EXPLICIT DATA OF INTERNET USERS: VALIDITY AND DECEPTION DISCLOSURE

Jitka Pokorná, Tereza Balcarová

Abstract: *This study is focused on validation of the explicit data provided by Internet users for the prediction of required on-line content. Content prediction represents the core of on-line personalization process. Whether explicit data quoted by users reflect the reality analyze this research. Research participants (n = 32) filled the electronic questionnaire in and were monitored to validate questionnaire data. The frequencies of overall weekly on-line time and doing particular on-line activity are measured. Evaluated on-line activities are: on-line communication including sharing, reading on-line news, on-line entertainment and information search. Frequency of Internet use was determined as the respondents estimation of the time spent on-line in hours per week. Five zero hypotheses were proposed and tested with Wilcoxon non-parametric paired test. Hypotheses were accepted in cases of total frequency of Internet usage, on-line communication, entertainment and information search. Respondents mostly differ in estimation of reading on-line news frequency. Frequency of reading on-line news the null hypothesis was rejected. In all cases of difference, respondents overestimated frequency weekly.*

Keywords: *Explicit data, Implicit data, Personalization, Validation, Internet use, On-line deception.*

JEL Classification: *D83.*

Introduction

Personalization includes the selection of dynamic content such as links, advertisement, evaluation or recommendation, which is of interest to a particular user or group [1]. According to previous study [5], personalization is managed conjunction of categorized content according to the profile of an Internet user. Personalization on the website [9] is defined as a process of change of content and structure of a page for the purpose of adaptation to specific needs, goals, interests and preferences of each user. Personalized website deliver potential interesting content to users, on the other side it helps make profits to on-line service providers also. In previous surveys 80 % of respondents said they like to adopt personalized content, especially when it comes to recommendations from other users and this preference tends to increase [4]. Provider [4] declares double growth of relations number at the suggested level of content, 10 % profit growth and 2,4 times as much longer time spend at the Internet page. However, negative aspects of personalization are mentioned in sub-chapter 1.1.

In previous studies the active (explicit) and passive (implicit) form of identifying the Internet users, which allows more accurate analysis of personalization techniques, were defined. Active approach is based on the explicit ways of data collections, realized mostly by questionnaire surveys or forms [16]. Active method requires activity from users during the data collection; implicit method is based on passive data collected automatically by the system.

Implicit personalization method

Implicit method or passive approach is more complex; because it requires prediction based on analyzing the previous user on-line behavior, most often detected by click logging or data on the user's service use. The aim of the implicit personalization system is systematically decide which content is the most relevant for the user according to the data about user's previous on-line activities [11].The monitored aspects can be, for example, sites or category of sites the user has visited in the past or which advertisement the user has reacted to in the past with a mouse click. The user profile is assembled on the basis of on-line behavior, content of visited websites or both. The on-line behavior model consists of monitoring user activities such as clicking, downloading, frequency of use or of a record of activity on a specific website [13]. Other researches summarize implicit metrics of targeted advertising: time stamp, IP address, set of attributes from cookies, click, and ad ID or ID of ad placement [11].

The main criticism of passive approach points to the user's privacy violation of the user's privacy [17]. Personal information on the user, such as his name, address, and e-mail or telephone number is not generally collected by the personalization system, and the monitored data can be considered anonymous in this regard. The system more often gathers records on website the user has already visited, how long he or she spent there or what user´s next interest is. The problem is a long-term nature of user monitoring and recording the user's interests and other activities. Cookies are the most common intermediary between information on the user and the Internet company [8]. Many users and legislators worldwide are facing issues dealing with the privacy of implicit techniques for targeting ads. There can also be seen interferences between the development of the implicit targeting by automatic learning systems and the claims of opponents for the right to provide personal data only with the permission of the user [8]. In March 2011 it was reported that the online advertising industry started cooperation with the administrative authorities in the control of on-line monitoring of user activity, and therefore changes in the techniques of data gathering systems can be expected [14].

Explicit personalization method

Explicit method or active approach requires user to state information about him on the website. The most common form of the explicit data is an electronic questionnaire or form. Users are not, however, always willing to fill in information or the information might not be true [17].

Research shows that a lie and deception on the Internet exists as long as the Internet itself. The impersonal nature of on-line communication increases the opportunity for deception such as the current activity on the Internet [10], [20]. The differences are evident in certain Internet activities. The e-mail communication can be observed as a tool frequently burdened with the deception, that is increasing with growing frequency of e-mail use [10], [3]. Although on-line chat users lie less frequently, than occasional chat visitors [21]. Generally, frequent Internet users deceive more than infrequent users [10], [3]. Previous studies show, that these Internet users mislead information about their physical characteristics, psychological characteristics and on-line activities [10], [3], [6]. The most common categories, which reflect false data include age, education, occupation, religion, income and gender. Lying within the on-line environment occurs more in men than in women [21]. The reasons for deceiving are also discussed. The motivation to lie on-line is mostly caused because of safety reasons [22], [3]. Other reasons, according to [12] may

be identity games and perhaps psychological disorder. The problem, which is still current, is the detection of such lies. The authors state that it is still very difficult to detect the on-line environment deception [20], [23]. For further use of the active learning of Internet users is essential to obtain true data.

1 Statement of a problem

User's identification parameters, preferences, requirements or the context are collected for the purposes of personalization primary [9]. The user may have varied interests at various times, and may also search information in different contexts. Component interests may, however, be motivated by the same interest at a higher level of abstraction [13]. Finding user typology on the Internet based on motives satisfied on-line was the focus of the previous study [15], where following groups of users were identified: overview, escape from the reality of life, information, entertainment and maintenance of social connections. On the basis of exploratory factor analysis, three ways of Internet use for Generation Y where defined [19]; the Internet as a resource of: entertainment, communication, and information. On the basis of theoretical research, the following Internet user types were identified [2]: non-users, sporadic users, discussing users, users being entertained, social, observers, instrumental and advanced. Those Internet user typologies are based on motives for which the medium is used. Found user categories are similar to on-line activities classification questioned in national surveys of Internet use [7], [18].

Research on information and communications technology use in the Czech Republic comes from investigation of a descriptive nature, is coordinated by the European Statistical Office and takes place each year in all member nations of the European Union and other selected European countries. In the Czech Republic, this investigation under the name of Selective Investigation on Information and Communications Technologies Use in households and amongst individuals is performed by the Czech Statistical Office [7]. Among other monitored variables, on-line activities that users have undertaken in the past 3 months are questioned (for on-line shopping in the past year) for private purposes: communications (sending/receiving e-mails, telephone and video conversations), searching for information (on goods and services, health, travel and accommodations, reading of on-line news reports, job searching), entertainment (playing and downloading games, watching or downloading movies, music, videos, downloading software, listening to on-line radio or television), on-line services (shopping, sales of goods or services, Internet banking, communication with the government, sending files). Netmonitor company is providing information on Internet visitor statistics and the socio-demographic profile of its visitors in the Czech Republic [18]. The implementer of this project is the Media search company in cooperation with Gemius S.A. For data collection the system use a hybrid method which measures on both server and client browser side. Respondents are questioned on the following on-line activities: communication, searching information on products and services, searching professional information for their work/studies, listening to music, watching videos on the Internet (for example, YouTube, stream, TV station archives, on-line radio).

Based on the Internet user typologies together with the on-line activity categories following activities are measured: reading on-line news, on-line communication including sharing (email, Facebook, Skype, etc.), information search and on-line entertainment (games, music, videos, stream, etc.). We can expect using Internet for professional purpose

especially at work. However, only home computers are monitored in this research. Professional purpose of searching information is not measured.

Accuracy of explicit personalization method depends on the validity of data provided by the user. Previous studies show [10], [3], [6], that the Internet users mislead information about their characteristic or on-line activities. Following up these results the aim of this study is to validate explicit data on on-line activities provided by Internet users. Summarizing the arguments above, following hypotheses are proposed:

$H1_0$: Median difference between questioned and monitored overall Internet use is null.

$H2_0$: Median difference between questioned and monitored frequency of on-line communication is null.

$H3_0$: Median difference between questioned and monitored frequency of news reading is null.

$H4_0$: Median difference between questioned and monitored frequency of on-line entertainment is null.

$H5_0$: Median difference between questioned and monitored frequency of on-line information search is null.

2 Methods

Identical primary data were collected using both explicit and implicit data collection. For these purposes method of questioning with electronic questionnaire and monitoring performed by overt, structured and indirect observation. The survey was intended to the Internet users above the age of 15 years. Selection of users was intended with the requirement of both gender representations. 32 Internet users participated in the questionnaire survey (19 men and 13 women) as well as monitoring. Research participants firstly responded to the electronic questionnaire. Frequency of Internet use is determined by the estimation of time spent on-line in hours per week stated by the respondent. The overall weekly on-line frequency and the frequency of dedication to a particular activity are measured (reading on-line news, on-line communication, information search and on-line entertainment). For the research purposes, gender and age were settled as questioned socio-demographic variables. Three filtration questions were related to exclusive usage of home PC, Mozilla FireFox browser and users regular rhythm of life within the monitoring time. These questions were principal for the subsequent observation. For the purpose of mentioned indirect observation a special monitoring application was developed for Mozilla Firefox web browser. On this condition only home computer Internet usage is monitored. The monitoring was conducted under the regular rhythm of life with regard to the reduction of irregular, random, unusual or extreme situations that could misrepresent the research results.

Monitoring of Internet users was performed to validate explicit data. The following variables were recorded: user ID, IP address, domain, date, relation start, end and duration in seconds and domain. For the duration of seven monitored days 86 176 relations (clicks) were collected. Statistical test were conducted using SPSS 19 program. Monitored data were filtred and coding in MS Excel into categories pursued in the questionnaire. To determine the validity of questioned data, estimated values were compared with the really measured. Hypotheses were tested by the Wilcoxon non-parametric paired test. The match

of questioned and monitored data was analyzed. For the research purposes, the level of significance was determined at α 0,05.

3 Problem solving

Whether or not the monitored data mentioned corresponds to the actual measured values was determined through Wilcoxon paired test. The match of following parameters was tested: differences in total Internet use frequency, communication, on-line news reading frequency entertainment and information search. In the tables below the coding of parameters were used: (Q) for estimated value obtained by questionnaires and (M) for real measured values obtained by monitoring. As stated at chapter 2 five zero hypotheses are tested.

Tab. 1: Wilcoxon paired test – overall Internet usage

Wilcoxon paired test: Marked tests are significant on the level p <,05000		
	Valid no.	p-value
Overall usage (Q) & Overall usage (M)	8	0,068740
Communication (Q) & Communication (M)	7	1,000000
News reading (Q) & News reading (M)	14	0,000892
Entertainment (Q) & Entertainment (M)	7	0,342811
Information (Q) & Information (M)	7	0,091270

Source: [own results]

Comparing data about overall Internet usage the calculated p-value (0,069) is greater than the set level of 0,05 and that is why we not rejected the zero hypothesis, that there is no significant difference between questioned and monitored overall Internet use.

In case of questioned and monitored data about on-line communication the calculated p-value (1,000000) is greater than the set level of 0,05 and that is why we not rejected the zero hypothesis, that there is no significant difference between questioned and monitored frequency of on-line communication.

Comparing data about on-line news reading the attained value of significance (0,000892) is less than the set level of 0,05 and that is why we reject the zero hypothesis that the difference between the estimated and actual time of reading on-line news is zero. From the results, it can be judged that users perceive time spent reading on-line news as longer than it is in reality.

Comparing data about on-line entertainment the calculated p-value (0,342811) is greater than the set level of 0,05 and that is why we not rejected the zero hypothesis, that there is no significant difference between the estimate of searching for on-line information and the actual measured value.

In case of questioned and monitored data about searching on-line information the calculated p-value (0,091270) is greater than the set level of 0,05 and that is why the zero hypothesis is not rejected. There is no significant difference between the estimate of searching for on-line information and the actual measured value.

4 Discussion

Primary data were gathered both by using electronic questionnaire (explicit data) and one-week observation (implicit data) of on-line activities performed by monitored respondents. The explicit data validity testing was conducted by Wilcoxon paired test that confirmed statistically significant compliance of both samples especially in following cases: overall Internet use frequency, on-line communication frequency, information search and also on-line entertainment.

Tab. 2: Results of hypotheses testing

Hypothesis	p-value	Testing result
$H1_0$	0,0687	Not rejected
$H2_0$	1,0000	Not rejected
$H3_0$	0,0008	Rejected
$H4_0$	0,3428	Not rejected
$H5_0$	0,0912	Not rejected

Source: [own results]

The difference of explicit data was statistically significant only in the frequency of reading on-line news. Monitored respondents overestimated time spent by reading on-line news. This result may lead a number of reasons. One of the reasons may be the implementation of monitoring on home computers. Results are different, due to the fact that respondents do not differ between reading on-line news at home and reading on-line news within working hours.

Hypotheses of a zero median difference between the actually measured and estimated data were not rejected in the case of on-line communication, entertainment, information search and the overall frequency of Internet use. Given the very low p-value for hypothesis $H1_0$ ($p = 0,0687$), the agreement of real and estimated overall time spent on-line is very low. The reason for this may be an overestimation of the total time spent on-line. Another cause can be the deceptive statement of the less time spent online that is also declared in the previous studies [3], [6], [10], [20]. Internet users are intentionally given untruthful data because of security reasons [3], [22]. This research is limited by the fact that only one computer device is monitored. Nowadays personal computers, notebooks, iPads, mobile phones or similar devices are used together for on-line activities. However, extending the research with new measured devices requires different monitoring software for both particular operating system and web search engine. That is a technically challenging issue. Due to private character of data (monitored on-line activity) not many people are willing to contribute to the research. Only the most used devices private on-line activity was monitored in this research. However, respondents did not distort the results significantly.

Conclusion

The aim of this study was to validate explicit data of on-line activities provided by Internet users. Whether explicit data provided by electronic questionnaire match the real monitored values was determined through pair value testing.

In case of overall Internet use, on-line communication, entertainment and information search, a statistically significant accordance with both values was confirmed. The difference

in estimated values and real data was determined for the frequency of on-line news reading. The difference may be caused by the fact that the monitored users might be reading on-line news during working hours. With the exception of reading on-line news frequency, the users confirmed the real estimated time period spent on the concrete on-line activity. In spite of all research limitations the results show that users are relatively well oriented in time spent on-line. They did not deceive with their on-line activities frequency. The on-line data about the frequency of overall usage, communication, information search, etc. describes the way how Internet is used by individuals. Understanding this can be used to adjust offered on-line content and to set personalization algorithm in particular conditions. For the further research, a longer time period should be considered to avoid unexpected influences on data.

Data about the frequency of overall usage, communication, information search, etc. can be gathered implicitly without action of users. Increasing importance of active method of collection relates to the future of behavioral targeting legislation. There is a world discussion about on-line privacy and implicit personalization methods. In recent years, there is a pressure on personalized system provider to make collection of data transparent. Implicit data collection methods are required to be used with explicit agreement of the user only. As the results show, almost the same data is possible to obtain explicitly from users. The results prove the potential of using explicit data for the purposes of personalization. On the other hand it is necessary to accept the fact that respondents must actively respond, therefore they must be willing to cooperate. Explicit data is perspective, due to the existence of arguments about the legality of obtaining the implicit data through monitoring.

References

[1] BARAGLIA, R., SILVESTRI, F. Dynamic personalization of web sites without user intervention. *In Communications of the ACM*, 2007, Vol. 50, Iss. 2, pp. 63-67. ISSN 0001-0782.

[2] BRANDTZAEG, P. Towards a unified Media-User Typology (MUT): A meta -analysis and review of the research literature on media-user typologies. *In Computers in Human Behaviour*. 2010. [cit. 2012-09-06]. Available from WWW: <http://www.journals.elsevier.com/computers-in-human-behavior/>

[3] CASPI, A., GORSKY, P. Online Deception: Prevalence, Motivation, and Emotion. *In Cyber Psychology & Behavior*, 2006, Vol. 9, Iss. 1, pp. 54-59. ISSN 2152-2715.

[4] CHOICESTREAM. *PERSONALIZATION SURVEY:Consumer Trends and Perception s*. 2007. [cit. 2012-09-06]. Available from WWW: <http://www.lazworld.com/whitepapers/internet_marketing_whitepapers/ChoiceStream_PersonalizationSurveyResults2007.pdf>

[5] CONER, A. Personalization and Customization in Financial Portals. *In Journal of the American Academy of Business,* 2003, Vol. 2, Iss. 2. ISSN 1540-1200.

[6] CORNWELL, B., LUNDGERN, D. C. Love on the Internet: Involvement and misrepresentation in romantic relationships in cyberspace vs. realplace. *In Computers in Human Behavior*, 2001, Vol. 17, Iss. 2 pp. 197–211. ISSN: 0747-5632.

[7] ČSÚ. *Využívání informačních a komunikačních technologií v domácnostech a mezi jednotlivci.* 2011. [cit. 2012-09-06]. Available from WWW: <http://www.czso.cz/csu/2011edicniplan.nsf/p/9701-11>

[8] DWYER, C. Behavioral Targeting: A Case Study of Consumer Tracking on Levis.com *In Proceedings of the15th Americas Conference on Information Systems,* California, USA, August, 2009.

[9] GARRIGÓS, I. et al. Specification of personalization in web application design. *In Information and Software Technology*, 2010, Vol. 5, Iss. 52. pp. 991-1010. ISSN: 0950-5849.

[10] HANCOCK, J. T., THOM-SANTELLI, J., AND RITCHIE, T. Deception and design: The impact of communication technology on lying behavior. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* Vienna: SIGCHI, Austria, 2004, pp. 129-134. ISBN:1-58113-702-8.

[11] HIGGS, B., RINGER, A. C. *Trends in consumer segmentation.* [online]. 2007 [cit. 2012-09-06]. Available from WWW: <http://vuir.vu.edu.au/874/1/Trends_in_Consumer_Segmentation-Final.pdf>

[12] JOINSON, A. M., DIETZ-UHLER, B. Explanations for the penetration of and reaction to deception in a virtual community. *In Social Science Computer Review*, 2002, Vol. 20,Iss. 3, pp. 275–289.ISSN:0894-4393.

[13] KIM, H. R., CHAN, P. K. Learning implicit user interest hierarchy for context in personalization. *In Proceedings of the 2003 International Conference on Intelligent User Interfaces (IUI'03).* Miami: ACM, USA, 2003, pp.101–108. ISBN:1-58113-586-6.

[14] LEE, K. Behavioural Targeting am Europäischen Verbrauchergipfel. *Adage* [online]. 2011 [cit. 2012-09-06]. Available from WWW: <http://adage.com/article/digital/behavioral-advertising-principles-enforced/149228/>

[15] LEUNG, L. Global Impacts of Net generation attributes, seductive properties of the Internet, and gratifications-obtained on Internet use. *In Cyberpsychology, Behavior, and Social Networking*, 2004, Vol. 7, Iss. 3, pp. 333-348. ISSN: 2152-2723.

[16] MONTGOMERY, A. L., SRINIVASAN, K. Learning about customers without asking, InN. Pal and A. Rangawamy (eds.), *The Power of One-Leverage Value from Personalization Technologies*, Penn State University: eBRC Press. 2002. [cit. 2013-01-04]. Available from WWW: <http://repository.cmu.edu/tepper/324/?utm_source=repository.cmu.edu%2Ftepper%2F324&utm_medium=PDF&utm_campaign=PDFCoverPages>

[17] MONTGOMERY, A. L., SMITH, D.M. Prospects for Personalization on the Internet, *In Journal of Interactive Marketing*. 2008. [cit. 2012-09-06]. Available from WWW: <http://repository.cmu.edu/heinzworks/46/?utm_source=repository.cmu.edu%2Fheinzworks%2F46&utm_medium=PDF&utm_campaign=PDFCoverPages>

[18] NETMONITOR. *Výzkum sociodemografie návštěvníků internetu v České Republice*. 2012. [cit. 2012-09-06]. Available from WWW: <http://www.netmonitor.cz/sites/default/files/vvnetmon/2012_06_total.p >

[19] POKORNÁ, J. Role nových médií v životě generace Internetu. *In Proceedings of the Think Together 2009 Conference.* 2009. Praha: ČZU. ISBN 978-80-213-1906-6.

[20] TOMA, C. L., HANCOCK, J. T. What Lies Beneath: The Linguistic Traces of Deception in Online Dating Profiles. *In Journal of Communication*, 2012, Vol. 62, Iss. 1, pp. 78-97. ISSN 1460-2466.

[21] WHITTY, M.T. Liar, liar! An examination of how open, supportive and honest people are in chat rooms. *In Computers in Human Behavior*, 2002, Vol. 18, Iss. 4, pp. 343 –352. ISSN: 0747-5632.

[22] WHITTY, M. T., GAVIN, J. Age/sex/location: uncovering the social cues in the development of online relationships. *In Cyber Psychology & Behavior*, 2001, Vol. 4, Iss. 5, pp. 623–630.ISSN: 2152-2715.

[23] WHITTY, M. T., JOINSON, A. *Truth, lies and trust on the Internet.* New York, NY: Psychology Press, 2009. ISBN: 978-1841695846.

Contact Address

Ing. Jitka Pokorná, Ph.D.
Czech University of Life Sciences, Faculty of Economics and Management
Kamýcká 129, 165 21, Prague 6 – Suchdol, Czech Republic
E-mail:pokornaj@pef.czu.cz
Phone number: +420 224 382 239

Ing.Tereza Balcarová, Ph.D.
Czech University of Life Sciences, Faculty of Economics and Management
Kamýcká 129, 165 21, Prague 6 – Suchdol, Czech Republic
E-mail:balcarova@pef.czu.cz
Phone number: +420 224 382 239

Received: 01. 05. 2013
Reviewed: 04. 06. 2013, 12. 08. 2013
Approved for publication: 04. 11. 2013