The second type may be "institutional" customers (like schools or associations) because many views of pages with shipping terms and conditions suggest that the customers want to make sure that time and form of delivery will be acceptable for them and complies with the conditions required for spending available funds. This is a typical case of public institutions where the legal framework for spending funds is very strict. Besides, many searches via the internal search engine may indicate that an institutional customer is searching for a specific list of products.

An interesting group is a "browsing" cluster 5, which contains solely non-buying sessions. For these sessions, the probability distribution reveals high peaks for attributes 5 and 19 that correspond to long mean time per page and many views of the entertainment contents. At the same time, these sessions do not contain operations related to the user registration, logging on, reading information about the bookstore or shipping conditions, searching for products, adding products to the shopping cart, and attempts to start the checkout process (attributes 7, 8, 9, 10, 11, 13, 16, 17, and 18). Thus, it is evident that cluster 5 represents Web users who do not plan to make a purchase but are only interested in the added value of the bookstore, that is, pages containing short movies, quizzes, online mini games, and so on.

Division into seven clusters

The last experimental case discussed in the paper concerns division of Web user sessions into seven clusters, that is, a scenario with the best number of clusters, resulting from the analysis of dependence of $I(B;C)$ and H on k (discussed at the beginning of "Experimental results" Section). Distributions of sessions of both classes for the best and worst case are illustrated in Figure 13. One can observe that even in the worst case, all the clusters except one have a clear dominance of one session class.

We continue depicting an image of evolution of the sample no. 25 depending on the number of clusters. Figure 14 juxtaposes distribution of sessions of both classes into five and seven clusters, along with cluster cardinalities corresponding to session membership with regard to the session class. One can see that when we move from five to seven clusters, three clusters remain unchanged: two clusters grouping buyers just acquainted with the store (cluster numbers 1 and 2 for more and less cautious customers, respectively) and the cluster of non-buyers extensively exploring the

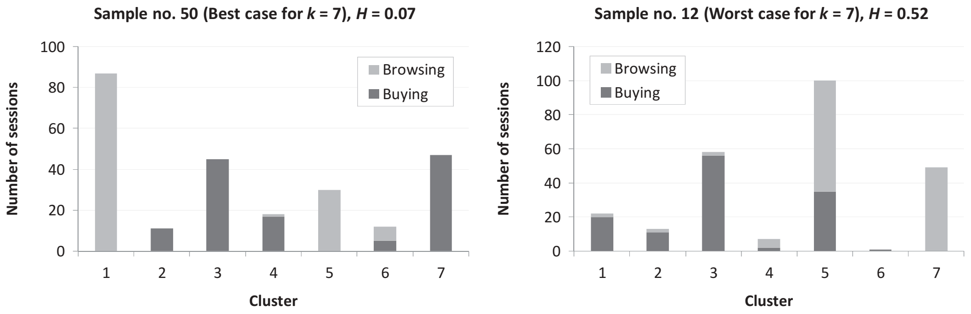

Figure 13. Distribution of browsing and buying sessions for seven clusters, for the best case (left) and for the worst case (right).

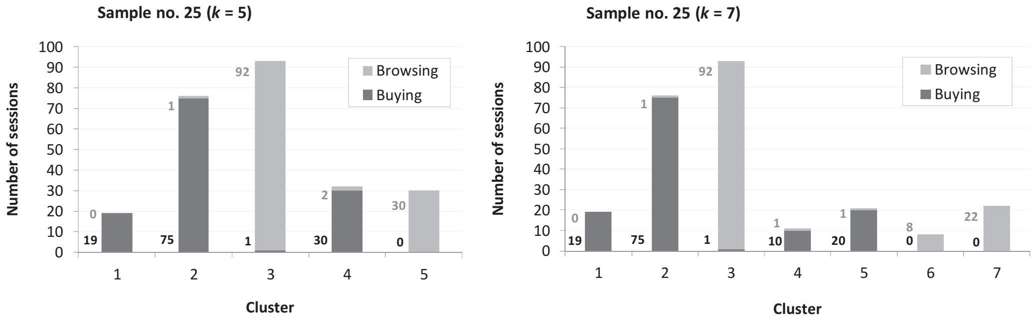

Figure 14. Comparison of distribution of browsing and buying sessions for five (left) and seven (right) clusters for sample no. 25. Clusters 1, 2, and 3 remain unchanged.

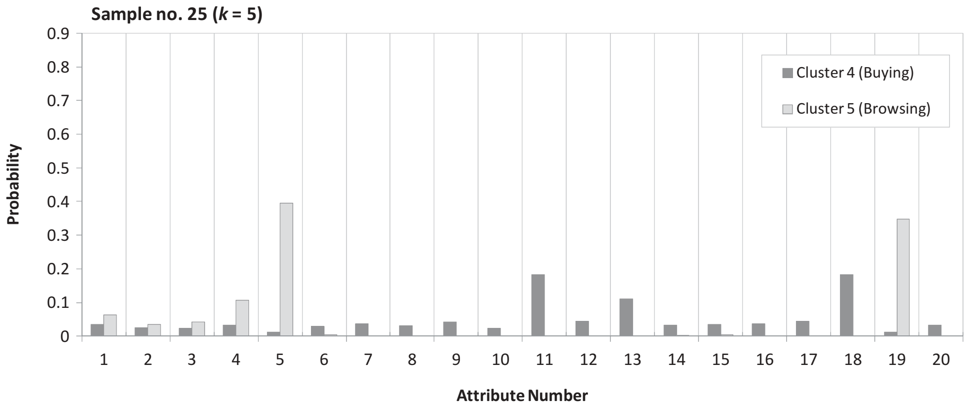

Figure 15. Probability distributions of session attributes in selected clusters (4 and 5), *k* = 5.

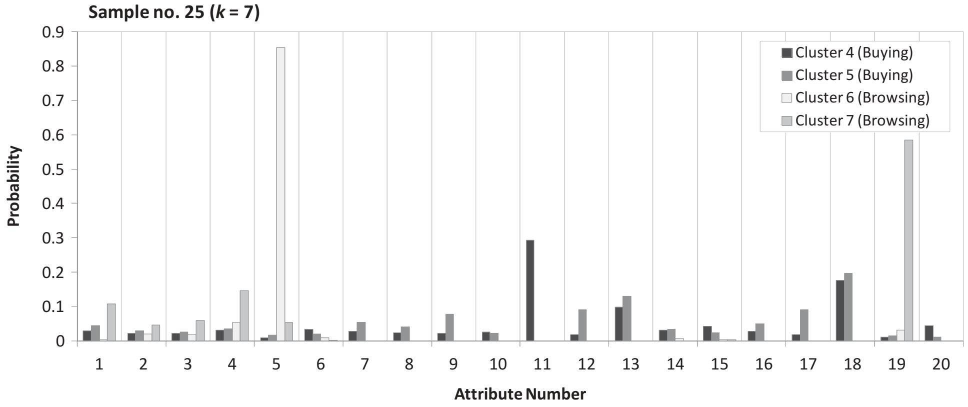

Figure 16. Probability distributions of session attributes in selected clusters (sample no. 25) which changed after increasing the number of clusters from five to seven, i.e., 4, 5 (the former cluster 4) and 6, 7 (the former cluster 5), *k* = 7.

bookstore (cluster no. 3). The former "buying" cluster 4 was entirely split into two new clusters (numbers 4 and 5), and the former "browsing" cluster 5 was split into new clusters 6 and 7.

Because visualization of probability distribution of session attributes for seven clusters is not well legible due to the large number of data, we illustrate only results for the clusters that changed.

Figure 15 corresponds to the results for k = 5, and Figure 16 presents the results for the newly created clusters for k = 7. The y-scale is the same on both graphs to facilitate observation of changes in values of probability distribution of the corresponding attributes.

As regard a group of 30 buyers from the former cluster 4 (with peaks for the attributes 11, 13, and 18 in Figure 15), the main difference between the corresponding new groups (clusters 4 and 5 in Figure 16) lies in the probability distribution of attribute 11. Ten buyers reassigned to cluster 4 often viewed pages containing shipping terms and conditions, whereas 20 buyers reassigned to cluster 5 did not. This suggests that buyers in cluster 4 may be both institutional customers, who want to make sure of the suitability of time and form of delivery, as well as other buyers who care about delivery time and want to make sure that they will receive ordered items on time. On the other hand, for cluster 5 we can observe higher values of probability distribution for the attributes 7, 8, 9, 12, and 17, which are related to the whole checkout process. This may suggest that this cluster represents less experienced customers. As regards possible practical implications of identifying such a customer group, an online retailer could personalize its service by introducing additional special information on viewed pages, facilitating the checkout process. If the retailer decides to use a real-time communication channel, like a live chat, such customers should be served first.

A group of 30 non-buyers from the former cluster 5 (with peaks for the attributes 5 and 19 in Figure 15) was split into two distinct subgroups. The first subgroup of eight users (cluster 6 in Figure 16) is characterized by the very high probability of spending a long time on visited pages (attribute 5) and the low probability of browsing the entertainment contents (attribute 19). Such a behavioral pattern may indicate users who are interested only in finding information on a specific product — they may have entered the site by following a search engine link, without the intention of making a purchase in this online bookstore (they may not be ready to order products at this moment but they may return to the bookstore in the future). The second subgroup of 22 users (cluster 7 in Figure 16) is evidently interested in the multimedia entertainment contents available on the site (attribute 19). Their sessions tend to last longer (attribute 4) and contain more page views (attribute 1) than in the case of the former subgroup of "information searchers."

To sum up, the analyzed case for the seven-cluster scenario (i.e., for the optimal number of clusters), seven user profiles were identified by using the unsupervised clustering technique:

(1) buyers well acquainted with the store offer and conditions (cluster 1)
(2) buyers well acquainted with the store, who are detail-oriented and cautious customers (cluster 2)
(3) institutional customers and first-time buyers very cautious with regard to shipping terms and conditions (cluster 4)
(4) other first-time buyers, who are less experienced customers (cluster 5)
(5) non-buyers extensively exploring the bookstore (cluster 3)
(6) non-buyers searching for information on specific products (cluster 6)
(7) non-buyers interested in multimedia entertainment contents (cluster 7)

In general, it can be noticed that the more clusters there are, the better separation of buying and browsing sessions between the clusters is achieved. For $k > 7$, even the worst case scenarios provide a very good separation of sessions of both classes and mixed clusters are small (the worst case entropy is less than 0.5, and its mean value is less than 0.25). However, the more clusters are generated, the more complicated is the analysis and visualization of common patterns of e-customer behavior within the clusters. Moreover, from the practical point of view, the cost of implementing a variety of marketing activities increases with the number of clusters.

Additional experiments, performed for larger samples (containing 873 sessions of each class), led to the same conclusions as the results obtained for 250-session samples. The only difference is less deviation from the mean entropy value observed for larger samples compared to the smaller ones. A huge disadvantage of using large data samples is very high computational cost. The computational

Table 2. Entropy of clusters generated by the IB and *k*-means algorithms.

	Max entropy (Worst case)		Min entropy (Best case)		Average entropy	
Number of clusters	*k*-means	IB	*k*-means	IB	*k*-means	IB
2	1.00	1.00	0.84	0.11	0.97	0.57
3	0.99	0.93	0.50	0.11	0.84	0.42
4	0.96	0.85	0.22	0.10	0.73	0.37
5	0.91	0.78	0.10	0.08	0.62	0.33
6	0.92	0.72	0.10	0.08	0.52	0.28
7	0.81	0.52	0.13	0.07	0.38	0.26
8	0.84	0.46	0.12	0.05	0.37	0.24
9	0.78	0.46	0.13	0.05	0.31	0.23
10	0.75	0.41	0.12	0.05	0.30	0.21

complexity of IB algorithm, $O(|A|^3|C|)$, means that the two-fold increase in the number of sessions causes more than the eight-fold increase in the computation time. For example, in the case of 1,746 sessions, this time is about 1 hour and 25 minutes for each sample.

Comparison of the efficiency of IB and k-means

After evaluating the ability of our Information Bottleneck approach to differentiate between user sessions of both classes, similar experiments were performed for the reference clustering algorithm, *k*-means, and the resulting entropy values achieved for both methods were compared. Table 2 juxtaposes minimum, maximum, and average entropy values for IB and *k*-means for the number of clusters ranging from 2 to 10. Figure 17 shows the improvement in the separation of buyers and non-buyers with the increase in the number of clusters and illustrates the supremacy of our approach over *k*-means.

One can observe that increasing the number of clusters generally leads to better results for *k*-means, like in the case of IB. The average efficiency of IB is significantly higher than that of *k*-means for all scenarios. This superiority is especially clear for lower number of clusters. For two clusters, *k*-means could hardly separate sessions of buyers and non-buyers (the average entropy of 0.97, compared to the IB average entropy of 0.57). In all experiments with two clusters, *k*-means generated one very small cluster containing a majority of buying sessions and one huge mixed cluster with sessions of both classes.

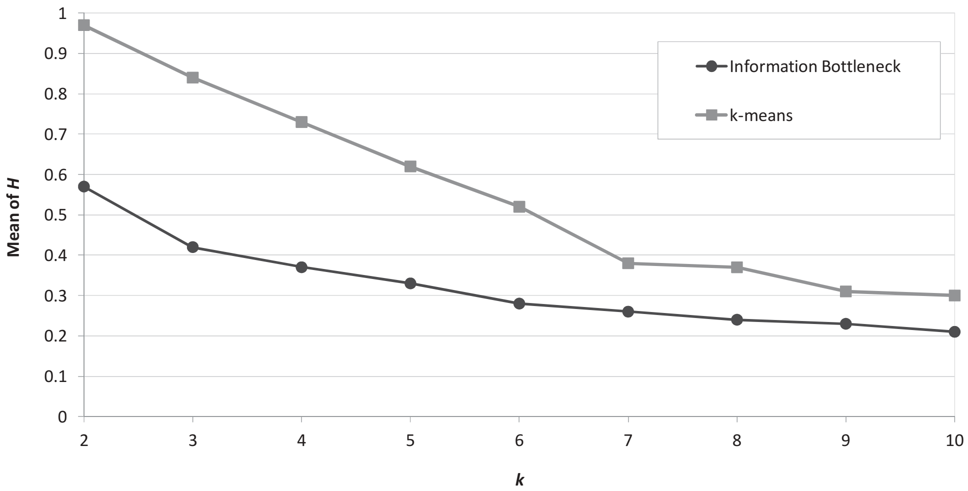

Figure 17. The comparison of average entropy for IB and *k*-means depending on the number of clusters.

The higher the number of clusters, the lower degree of superiority of IB over *k*-means, but even for the ten-cluster scenario, the difference in the average entropy is significant: 0.21 for IB compared to 0.30 for *k*-means.

Concluding remarks

In this paper, we presented our novel framework, based on agglomerative Information Bottleneck, to deal with the problem of discovering various user profiles in a Web store. We analyzed click-stream data from a real e-commerce site to derive and investigate behavioral patterns of buyers and non-buyers. Our main contribution is adaptation of the agglomerative Information Bottleneck algorithm to cluster e-customer sessions based on the observed click-steam user behavior, as well as implementation of a case study of clustering e-customer sessions reconstructed from server log data for a real online bookstore.

The results of our study show the high efficiency of the our approach, which makes it possible not only to discover various profiles of e-customers but also investigate hidden knowledge about specific characteristics of the identified profiles. The comparison of clustering results achieved for Information Bottleneck and the most popular clustering technique, *k*-means, showed a clear superiority of IB in terms of the ability to differentiate between various profiles of buyers and non-buyers.

The approach introduced here has significant practical potential. It can be used to develop tools supporting personalized marketing techniques, for example, a CF recommendation system, targeted mailing, or a system of CF-based discount coupons, stored in the retailers' database and integrated with the e-commerce module. The results may also be useful for the customization of real-time services, like live chat, in which users identified as less experienced customers should be served first. Another possible implementation of the approach might be an intelligent approach to classify new sessions to user profiles identified for a given website to provide users with tailored service. We leave these issues to our future work.

References

Abbot, D. 2014. *Applied predictive analytics. Principles and techniques for the professional data analyst*. Indianapolis, IN: Wiley.

Adnan, M., M. Nagi, K. Kianmehr, R. Tahboub, M. Ridley, and J. Rokne. 2011. Promoting where, when and what? An analysis of Web logs by integrating data mining and social network techniques to guide ecommerce business promotions. *Social Network Analysis and Mining* 1 (3):173–85. doi:10.1007/s13278-016-0375-4.

Adomavicius, G., and A. Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* 17 (6):734–49. doi:10.1109/TKDE.2005.99.

Carmona, C. J., S. Ramírez-Gallego, F. Torres, E. Bernal, M. J. Del Jesus, and S. García. 2012. Web usage mining to improve the design of an e-commerce website: OrOliveSur.com. *Expert Systems with Applications* 39 (12):11243–49. doi:10.1016/j.eswa.2012.03.046.

Chang, H.-J., L.-P. Hung, and C.-L. Ho. 2007. An anticipation model of potential customers' purchasing behavior based on clustering analysis and association rules analysis. *Expert Systems with Applications* 32 (3):753–64. doi:10.1016/j.eswa.2006.01.049.

Chen, Z., A. W.-C. Fu, and F. C.-H. Tong. 2003. Optimal algorithms for finding user access sessions from very large Web logs. *World Wide Web* 6:259–79. doi:10.1023/A:1024606901978.

Chen, Z.-Y., Z.-P. Fan, and M. Sun. 2012. A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *European Journal of Operational Research* 223 (2):461–72. doi:10.1016/j.ejor.2012.06.040.

Cheng, C.-H., and Y.-S. Chen. 2009. Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications* 36 (3),Part I):4176–84. doi:10.1016/j.eswa.2008.04.003.

Cho, Y. H., and J. K. Kim. 2004. Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce. *Expert Systems with Applications* 26 (2):233–46. doi:10.1016/S0957-4174(03)00138-6.

Chou, P.-H., P.-H. Li, -K.-K. Chen, and M.-J. Wu. 2010. Integrating web mining and neural network for personalized e-commerce automatic service. *Expert Systems with Applications* 37 (4):2898–910. doi:10.1016/j.eswa.2009.09.047.

Du, W., and S. Tan. 2009. An iterative reinforcement approach for fine-grained opinion mining. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'09)*, ed. U. Germann, C. Shah, S. Stoyanchev, C. P. Rosé, and A. Sarkar, 486–93. Boulder, CO: Association for Computational Linguistics.

Ganu, G., Y. Kakodkar, and A. Marian. 2013. Improving the quality of predictions using textual information in online user reviews. *Information Systems* 38 (1):1–15. doi:10.1016/j.is.2012.03.001.

Gordon, S., H. Greenspan, and J. Goldberger. 2003. Applying the Information Bottleneck principle to unsupervised clustering of discrete and continuous image representations. In Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV'03), 370–77. New York, NY: IEEE.

Hecht, R. M., and N. Tishby. 2005. Extraction of relevant speech features using the Information Bottleneck method. Paper presented at the 9th European Conference on Speech Communication and Technology (INTERSPEECH'05 - Eurospeech), Lisbon, Portugal, September 4–8.

Hop, W. 2013. Web-shop order prediction using machine learning. MsC. thesis, Erasmus University Rotterdam.

Huiying, Z., and L. Wei. 2004. An intelligent algorithm of data pre-processing in Web usage mining. In *Proceedings of the 5th World Congress on Intelligent Control and Automation (WCICA'04)*, ed. Z. Da Xue, vol. 4, 3119–23. New York, NY: IEEE. doi:10.1109/WCICA.2004.1343095.

Ikeda, M., M. Yoshida, S. Ono, H. Nakagawa, and I. Sato. 2009. Person name disambiguation on the Web by two-stage clustering. Paper presented at the 18th International Conference on World Wide Web (WWW'09), Madrid, Spain, April 20–24.

Jenamani, M., P. K. J. Mohapatra, and S. Ghose. 2003. A stochastic model of e-customer behavior. *Electronic Commerce Research and Applications* 2 (1):81–94. doi:10.1016/S1567-4223(03)00010-3.

Jiang, X.-M., W.-G. Song, and W.-G. Feng. 2006. Optimizing collaborative filtering by interpolating the individual and group behaviors. In *Proceedings of the 8th Asia-Pacific Web Conference (APWeb'06), LNCS 3841*, ed. X. Zhou, J. Li, H. T. Shen, M. Kitsuregawa, and Y. Zhang, 568–78. Berlin Heidelberg, Germany: Springer-Verlag.

Joshi, A., K. Joshi, and R. Krishnapuram. 2000. On Mining Web access logs. Paper presented at the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'00), Dallas, TX, May 14.

Kumar, R., and P. K. Bala. 2017. Recommendation engine based on derived wisdom for more similar item neighbors. *Information Systems and e-Business Management* 15 (3):661–87. doi:10.1007/s10257-016-0322-y.

Kwan, I. S. Y., J. Fong, and H. K. Wong. 2005. An e-customer behavior model with online analytical mining for internet marketing planning. *Decision Support Systems* 41 (1):189–204. doi:10.1016/j.dss.2004.11.012.

Ling, X., G.-R. Xue, W. Dai, Y. Jiang, Q. Yang, and Y. Yu. 2008. Can chinese Web pages be classified with English data source? Paper presented at the 17th International Conference on World Wide Web (WWW'08), Beijing, China April 21–25.

Menascé, D. A., V. A. F. Almeida, R. Fonseca, and M. A. Mendes. 1999. A methodology for workload characterization of e-commerce sites. Paper presented at the 1st ACM conference on Electronic Commerce, Denver, CO, November 3–5. doi: 10.1145/336992.337024.

Mohammadnezhad, M., and M. Mahdavi. 2012. Providing a model for predicting tour sale in mobile e-tourism recommender systems. *International Journal of Information Technology Convergence and Service* 2 (1):1–8. doi:10.5121/ijitcs.2012.2101.

Nenava, S., and V. Choudhary. 2013. Hybrid personalized recommendation approach for improving mobile e-commerce. *International Journal of Computer Science & Engineering Technology* 4 (5):546–52.

Ofran, Y., and H. Margalit. 2006. Proteins of the same fold and unrelated sequences have similar amino acid composition. *Proteins* 64:275–79. doi:10.1002/prot.20964.

Panda, M., and M. R. Patra. 2009. A novel classification via clustering method for anomaly based network intrusion detection system. *International Journal of Recent Trends in Engineering* 2 (1):1–6.

Pereira, F. C., N. Tishby, and L. Lee. 1993.Distributional clustering of English words. Paper presented at the 31st Annual Meeting on Association for Computational Linguistics (ACL '93), Columbus, OH, June 22–26. doi: 10.3115/981574.981598.

Poggi, N., T. Moreno, J. L. Berral, R. Gavaldà, and J. Torres. 2007. Web customer modeling for automated session prioritization on high traffic sites. In *Proceedings of the International Conference on User Modeling (UM'07)*, ed. C. Conati, K. McCoy, and G. Paliouras, LNCS 4511, 450–54. Berlin Heidelberg, Germany: Springer-Verlag. doi:10.1007/978-3-540-73078-1_63.

Schneidman, E., N. Slonim, N. Tishby, R. R. de Ruyter Van Steveninck, and W. Bialek. 2002. Analyzing neural codes using the Information Bottleneck method. Technical report, The Hebrew University, Jerusalem, Israel.

Shim, B., K. Choi, and Y. Suh. 2012. CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns. *Expert Systems with Applications* 39 (9):7736–42. doi:10.1016/j.eswa.2012.01.080.

Slonim, N. 2002. The Information Bottleneck: Theory and Applications. PhD Thesis, Hebrew University, Jerusalem, Israel. http://www.cs.huji.ac.il/labs/learning/Theses/Slonim_PhD.pdf.

Slonim, N., N. Friedman, and N. Tishby. 2002. Unsupervised document classification using sequential information maximization. Paper presented at the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02), Tampere, Finland, August 11–15 doi: 10.1145/564376.564401.

Slonim, N., R. Somerville, N. Tishby, and O. Lahav. 2001. Objective classification of galaxy spectra using the Information Bottleneck method. *Monthly Notices of the Royal Astronomical Society* 323 (2):270–84. doi:10.1046/j.1365-8711.2001.04125.x.

Slonim, N., and N. Tishby. 1999. Agglomerative information Bottleneck. In *Advances in Neural Information Processing Systems 12*, ed. S. A. Solla, T. K. Leen, and K. Müller, 617–23. Cambridge, MA: MIT Press.

Slonim, N., and N. Tishby. 2000. Document clustering using word clusters via the Information Bottleneck method. Proceedings of SIGIR'00, 208–15. New York, NY: ACM.

Soiraya, B., A. Mingkhwan, and C. Haruechaiyasak. 2008. E-commerce Web site trust assessment based on text analysis. *International Journal of Business and Information* 3 (1):86–114.

Soliman, T. H. A., M. A. Elmasry, A. R. Hedar, and M. M. Doss. 2012. Utilizing support vector machines in mining online customer reviews. Paper presented at the 22nd International Conference on Computer Theory and Applications (ICCTA'12), Alexandria, Egypt, October 13–15. doi: 10.1109/ICCTA.2012.6523568.

Song, Q., and M. Shepperd. 2006. Mining Web browsing patterns for e-commerce. *Computers in Industry* 57 (7):622–30. doi:10.1016/j.compind.2005.11.006.

Su, Q., and L. Chen. 2015. A method for discovering clusters of e-commerce interest patterns using click-stream data. *Electronic Commerce Research and Applications* 14 (1):1–13. doi:10.1016/j.elerap.2014.10.002.

Suchacka, G. 2014. Analysis of aggregated bot and human traffic on e-commerce site. In *Annals of Computer Science and Information Systems* (*ACSIS*)*, Vol. 2: Proceedings of the Federated Conference on Computer Science and Information Systems* (*FedCSIS'14*), ed. M. Ganzha, L. Maciaszek, and M. Paprzycki, 1123–30. New York, NY: IEEE. doi: 10.15439/2014F346.

Suchacka, G., and L. Borzemski. 2013. Web server support for e-customer loyalty through QoS differentiation. *Transactions on Computational Collective Intelligence* (XII):89–107. doi:10.1007/978-3-642-53878-0_5.

Suchacka, G., and G. Chodak. 2016. Using association rules to assess purchase probability in online stores. *Information Systems and e-Business Management* 15 (3):751–80. doi:10.1007/s10257-016-0329-4.

Suchacka, G., M. Skolimowska-Kulig, and A. Potempa. 2015a. A k-Nearest Neighbors method for classifying user sessions in e-commerce scenario. *Journal of Telecommunications and Information Technology* 3:64–69.

Suchacka, G., M. Skolimowska-Kulig, and A. Potempa. 2015b. Classification of e-customer sessions based on Support Vector Machine. In *Proceedings of the 29th European Conference on Modelling and Simulation* (*ECMS'15*), ed. V. M. Mladenov, G. Spasov, P. Georgieva, and G. Petrova, 594–600. European Council for Modelling and Simulation. doi: 10.7148/2015-0594.

Suchacka, G., and S. Stemplewski. 2017. Application of neural network to predict purchases in online store. In *Advances in Intelligent Systems and Computing (AISC) 524, ISAT'16 - Part IV*, ed. Z. Wilimowska, L. Borzemski, A. Grzech, and J. Świątek, 221–33. Cham, Switzerland: Springer. doi:10.1007/978-3-319-46592-0_19.

Tan, P.-N., M. Steinbach, and V. Kumar. 2006. *Introduction to data mining*. Boston, MA: Pearson Addison-Wesley.

Tishby, N., F. C. Pereira, and W. Bialek. 1999. The Information Bottleneck method. Paper presented at the 37th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, September 22–24.

Wang, Y. Y., A. Luse, A. M. Townsend, and B. E. Mennecke. 2015. Understanding the moderating roles of types of recommender systems and products on customer behavioral intention to use recommender systems. *Information Systems and e-Business Management* 13 (4):769–99. doi:10.1007/s10257-014-0269-9.

Zhang, X., J. Edwards, and J. Harding. 2007. Personalised online sales using Web usage data mining. *Computers in Industry* 58 (8–9):772–82. doi:10.1016/j.compind.2007.02.004.

Zhou, X., J. Wei, and C.-Z. Xu. 2006. Resource allocation for session-based two-dimensional service differentiation on e-commerce servers. *IEEE Transactions on Parallel and Distributed Systems* 17 (8):838–50. doi:10.1109/TPDS.2006.111.

Ziv, E., M. Middendorf, and C. H. Wiggins. 2005. An information-theoretic approach to network modularity. *Physical Review E* 71:046117. doi:10.1103/PhysRevE.71.046117.

Notes on contributors

Jacek Iwanski is an assistant professor and the head of the IT section in the Institute of Mathematics and Informatics at the University of Opole, Poland. He received the M.Sc. degree in Physics from the University of Wroclaw, Poland. In 1994 he received the Ph.D. degree in Physics from the University of Hasselt, Belgium. His research interests include artificial intelligence and machine learning methods in real-life applications, embedded systems programming and construction, and sensor networks.

Grażyna Suchacka is an assistant professor in the Institute of Mathematics and Informatics at the University of Opole, Poland. She received the M.Sc. degrees in Computer Science and in Management from Wroclaw University of Science and Technology, Poland. In 2011 she received the Ph.D. degree in Computer Science with distinction from Wroclaw University of Science and Technology. Dr. Suchacka's research interests include analysis and modeling of Web traffic, Web mining, and Quality of Web Service with special regard to electronic commerce support and Web bot recognition.

Grzegorz Chodak is an associate professor at the Department of Operations Research, Finance and Applications of Computer Science at the Wroclaw University of Science and Technology, Poland. He is an author and co-author of over 70 scientific publications, mainly in the field of electronic commerce, logistics, data mining, and social media. He specializes in issues related to online stores and the publishing market. He also has a practical experience in electronic commerce.