CHARACTERIZING WEB SESSIONS OF E-CUSTOMERS INTERESTED IN TRADITIONAL AND INNOVATIVE PRODUCTS

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

Grzegorz Chodak
Department of Operational Research
Wroclaw University of Technology
Wybrzeże Wyspiańskiego 27
50-370 Wrocław, Poland
grzegorz.chodak@pwr.wroc.pl

KEYWORDS

Web traffic analysis, customer behavior, user session, click-stream analysis, log file analysis, Web server, e-commerce, innovative products

ABSTRACT

Web traffic characterization and modeling is currently a hot research issue. Low-level analysis of HTTP traffic on the server allows one to build adequate traffic models to be used in server benchmarking. High-level analysis of Web user behavior allows one to optimize website structure and develop personalized service strategies. In this paper, analysis of customer sessions in an online store is performed using Web server log data. The goal is to explore possible differences between sessions of customers viewing and purchasing innovative products, and customers only interested in traditional products.

INTRODUCTION

Electronic commerce has revolutionized the customers' approach to searching and buying products and services. Web users may access online stores regardless of time and space limits. A single visit of a customer to an online store corresponds to a series of Web pages opened by the customer via their Internet browser (a customer click-stream) and is called a *customer session* or a *Web session*.

The electronic environment makes it possible to collect detailed data on customers' behavior during their visit to an online store. This data may concern customers' searches, items viewed or added to the shopping cart, actions of purchase confirmation or withdrawal from the purchase, time spent in the store on performing various actions, etc. Basic data at HTTP level is recorded in standard Web server access log files. Reconstruction of customer sessions from logs is not a trivial task but it is worth an effort as it gives a researcher the ability to perform detailed analyses of many aspects of the Web traffic.

Various data mining methods have been applied so far to extract valuable knowledge on Web users' behavior and to predict customers' needs. The results have been used e.g., to improve customer relationship management (Xu and Wang 2011), to develop product recommendation and search support systems (Cho et al. 2013), (Kuang and Li 2014), (Huk et al. 2015), to manage the store inventory (Chodak 2011), or to predict sales (Mohammadnezhad and Mahdavi 2012), (Suchacka and Chodak 2013).

A significant aspect of a click-stream analysis has been the discovery of user navigation patterns and developing models of user sessions (Krishnamurthy et al. 2010), (Kwan et al. 2005), (Nenava and Choudhary 2013), (Shim et al. 2012). Traffic models may be used to generate a synthetic click-stream at the server input to test the server performance under realistic and controllable workloads. The models may also be used to develop mechanisms aiming at improving the quality of Web service and predicting Web performance (Borzemski and Kamińska-Chuchmała 2012), (Borzemski and Suchacka 2010), (Zatwarnicki and Zatwarnicka 2014), (Zhou et al. 2006).

In this paper, we focus on the statistical analysis of sessions performed by customers interested in innovative products compared to sessions of customers only interested in traditional products available in an online store. The comparison is done in terms of such session characteristics as the session length, session duration, and mean time per page. Our motivation was to characterize the behavior of e-customers (especially buyers) interested in innovative products as a group which potentially copes better than other users in a virtual environment, which itself has already an innovative character. The results may be interesting and useful for an online store manager.

ANALYSIS OF CUSTOMER SESSIONS IN AN ONLINE STORE

In the e-commerce environment particular attention should be paid to the analysis of sessions ending with a purchase (i.e. *buying sessions*) as they provide the online retailer with very important information, related to revenues and profits. This information concerns not only products bought by customers, but also the earlier customers' behavior in the store and sources of their visits (a reference from an organic or paid search engine result, a reference from an e-mail newsletter, entrance through a social media website, etc.). From a store manager's point of view, the analysis of buying sessions has the following objectives:

Proceedings 30th European Conference on Modelling and Simulation ©ECMS Thorsten Claus, Frank Herrmann, Michael Manitz, Oliver Rose (Editors)
ISBN: 978-0-9932440-2-5 / ISBN: 978-0-9932440-3-2 (CD)

- analysis of sources that referred customers to the store (this allows the retailer to assess the effectiveness of various marketing channels and their rates of return);
- customer segmentation in respect of further marketing activities (mailings, recommendation system, re-marketing activities, etc.);
- optimization of the store offers in terms of potential demand (e.g., based on information about categories of products viewed by customers and keywords typed into the store's search engine);
- optimization of the website's content, including the user interface;
- optimization of shipping and payment forms.

In practice, customers visiting online stores reveal differentiated navigational and behavioral patterns. This fact was a motivation to apply various segmentation or clustering techniques to determine customer profiles (Nenava and Choudhary 2013), (Song and Shepperd 2006), (Tanna and Ghodasara 2012), (Wang et al. 2004). In other studies VIP customers (Shim et al. 2012) or loyal customers (Chang et al. 2007) have been identified.

Since we wanted to investigate whether there are differences in traffic characteristics for customers interested in innovative and traditional products, we decided to distinguish between two subgroups of e-customers taking into account their preferences for types of viewed products. We have defined two types of customers in an online bookstore :

1) *Innovative customers* (*I*) are users who viewed some products considered "innovative", i.e. audio-books and multimedia products;
2) *Traditional customers* (*T*) are users who did not view any innovative products, but only traditional products, i.e. printed books.

RESEARCH METHODOLOGY

Raw data used for the analysis had been recorded from 1 April to 30 September 2014 in access logs of a Web server hosting the online bookstore (the address of the store website is not given in the paper due to a non-disclosure agreement). Data was written in the NCSA Combined log format. Each single HTTP request was described with the following data:

- IP address of the HTTP client (Web browser),
- date and time stamp meaning the time of the request coming to the server,
- HTTP method,
- version of HTTP protocol (1.0 or 1.1),
- HTTP status code,
- URI identifying the requested server resource,
- volume of data (in bytes) sent from the server in response to the request,
- URL of a site which linked the user to the store,
- user agent string, containing the name and version of the client Web browser.

A dedicated computer program was written in C++ to read, preprocess, clean, and analyze the data. Of all the HTTP requests we left only requests corresponding to page views (user clicks), eliminating requests for embedded objects (graphics and video files, etc.). Based on page view requests, click sequences for individual Web users were identified. Each unique user was identified based on two request data fields combined: the IP address of the client and the user agent string. Afterwards, Web sessions for each user were identified assuming that a minimum interval between two subsequent sessions of a given user is 30 minutes.

Since our goal was the analysis of sessions performed by customers viewing and buying products in the online store, we eliminated sessions issued by the website administrator and sessions performed by Internet bots. Bots were identified using a methodology proposed in (Suchacka 2014). We also eliminated sessions containing only one page view and/or lasting less than two seconds.

To verify and refine the description of customer sessions, we used data that had been gathered by the tracking software, SuperTracker, for the analyzed website. To obtain information on categories associated with products viewed in customer sessions, we combined data from three sources: server logs, SuperTracker database, and product database.

The entire data set contained 33 354 customer sessions. The subset of *innovative customer* sessions contained 6 171 sessions (including 466 buying ones) and the subset of *traditional customer* sessions contained 5 415 sessions (including 207 buying ones).

For each session three features were determined:

- session length – the number of pages opened during the session,
- session duration – the time interval between the last and the first customer's clicks in session,
- mean time per page – the average time of browsing a single page in session.

Such features of e-customers' visits may be obtained via some analytical tools, e.g., Google Analytics (GA). However, GA statistics may be less accurate than information gathered directly from logs because they are based on a sampling procedure (Google 2016). Session sampling may result in inaccurate results especially in the conversion analysis where conversions constitute a small fraction of all sessions. Furthermore, online retailers are not willing to share high-level information about traffic on their websites. Our software has the advantage that it uses low-level data written in server logs which are much easier to obtain.

RESULTS AND DISCUSSION

We analyzed session characteristics for *innovative* and *traditional* customers both for all sessions in each group and only for buying sessions in each group. The results are shown in Tab. 1-3.

Fig. 1 and Fig. 2 show histograms of session lengths, session durations, and mean times per page for all *innovative* and *traditional* customer sessions, respectively. To compare the results for both customer types, histogram intervals have a fixed width (the figures show up to fifteen first intervals).

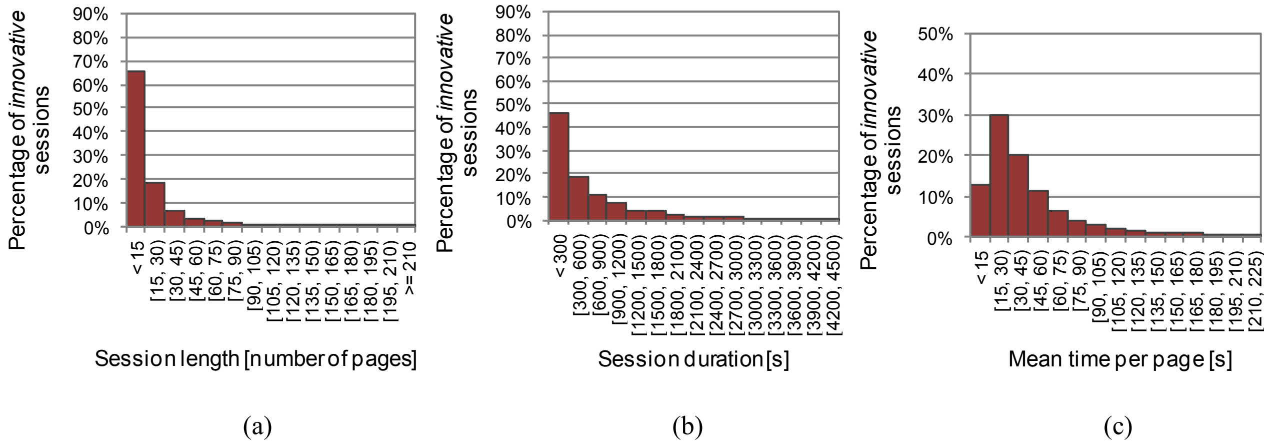


Figure 1: Data distributions for all *innovative customer* sessions
(**a**) Histogram of session lengths; (**b**) Histogram of session durations; (**c**) Histogram of mean times per page

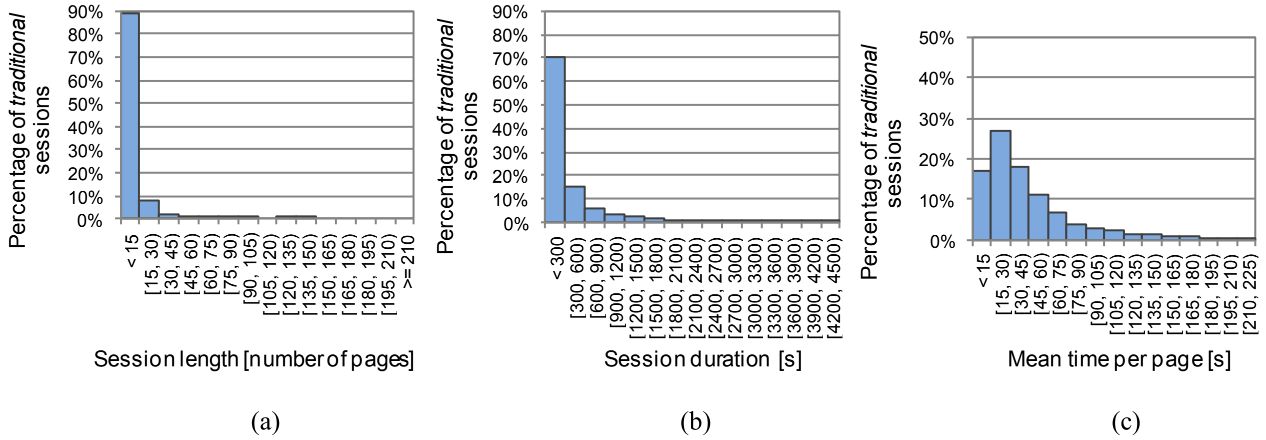


Figure 2: Data distributions for all *traditional customer* sessions
(**a**) Histogram of session lengths; (**b**) Histogram of session durations; (**c**) Histogram of mean times per page

Table 1: Session Length Statistics
(in Number of Pages in Session)

	All sessions		**Buying sessions**	
Statistics	***I***	***T***	***I***	***T***
Mean	17.4	7.7	60.9	37.3
Median	9	5	52	33
Mode	3	3	31	23
Std. dev.	23.7	9.3	36.9	19.2
Minimum	2	2	9	13
Maximum	286	145	264	145

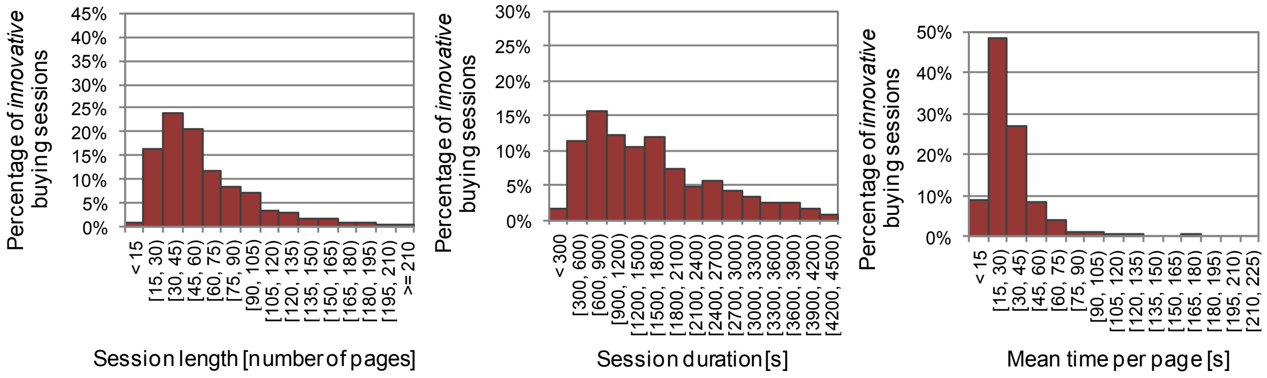
Table 2: Session Duration Statistics (in Seconds)

	All sessions		**Buying sessions**	
Statistics	***I***	***T***	***I***	***T***
Mean	680	327	1 801	1 164
Median	344	147	1 476	859
Mode	39	23	1 290	618
Std. dev.	893	491	1 337	926
Minimum	2	2	195	142
Maximum	10 043	7 544	10 043	5 772

Table 3: Mean time per page statistics (in Seconds)

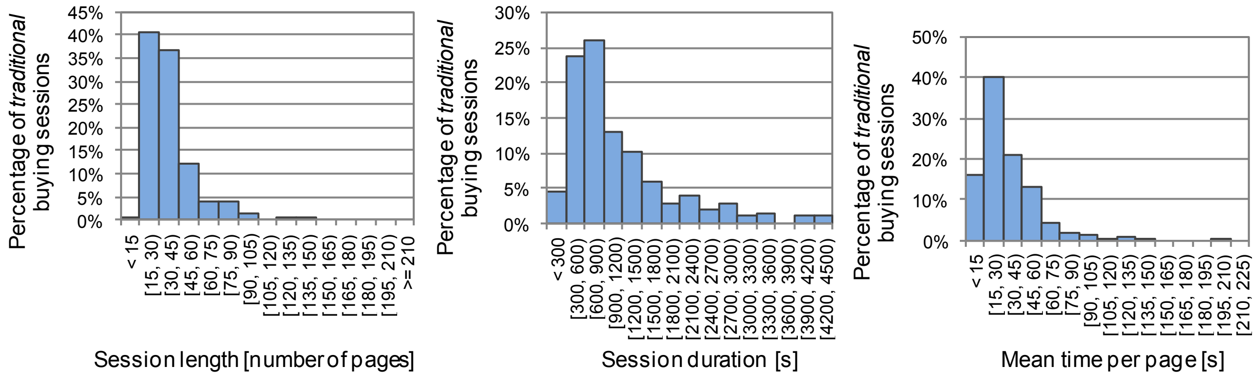
	All sessions		**Buying sessions**	
Statistics	***I***	***T***	***I***	***T***
Mean	69.4	67.7	31.8	34.1
Median	34.6	34.0	27.1	27.5
Mode	15.0	21.0	37.6	36.4
Std. dev.	131.2	126.6	19.0	25.4
Minimum	0.5	0.5	5.9	6.6
Maximum	1 788.0	1 641.0	171.0	197.6

One can notice in Tab. 1 that visitors interested in traditional, printed books typically open two times less pages in a session (7.7 on average) than customers searching for audio-books (17.4 on average). For *innovative customer* sessions both the mean and the median are much higher and their session lengths are more differentiated (as indicated by very high value of the standard deviation). For both customer groups histograms of session lengths (Fig. 1a, Fig. 2a) are right-skewed and long-tailed – very short sessions dominate especially in the case of *traditional customers*.



(a) (b) (c)

Figure 3: Data distributions for buying *innovative customer* sessions
(**a**) Histogram of session lengths; (**b**) Histogram of session durations; (**c**) Histogram of mean times per page



(a) (b) (c)

Figure 4: Data distributions for buying *traditional customer* sessions
(**a**) Histogram of session lengths; (**b**) Histogram of session durations; (**c**) Histogram of mean times per page

Session lengths correspond to session durations (Tab. 2, Fig. 1b, Fig. 2b). *Innovative customers* spend on average much more time interacting with the site (11.3 minutes) than *traditional* ones (5.4 minutes). However, statistics for the mean time per page (Tab. 3) are very similar for both groups – the mean is equal to about 1 minute (69.4 seconds for *innovative customers* and 67.7 for *traditional* ones). Such results suggest that generally customers do not differ much in the way of browsing pages and analyzing Web store content; however, they may have differentiated needs and expectations which cause *innovative customers* to perform more searching and browsing operations in the bookstore.

Differences in session characteristics for both customer groups are even more visible in the case of buying sessions (c.f. statistics for buying sessions in Tab. 1-3). Moreover, for each group the buying session characteristics differ from the general results in each group: buyers open many more pages (*innovative*: 60.9 pages compared to 17.4 pages, *traditional*: 37.3 pages compared to 7.7 pages) and spend much more time on the site (*innovative*: 30 minutes compared to 11.3 minutes, *traditional*: 19.4 minutes compared to 5.4 minutes), whereas the mean time per page is much shorter (more than half a minute compared to more than one minute for both customer groups). Buyers seem to only recently be familiar with the offers of the online store and do not browse information pages so intensively (moreover, the checkout process itself includes several pages informing them about the ordered products, the total cost, conditions and address of delivery, etc., and interaction with these sites is usually not very time-consuming, especially for regular customers).

For buyers, distributions of session lengths (Fig. 3a, Fig. 4a), session durations (Fig. 3b, Fig. 4b) and mean times per page (Fig. 3c, Fig. 4c) differ between the customer groups. In the case of *traditional* buyers short sessions (containing 15 – 45 pages, lasting 5 – 15 minutes) clearly dominate; furthermore, the session length (Fig. 4a) and session duration (Fig. 4b) distributions are clearly right-skewed, whereas the shapes of these histograms for *innovative* buyers (Fig. 3a, Fig. 3b) are different. As regards the mean time per page, it tends to be higher for *innovative* buyers than for *traditional* ones (Fig. 3c, Fig. 4c).

The differences between the statistics for two groups of buyers may be caused by the innovative products' descriptions which often include multimedia samples or film presentations (trailers, etc.). The multimedia content located on the product description page usually requires the customer to devote more time to them than other product description pages. Another hypothetical explanation may be such that *innovative customers* are more computer-oriented and therefore they spend more time in the online store as they feel more comfortable in a digital online store than *traditional customers*.

CONCLUSIONS

The paper discusses results of the statistical analysis of sessions performed by customers viewing and buying innovative and traditional products in an online bookstore. The results show that customers viewing and buying innovative products open on average many more pages and spend much more time interacting with the site than in the case of traditional products. This tendency is even more visible if we confine ourselves only to buying sessions.

The resulting distributions of the session lengths, session durations, and mean times per page may be used in simulation models. They may also be useful in setting input values to other data mining methods, e.g. association rules, applied to online store data. In a broader perspective the results may be used to personalize and improve the future customers' online shopping experience. It should be kept in mind, however, that user visits to different online stores may be characterized by different navigation patterns and thus, the results of our analysis cannot be automatically generalized to other e-stores.

In future work we are planning to extend our research to datasets from other e-commerce sites. Furthemore, as we considered a very limited number of session features in this study, we are planning to investigate intra-session dependencies between other features of sessions performed by customers interested in innovative products. Finally, it would be very interesting to apply some unsupervised machine learning methods to determine customer groups based on product categories rather than relying on arbitrary defining two customer subgroups. In this respect, we are planning to apply clustering methods to automatically divide e-customers into clusters using information on product categories.

ACKNOWLEDGEMENT

This work was partially supported by the National Science Centre (NCN) in Poland under grant no. 2013/11/B/HS4/01061.

REFERENCES

Borzemski, L. and A. Kamińska-Chuchmała. 2012. "Client-perceived Web performance knowledge discovery through turning bands method." *Cybernetics and Systems* 43, No.4, 354-368.

Borzemski, L. and G. Suchacka. 2010. "Discovering and usage of customer knowledge in QoS mechanism for B2C Web server systems." In *Proceedings of the 14th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (Cardiff, Wales, Sep.08-10), Lecture Notes in Artificial Intelligence 6277, Part II. Springer, Berlin Heidelberg, 505-514.

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, No.3, 753-764.

Cho, Y.S.; S.C. Moon; S.-p. Jeong; I.-B. Oh; and K.H. Ryu. 2013. "Clustering method using item preference based on RFM for recommendation system in u-commerce." *Lecture Notes in Electrical Engineering* 214, *Ubiquitous Information Technologies and Applications*, 353-362.

Chodak, G. 2011. "ABC analysis in an internet shop: a new set of criteria." In *Proceedings of the IADIS International Conference* (Avila, Spain, Mar.10-13), 196-204.

Google. 2016. "How sampling works." Analytics Help, https://support.google.com/analytics/answer/2637192?hl=en (access: Feb.11).

Huk, M.; J. Kwiatkowski; D. Konieczny; M. Kędziora; and J. Mizera-Pietraszko. 2015. "Context-sensitive text mining with fitness leveling genetic algorithm." In *Proceedings of the 2nd International Conference on Cybernetics* CYBCONF'15 (Gdynia, Poland, Jun.24-26). IEEE, New York, USA, 183-188.

Krishnamurthy, D.; M. Shams; and B.H. Far. 2010. "A model-based performance testing toolset for Web applications." *Engineering Letters* 18, No.2, 92-106.

Kuang, G. and Y. Li. 2014. "Using fuzzy association rules to design e-commerce personalized recommendation system." *TELKOMNIKA Indonesian Journal of Electrical Engineering* 12, No.2 (Feb), 1519-1527.

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, No.1, 189-204.

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 Services* (IJITCS) 2, No.1 (Feb), 1-8.

Nenava, S. and V. Choudhary. 2013. "Hybrid personalized recommendation approach for improving mobile e-commerce." *International Journal of Computer Science & Engineering Technology* (*IJCSET*) 4, No.5, 546-552.

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, No.9, 7736-7742.

Song, Q. and M. Shepperd. 2006. "Mining Web browsing patterns for e-commerce." *Computers in Industry* 57, No.7, 622-630.

Suchacka, G. and G. Chodak. 2013. "Practical aspects of log file analysis for e-commerce". In *Proceedings of the 20th International Conference Computer Networks* CN'13 (Lwówek Śląski, Poland, Jun.17-21), Communications in Computer and Information Science 370. Springer, Berlin Heidelberg, 562-572.

Suchacka, G. 2014. "Analysis of aggregated bot and human traffic on e-commerce site." In *Proceedings of Federated Conference on Computer Science and Information Systems* FedCSIS'14 (Warsaw, Poland, Sep.7-10), ACSIS, Vol. 2. IEEE, New York, USA, 1123-1130.

Tanna, P. and Y. Ghodasara. 2012. "Exploring the pattern of customer purchase with Web usage mining." In *Proceedings of the International Conference on Advances in Computing* ICAdC'12 (Bangalore, India, Jul.4-6), AISC 174. Springer India, New Delhi, India, 935-941.

Wang, Q.; D.J. Makaroff; and H.K. Edwards. 2004. "Characterizing customer groups for an e-commerce website." In *Proceeding of the 5th ACM Conference on Electronic Commerce*. ACM Press, New York, 218-227.

Xu, H. and L. Wang. 2011. "Application of analysis CRM based on association rules mining in variable precision rough set". In *Proceedings of International Conference on Computer Science, Environment and Ecoinformatics* (Wuhan, China, Aug.21-22), Communications in Computer and Information Science 216. Springer, Berlin Heidelberg, 418-423.

Zatwarnicki, K. and A. Zatwarnicka. 2014. "The cluster-based time-aware Web system." In *Proceedings of the 21th International Conference Computer Networks* CN'14 (Lwówek Śląski, Poland, Jun.23-27), Communications in Computer and Information Science 431. Springer, Berlin Heidelberg, 37-46.

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, No.8, 838-850.

AUTHOR BIOGRAPHIES

GRAŻYNA SUCHACKA received the MS degrees in Computer Science and in Management from Wroclaw University of Technology, Poland. She received her Ph.D. degree in Computer Science from Wroclaw University of Technology. Now she is an assistant professor in the Institute of Mathematics and Informatics at Opole University, Poland. Her research interests include Web mining, Web analytics, and Quality of Web Service with special regard to electronic commerce. Her e-mail address is: gsuchacka@math.uni.opole.pl.

GRZEGORZ CHODAK received the MS degree in Computer Science from Wroclaw University of Technology, Poland. He received his Ph.D. degree and habilitation in Management from Wroclaw University of Technology. Now he is an assistant professor in the Department of Operational Research at Wroclaw University of Technology. His research interests include e-commerce, logistics with regard to online stores. His e-mail address is: grzegorz.chodak@pwr.wroc.pl.