Practical Aspects of Log File Analysis for E-Commerce

Grażyna Suchacka[1] and Grzegorz Chodak[2]

[1] Institute of Mathematics and Informatics, Opole University,
Opole, Poland
[2] Institute of Organisation and Management, Wroclaw University of Technology,
Wroclaw, Poland
gsuchacka@uni.opole.pl, grzegorz.chodak@pwr.wroc.pl

Abstract. The paper concerns Web server log file analysis to discover knowledge useful for online retailers. Data for one month of the online bookstore operation was analyzed with respect to the probability of making a purchase by e-customers. Key states and characteristics of user sessions were distinguished and their relations to the session state connected with purchase confirmation were analyzed. Results allow identification of factors increasing the probability of making a purchase in a given Web store and thus, determination of user sessions which are more valuable in terms of e-business profitability. Such results may be then applied in practice, e.g. in a method for personalized or prioritized service in the Web server system.

Keywords: web server, log file analysis, statistical inference, e-commerce, B2C, Business-to-Consumer, web store.

1 Introduction

The analysis of historical data recorded in Web server log files is the basic way of capturing knowledge of the Web server workload and the behavior of Web users. In e-commerce environment such analyses have been performed at multiple levels, including the lowest, protocol level (corresponding to HTTP requests), the application level (corresponding to Web page requests or business-related Web interactions), and the user level (corresponding to user sessions) [1–3]. From the online retailers' point of view the application and user level analyses are of the highest practical value because understanding the way in which customers use the site and navigate through the store, especially in the context of successful purchase transactions, may lead to better organization of e-commerce service and more efficient business decisions.

This was our motivation for exploring dependencies between different characteristics of user visits to the site of a Web store and their probability of being ended with a purchase. Three groups of visits differing in user behavior were considered: visits of anonymous users, visits of users who logged on but did not buy anything, and visits of users who decided to make a purchase.

A. Kwiecień, P. Gaj, and P. Stera (Eds.): CN 2013, CCIS 370, pp. 562–572, 2013.

The rest of the paper is organized as follows. Section 2 overviews related work on analyses of historical data for Web stores. Section 3 discusses some practical aspects and possibilities of data analysis for different user groups in the Web store. Section 4 presents our research methodology and Sect. 5 discusses results of the research. Key findings are summarized in Sect. 6.

2 Related Work

An important aspect of e-commerce workload characterization is a click-stream analysis, which concerns discovering user navigational patterns at a Web site [1]. The click-stream analysis is often combined with segmentation or clustering methods to determine different customer profiles [4, 2, 5, 3]. Discovery of meaningful usage patterns characterizing the browsing behavior of Internet users realized by applying data mining techniques is called Web usage mining.

Segmentation methods have been especially intensively explored for Web store historical data. Some techniques used in CRM (*Customer Relationship Management*) in traditional commerce have been also applied in e-commerce. For example, RFM (*Recency, Frequency, Monetary*) analysis was combined with such data mining techniques as Apriori algorithm [6], rough sets and K-means algorithm [7], and approximate reasoning from the set of fuzzy rules [8]. The vector quantization-based clustering and the "Apriori"-based association rule mining algorithm was applied to classify e-customers based on their RFM values and to find out relationships among their purchases [9]. A design of the Web mining algorithm based on variable precision rough sets to classification tasks was proposed in [10].

As a consequence, some models of a user session in a Web store have been proposed as probabilistic finite state machines for different user profiles. Examples include Customer Behavior Model Graph [11], TPC-W session model [12], and Extended Finite State Machine [13].

Association rules have also been successfully applied to e-commerce, combined with collaborative filtering methods [14, 15]. Association rules for products viewed or purchased by different customers are widely used in contemporary Web stores in product recommendation. In this paper, we focus on applying statistical analysis and association rules to identify these sessions attributes and characteristics which increase the probability of making a purchase.

3 Key Questions in the Context of Different Web Store User Groups

Although most online retailers today use analytical applications (e.g. Google Analytics), which provide much valuable information, they are not able to answer many questions concerning a detailed behavior of three user groups: users, who do not log on, users who log on but do not buy anything, and users who log on and finalize purchase transactions.

1. **Anonymous users.**
   Questions concerning the behavior of unlogged users aim at finding reasons for which the users – potential customers – do not log on and do not place an order. This group is the most numerous. It contains many users who visit only one page in the session. Some information on these users is available in Google Analytics, e.g. their share in overall traffic (called the bounce rate) and sources of their visits. Since these users do not stay in the Web store, no information on their behavior can be read from log files. Therefore, such sessions were not under consideration in this paper.
   However, some anonymous users navigate through the site for some time and it gives the chance to analyze their usage patterns: sequences of visited pages, usage of specific tools, visited featured places of the Web store (new products, bestsellers, promotions). The analysis should aim at finding reasons for which the user do not decide to log on and make a purchase.
2. **Users who log on but do not buy anything.**
   The analysis of these users' behavior should be more detailed because they put some effort into the interaction, which shows they are interested in the offered products, but they did not decide to purchase for unknown reasons. First of all, one should analyze the last pages statistics in their sessions, including the following page types: a product page, a product category page, a product review page, the shopping cart page, the page informing about shipping charges and conditions, the page containing the store regulations, the contact page, the home page, or a page containing search results. One aspect of possible practical analysis concerns the percentage of logged users who added products to the carts but did not buy them. Other aspect is connected with the usage of the search engine at the store site and the analysis of search results: maybe users did not find products satisfying their searching criteria, or maybe they got some results which were not satisfactory for them.
3. **Users who log on and finalize purchase transactions.**
   These analyses are the most valuable for the online retailer. Customers may be classified in terms of their navigational patterns, viewed and/or purchased products, value of the shopping cart, source of the customer visit, attributes captured in the registration process (age, sex, geographical location, etc.). Not all that information may be read from logs – some of them require installing additional software able to intercept data from registration forms (like Google Analytics does).

Some questions which can be answered based on log file analysis are the following. Is there a relationship between the number or value of purchased products and the source of the visit? Is there a relationship between the time users spend in the Web store and the percentage of customers making a purchase? Is there a relationship between the number of viewed products and the percentage of customers making a purchase? What is the mean time users spend on individual pages of the checkout process? What percentage of customers, who had checked information about shipping costs before starting the checkout process, decide to confirm the purchase?

The comparison of the aforementioned user groups may provide information on significant differences between them. It is especially important to identify features distinguishing customers in the third group and to determine factors increasing the probability of making a purchase.

To illustrate the practical aspect of such research we performed analyses in order to answer the following questions:

1. How many logged users confirm a purchase and how many users give up after reading shipping charges in the checkout process? How many users from these both groups checked shipping charges before the checkout process?
2. How many users, who had checked information on shipping charges before the checkout process, decided to make a purchase?
3. How many (logged and unlogged) users, who added products to the shopping carts, give up purchasing them?

4 Research Methodology

The analysis was performed for one month data in Web bookstore log files, recorded in December 2011. The Web server was an Apache HTTP server on Linux with PHP and MySQL support. Using a C++ computer program, data was read from log files, preprocessed, and cleaned; user sessions were reconstructed, key session states were distinguished, and data was analyzed by means of statistical analysis and association rules.

4.1 Data Preprocessing

First, raw data were read from Web server log files and for each HTTP request the following data were distinguished: the IP address of the Web client, the identifier of the Web client, the user identifier, the timestamp, the HTTP method, the URI of the resource requested, the version of HTTP protocol, the HTTP status code, the size of the object sent to the client, the address of the referring page, and the client browser information. Some data was then transformed, e.g. the timestamps had to be converted to integer values so that the comparison of request interarrival times could be performed.

4.2 Data Cleaning

First stage of data cleaning was connected with elimination of useless data. Since our analysis concerns the behavior of users and involves a click-stream analysis, the following requests have been excluded from analysis: hits for embedded objects (e.g. images), automatically generated by Web client browsers, requests generated by Web bots (e.g. Web crawlers), and requests connected with administrative tasks.

The original data set contained 1 600 964 HTTP requests. 1 540 812 requests were excluded from analysis and only 60 152 requests (3.76 %) were left, as directly resulting from users' clicks.

4.3 Reconstruction of User Sessions

To analyze user behavior, Web users were identified based on IP addresses combined with the client browser information, included in each request. Then user sessions were reconstructed taking into consideration request arrival times. A *user session* means a sequence of Web page requests issued by the user during their single visit to the site. We assumed that the intervals between Web page requests in a session does not exceed 30 minutes.

16 081 user sessions were identified. For each session three basic characteristics were determined:

- the *session length*, i.e. the number of pages requested in the session,
- the *session duration*, i.e. the time interval between arrivals of the first and the last Web page requests in the session (with one second accuracy),
- the *mean time per page*, i.e. the average time of browsing a page by user in the session, calculated as the session duration divided by the session length.

The determined session duration is shorter than the actual time of the user-Web site interaction, because it does not include the time of browsing the last page in session, which cannot be determined at the server side. Therefore, the *session duration* and the *mean time per page* could not be determined for sessions containing only one page.

4.4 Elimination of Outlier Sessions

Session lengths and durations were used to find and exclude outlier sessions, diverging from a general trend for the sessions. In our case, such sessions may represent incorrect entries in log files or may suggest robot-generated interactions, which have not been identified. Using a graphical method [16], four sessions were eliminated. Thus, 16 077 user sessions were analyzed.

4.5 Identification of Key Session States

A user visits multiple Web pages in a session and performs multiple Web interactions, such as searching products according to given keywords, browsing detailed information on a selected product, etc. Pages may be grouped by functionality into different session states. The most desirable state is connected with making a purchase, i.e. with purchase confirmation:

- *Checkout_ success* state occurs when a user successfully realizes the checkout process, accepting total transaction cost, giving the shipping address, and finally confirming the purchase.

Taking into account goals of our analysis, we identified four other key session states, which may affect the probability of making a purchase by a user:

- *Login* means user's registration or logging into the site. Users may browse the site content without being logged on; logging into the site is necessary if they want to finalize a purchase transaction. However, not every user who is logged on will decide to buy something finally.

- *Shopping_ cart* means adding a product to the virtual shopping cart. Adding at least one item to the cart is a prerequisite for purchase transaction; however, products in cart do not guarantee their purchase. In fact, shopping cart abandonment is a huge problem for many online retailers [17].
- *Shipping_ info* means the situation when a user checks the information on shipping charges and conditions before starting the checkout process. Distinguishing this state is important because it may affect a purchase transaction: a user who is aware of shipping charges, may take them into consideration while adding products to cart and thus they will not be surprised by the total transaction cost at checkout.
- *Shipping_ info_ at_ checkout* is an integral stage of the purchase transaction and it occurs when a user is informed about possible means and charges of shipping during the checkout process. The shipping cost is automatically added to product costs.

Our analysis revealed that the prevailing majority of Web interactions in the Web store is connected with browsing and searching operations, whereas very few users perform Web interactions connected with key session states. For all 16 077 user sessions, visits to key session states have been rare (Fig. 1).

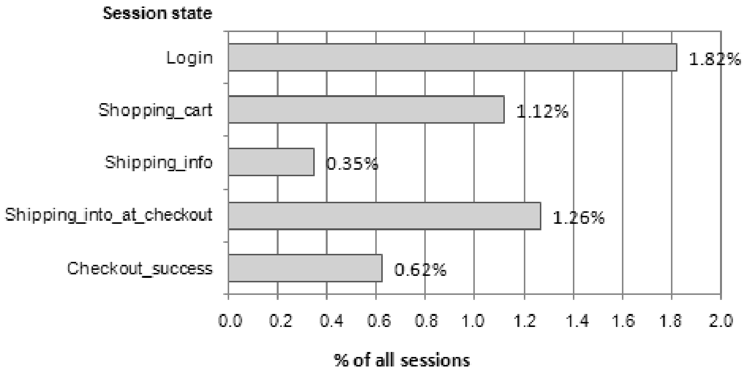

Fig. 1. Percentages of sessions with visits to key session states

As it can be seen, only 1.12 % of users added some products to the shopping carts and 0.62 % of users decided to buy them. The checkout process was started by at least 1.26 % of users (it can be inferred from the percentage of visits to the session state *Shipping_ info_ at_ checkout*, which is a part of this process), however, more than half of them gave up confirming the purchase.

It seems surprising that although only 1.12 % of users added products to the carts, as much as 1.26 % started the checkout process (because having products in carts is essential for starting this process). The explanation is the fact that products added to the user's cart during one visit are remembered till the next user's visit (unless a cookie mechanism is blocked in the user's Internet browser). Thus, some users had to add products to the carts during their former visits.

5 Analysis of Dependencies between Session States

The next step of analysis concerned discovery of relationships between the key session states in order to answer three questions formulated in Sect. 3. Main results are presented in Fig. 2.

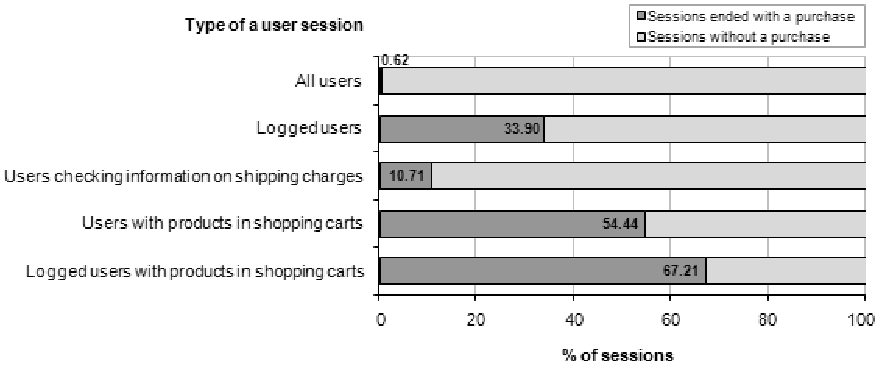

Fig. 2. Percentages of sessions ended with a purchase depending on a session type

- Question 1: How many logged users confirm a purchase and how many users give up after reading shipping charges in the checkout process? How many users from these both groups checked shipping charges before the checkout process?
  From among all users entering the site only 292 logged on. Only 99 logged users (33.9 %) decided to confirm the purchase: 6 of them (6.06 %) had checked the shipping information before starting the checkout process; 93 of them (93.94 %) had not checked the shipping information before.
  These results suggest that when a user logs on, the probability of making a purchase by them increases: in the set of all users only 0.62 % confirmed the purchase, whereas in the set of logged users as much as 33.9 % confirmed the purchase, which is more than one third.
  Moreover, 103 logged users (i.e. 35.52 % of all logged users) started the checkout process (they visited state *Shipping_ info_ at_ checkout*) but did not finish it: 26 of them (25.24 %) had checked the shipping information before starting the checkout process; 77 of them (74.76 %) had not checked the shipping information before, so maybe they were surprised by the additional shipping cost at checkout.
  However, only three from among these 103 users finished their sessions just after checking shipping information at checkout.

- Question 2: How many users, who had checked information on shipping charges before the checkout process, decided to make a purchase?
  56 users read the shipping information before starting the checkout process. From among them only six (10.71 %) decided on a purchase. Thus, checking the shipping information during the navigation through the site before starting the checkout process increase the probability of making a purchase from 0.62 % to 10.71 %.
- Question 3: How many (logged and unlogged) users, who added products to the shopping carts, give up purchasing them?
  Products were added to carts by 180 users; 98 of them (54.44 %) did not decide on the purchase: 122 logged users added products to the carts and 40 of them (32.79 %) did not buy the products; 58 unlogged users added products to the carts and 58 of them (100%) did not buy the products.
  It can be noticed that products have been added to carts by much more logged users than the unlogged ones. Moreover, much more logged users decided to buy the products. One can infer that for sessions with not empty carts the fact of user's logging into the site significantly increases the probability of purchase.

To formally describe relationships between the key session states, we applied association rules. We implemented the a priori algorithm using the methodology and notation used in [18].

We determined eight characteristics of a user session which may matter in the context of the purchase probability. The first group of characteristics include visits to four key session states, defined above. The second group of characteristics was determined based on the *session length*, the *session duration*, and the *mean time per page*. Based on results of analyses comparing characteristics of all sessions and the sessions ended with a purchase (which were not discussed in this paper due to limited place), we replaced numeric values of the session characteristics with qualitative values in the following way:

$$\textit{session length} = \begin{cases} \text{short,} & \text{if the number of requests } = 1 \\ \text{medium,} & \text{if the number of requests } \in [2, 13] \\ \text{long,} & \text{if the number of requests } \geq 14 \end{cases} \quad (1)$$

$$\textit{session duration} = \begin{cases} \text{short,} & \text{if the session duration } < 60\,\text{s} \\ \text{medium,} & \text{if the session duration } \in [60, 960]\,\text{s} \\ \text{long,} & \text{if the session duration } > 960\,\text{s} \end{cases} \quad (2)$$

$$\textit{mean time per page} = \begin{cases} \text{short,} & \text{if mean time per page } < 60\,\text{s} \\ \text{medium,} & \text{if mean time per page } \in [60, 180]\,\text{s} \\ \text{long,} & \text{if mean time per page } > 180\,\text{s} \end{cases} . \quad (3)$$

We define a set I containing key session characteristics: I = {"*Login* = true", "*Shopping_ cart* = true", "*Shipping_ info* = true", "*Checkout_ success* = true", "*session length* = medium", "*session length* = long", "*session duration* = medium", "*session duration* = long", "*mean time per page* = short", "*mean time per page*

= medium", "*mean time per page* = long"}. Characteristics "*session length* = short" and "*session duration* = short" were excluded from the set because it is impossible to make a purchase in a session containing only one page request or lasts less than one minute.

Let D denote a set of all user sessions. Each session in D is represented as a set of session characteristics from I. For example, if a user logged on, searched for a few books, added two books to the shopping cart, and then finished the session; the session consisted of 12 page requests and was lasting 10 minutes (which gives the mean time per page equal to 54.5 s), the session will be represented as the set {"*Login* = true", "*Shopping_ cart* = true", "*session length* = medium", "*session duration* = medium", "*mean time per page* = short"}.

An association rule is the implication $A \Rightarrow B$ (*if A then B*), where the antecedent $A \subseteq I$, the consequent $B \subseteq I$, and $A \cap B = \varnothing$. Moreover, each association rule is described with two measures: *support* and *confidence* [18].

The *support* for a given association rule $A \Rightarrow B$ is the percentage of sessions in D which contain both A and B:

$$support = P(A \cap B) = \frac{number\ of\ sessions\ containing\ both\ A\ and\ B}{number\ of\ all\ sessions}\ . \qquad (4)$$

The *confidence* for a given association rule $A \Rightarrow B$ is the percentage of sessions containing A which also contain B:

$$confidence = P(B \mid A) = \frac{number\ of\ sessions\ containing\ both\ A\ and\ B}{number\ of\ sessions\ containing\ A}\ . \qquad (5)$$

A rule is *strong* if it meets or surpasses certain minimum support and confidence criteria. Since we consider associations between session states in the context of making a purchase and very small percentage of all sessions (0.62 %) ends with a purchase, we assumed the minimum support of 0.5 % and the minimum confidence of 70 %. Rules resulting from the application of a priori algorithm for such conditions are presented in Table 1.

The support of the first rule, R_1, is 0.5 %, meaning that the rule applies to 0.5 % of sessions in the analyzed dataset. It is relatively low level of support, however we have to keep in mind that only 0.62 % of all sessions ended with a purchase. 100 sessions ended with a purchase and of these sessions, 80 fulfilled the antecedent condition, i.e. the user logged on, added some products to cart (what is obvious as these events occur for each purchase transaction), the user opened at least 14 pages, and the mean time per page did not exceeded one minute.

The confidence of the rule R_1 equal to 73 % means that of all 16 077 sessions some group of sessions (0.68 %, i.e. 109 sessions) fulfilled the antecedent condition; 73 % of these 109 sessions (i.e. 80 sessions) ended with a purchase. Taking into account that for all sessions in data set the probability of purchase is only 0.62 %, one can say that the rule has very high level of confidence.

Table 1. Strong association rules determined for the analyzed data set

Rule	Antecedent	Consequent	Support [%]	Confidence [%]
R_1	{“*Login* = true”, “*Shopping_cart* = true”, “*session length* = long”, “*mean time per page* = short”}	“*Checkout_success* = true”	0.5	73
R_2	{“*Login* = true”, “*Shopping_cart* = true”, “*session length* = long”}	“*Checkout_success* = true”	0.5	72
R_3	{“*Login* = true”, “*Shopping_cart* = true”, “*mean time per page* = short”}	“*Checkout_success* = true”	0.5	70

6 Conclusions

Motivated by the need of precise characterization of e-customer behavior, we analyzed data for a Web bookstore applying statistical analysis and association rules. The main goal of the analysis was identification of these sessions characteristics which increase the probability of making a purchase and thus, determination of user sessions which are more valuable in terms of e-business profitability.

Results show that only 0.6 % of user sessions ends with a purchase. When a user checks the shipping information, the probability of a purchase increases to almost 11 %, whereas the fact of the user’s logging on increases this probability to 34 %. It can be observed that logged users are much more likely to add products to carts than unlogged users. Unlogged users usually do not decide to buy the products they have in carts, whereas from among logged users with not empty shopping carts almost two thirds finalize purchase transactions.

Application of association rules confirmed that factors increasing the probability of making a purchase include user’s logging on, adding a product to the shopping cart, the mean time per page not exceeding one minute, and at least 14 page requests in the session. One can formulate some relationships, e.g. a logged user with not empty shopping cart, spending on average less than one minute per page, who visited at least 14 pages in the session, will decide to confirm the purchase with probability of 73 %.

Though our results are promising, the analysed data set was rather small so the correctness of our approach should be verified for a bigger data set. The findings may be used in practice, e.g. in a method for personalized user service in the Web store, encouraging “likely buyers” to finalize purchase transactions, or in a method for prioritized service in the Web server system, aiming to offer the best quality of service to probable buyers.

References

1. Kurz, C., Haring, G.: E-Business Benchmarking Based on Hierarchical Customer Behavior Characterization. In: 5th ICECR (2002)
2. Menascé, D.A., Almeida, V.A.F., Riedi, R., Ribeiro, F., Fonseca, R., Meira Jr., W.: A Hierarchical and Multiscale Approach to Analyze E-Business Workloads. Perform. Eval. 54(1), 33–57 (2003)
3. Wang, Q., Makaroff, D.J., Edwards, H.K.: Characterizing Customer Groups for an E-commerce Website. In: 5th ACM EC, pp. 218–227. ACM Press, New York (2004)
4. Joshi, A., Joshi, K., Krishnapuram, R.: On Mining Web Access Logs. In: ACM SIGMOD Workshop DMKD, pp. 63–69 (2000)
5. Song, Q., Shepperd, M.: Mining Web Browsing Patterns for E-commerce. Comput. Ind. 57(7), 622–630 (2006)
6. Chen, Y.-L., Kuo, M.-H., Wu, S.-Y., Tang, K.: Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. Electron. Commer. R. A. 8(5), 241–251 (2009)
7. Cheng, C.H., Chen, Y.S.: Classifying the segmentation of customer value via RFM model and RS theory. Expert Syst. Appl. 36(3), 4176–4184 (2009)
8. Chan, C.-C.H., Cheng, C.-B., Hsu, C.-H.: Bargaining strategy formulation with CRM for an e-commerce agent. Electron. Commer. R. A. 6(4), 490–498 (2007)
9. Tanna, P., Ghodasara, Y.: Exploring the Pattern of Customer Purchase with Web Usage Mining. Advances in Intelligent Systems and Computing 174, 935–941 (2012)
10. Zhang, Z., Zhang, S.: The Research of Web Mining Algorithm Based on Variable Precision Rough Set Model. In: Jin, D., Lin, S. (eds.) Advances in FCCS, Vol. 1. AISC, vol. 159, pp. 573–578. Springer, Heidelberg (2012)
11. Menascé, D.A., Almeida, V.A.F., Fonseca, R., Mendes, M.A.: A Methodology for Workload Characterization of E-Commerce Sites. In: 1st ACM EC, Denver, CO, USA, pp. 119–128 (1999)
12. García, D.F., García, J.: TPC-W E-Commerce Benchmark Evaluation. IEEE Computer 36(2), 42–48 (2003)
13. Krishnamurthy, D., Shams, M., Far, B.H.: A Model-Based Performance Testing Toolset for Web Applications. Engineering Letters 18(2), 92–106 (2010)
14. Adomavicius, G., Tuzhilin, A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. Knowl. Data Eng. 17(6), 734–749 (2005)
15. Kim, J.-K., Cho, Y.H.: Using Web Usage Mining and SVD to Improve E-commerce Recommendation Quality. In: Lee, J.-H., Barley, M.W. (eds.) PRIMA 2003. LNCS (LNAI), vol. 2891, pp. 86–97. Springer, Heidelberg (2003)
16. Larose, D.T.: Discovering Knowledge in Data: An Introduction to Data Mining. Wiley-Interscience (2004)
17. Maravilla, N.: 8 Tips to reduce shopping cart abandonment (2012), `http://www.powerhomebiz.com/online-business/shop/shopping-cart-abandonment.htm` (access date: January 20, 2013)
18. Markov, Z., Larose, D.T.: Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage. Wiley-Interscience (2007)