Web Personalization and the Privacy Concern

Konstantinos Markellos, Penelope Markellou, Maria Rigou, Spiros Sirmakessis, Athanasios Tsakalidis

Abstract

On the road to enhance website user's experience and treat each user individually, personalization plays a central role. Personalization is striving to identify the user, record the user's online behaviour in as much detail as possible and extract needs and preferences in a way the user cannot notice, understand or control. It is a process that can put user's privacy in jeopardy. This paper discusses the fine line between web personalization and personal intrusion and control when recording and modelling the user. It also investigates when this line is crossed and indicates the available technological solutions and standards for protecting user's privacy online.

1. Introduction

Nowadays, more and more people have access to the Internet but as the web is growing exponentially, human ability to find, read, and understand content remains constant. *Web personalization* is perceived as one of the most promising approaches to alleviating this problem of information overload providing users with tailored experiences [17]. It comprises a broad scientific and technological area, also covering recommendation systems, customization, one-to-one marketing, and adaptive websites [1], [12], [18].

Various definitions of personalization can be found in the relative literature; i.e. *"personalization is the provision to each individual of tailored information, products, or services"* [11] or *"personalization is defined as any action that adapts the information or services provided by a website to the knowledge gained from the users' navigational behaviour and individual interests, in combination with the content and the structure of the website"* [3]. Generally, personalization is of high importance for the design and implementation of intelligent web applications e.g. commercial websites, information portals, e-commerce sites, e-learning systems, etc. In short, the personalization process includes the steps of gathering and storing information about website visitors, analyzing the information in order to extract user patterns, habits and preferences, and -based on this analysis- delivering the right information to each visitor at the right time.

The benefits for both businesses and users can be significant [14]. More specifically, businesses can use personalization as a means for increasing site usability and user's satisfaction, promoting loyalty, establishing one-to-one relationships, converting users to buyers, retaining current users, re-engaging users, and penetrating new markets. Users on the other hand, need to feel that they have a unique personal relationship with the websites they come in contact with. They have now the ability to visit personalized website pages that allow them to find information or choose products and services for buying fast, easily and according to their preferences. Moreover, they can receive e-mail, newsletter or other information that suits their personal interests. This can play an important role in customer's faith and loyalty to the business behind the website.

Achieving the aforementioned benefits, it is obvious that personalization techniques require collecting and storing far more personal data than those used by ordinary, non-personalized websites. A login (usually user's name) and a password are not enough for constructing and

updating user profiles. However, most web-users are not willing to reveal more information about themselves. They want to be reassured that their personal information will not be shared with anyone else without their prior explicit permission and this forces a number of ethical restrictions to the design, implementation and delivery of personalization and causes considerable controversy. While personalization looks important and appealing for the web experience, several issues related to *privacy* still remain unclear [22]. Even though this term has many connotations in the society the following definition can be adopted for the case of web *"privacy is the subjective condition a person experiences when two factors are in place, firstly he must have the power to control information about himself/herself and secondly he must exercise that control consistent with his/her interests and values"* [16].

The 6th WWW User Survey conducted by the Graphics, Visualization and Usability Center of the Georgia Institute of Technology showed that the main reason for not registering in a website is that the terms/conditions of how the collected information is going to be used are not clearly specified (70%) [5]. Another survey conducted by the Personalization Consortium indicated that privacy issues are important for the users but they would share personal information in exchange for better services [15]. Moreover, 58% of users require a privacy statement from the website and even 51% read it before registering on the site. Furnell and Karweni [4] in their study found that 87.5% of surveyed consumers expect to see comprehensive information regarding the privacy policy when visiting a commerce website. On the other hand the survey in [8] examined websites of the Fortune 500 and showed that slightly more that 50% of sites provide privacy policies on their home pages.

Personalization and the closely relating web mining techniques have to overcome the privacy problem in order to present satisfactory results [20]. These novel technologies retrieve and analyze data from different sources in order to extract "hidden" information and knowledge. In this framework it is obvious that privacy jeopardised since for constructing and updating individual or group profiles such systems gather a set of personal user data and also monitor and record user online activity. There are many ways in which users leave traces on the web. To maximize data gathering opportunities, websites collect data from every user touch point, online (registration, transactions, sign-ups, profiles, preferences, surveys, services, web log files, advertising banners, sweepstakes and other promotions requiring user's data) and offline (services by phone, in-store transactions, paper submissions like sweepstake or promotion entries). Then they deploy mining techniques (such as association rules, clustering algorithms, classification techniques, collaborative filtering, patterns discovery, log files analysis, etc. [21]) for producing personalized output (i.e. content, structure or presentation and media format). Many points in this process can present threats to customer privacy. Both users and websites can take measures against these threats, for example privacy policies, privacy seals, or privacy-protecting technologies (proxies or anonymizers, P3P, client-side profiling, pseudonymity, identity management), etc.

In this framework, the purpose of the paper is firstly to present some theoretical issues concerning the meaning and the significance of privacy. Then we discuss the fine line between web personalization and personal intrusion and control when recording and modelling the user. The cases when this line is crossed are also identified. The need for privacy has also created a market of products designed to protect user's personal information. The solutions that technology can offer to website owners and visitors and the most well-known domain standards are presented. Finally, we discuss open research issues since apart from the privacy threats that may emerge from the use of web mining for personalization, it is the same technology that can also be deployed to identify privacy violations.

2. Privacy threats due to personalization

2.1 The Personalization process decomposed

The overall personalization process can be decomposed in three discrete modules, namely *data acquisition*, *data analysis* and *personalized output* [10]. This section describes the objectives and the tasks of each module and investigates the potential hazards for privacy.

2.1.1 Data acquisition

In the large majority of cases, web personalization is a data-intensive task that is based on three general types of data; data about the user, data about the website usage and data about the software and hardware available on the user's side.

User data regard personal characteristics of the user and typically include demographics (name, phone number, geographic information, age, sex, education, income, etc.), skills, capabilities, interests, preferences, goals and plans, depending on the types of adaptations the system has to deliver for the personalization effect. There are two general approaches for acquiring user data: either the user is asked *explicitly* to provide the data (using questionnaires, fill-in preference dialogs, or even via machine readable data-carriers, such as smart cards), or the system *implicitly* derives such information without initiating any interaction with the user (using acquisition rules, plan recognition, and stereotype reasoning).

Usage data may be directly observed and recorded, or acquired by analyzing observable data (whose amount and detail varies depending on the technologies used during website implementation, i.e. java applets, etc.), a process termed as web usage mining. Usage data may either be:

- Observable data comprising selective actions like clicking on an link, data regarding the temporal viewing behavior, ratings (using a binary or a limited, discrete scale) and other confirmatory or disconfirmatory actions (making purchases, e-mailing/saving/printing a document, bookmarking a web page and more), or
- Data that derive from further processing the observed and regard usage regularities (measurements of frequency of selecting an option/link/service, production of suggestions/recommendations based on situation-action correlations, or variations of this approach, for instance recording action sequences).

Environment data addresses information about the available *software* and *hardware* at the client computer (browser version and platform, availability of plug-ins, firewalls preventing applets from executing, available bandwidth, processing speed, display and input devices, etc.), as well as *locale* (geographical information in order to adjust the language, or other locale specific content).

The acquired data need to be transformed into some form of internal representation (modeling) that will allow for further processing and easy update, in order to be used as the basis for the personalization decisions. Such internal representation models are used for constructing individual or aggregate (when working with groups of users) profiles, a process termed *user profiling* in the relative literature. Profiles may be *static* or *dynamic* based on whether -and how often- they are updated. Static profiles are usually acquired explicitly while dynamic ones are acquired implicitly by recording and analyzing user navigational behavior.

2.1.2 Data analysis

The data analysis module comprises the techniques that may be applied for further analyzing and expanding user profiles so as to derive secondary inferences and thus accomplish more sophisticated personalization. The available techniques vary and come from numerous scientific areas that comprise artificial intelligence, machine learning, statistics, and information retrieval. The kind of analysis that can be applied after data acquisition is affected dramatically by user profiling (as described in the previous section).

Some personalised systems can get along with fairly simple model structures. Other such systems, have higher demands with respect to user and usage model representation, and also need to employ inferences to further augment the user and usage model based on initial acquisition results and such inferences do not consider the current user input any more. Typical approaches comprise logic-based methods, inductive reasoning, neural networks, and machine-learning techniques.

2.1.3 Personalized output

This module determines the kind(s) of adaptations deployed by a web application in order to personalize itself. Adaptations can take place at different levels:

- **Content**: Typical applications of such adaptations are optional explanations and additional information, personalized recommendations, theory driven presentation, and more. Techniques used for producing such adaptations include adaptive selection of web page (or page fragment) variants, fragment coloring, adaptive stretch-text, and adaptive natural language generation.
- **Structure**: It refers to changes in the link structure of hypermedia documents or their presentation. Techniques deployed for producing this kind of adaptations comprise adaptive link sorting, annotation, hiding and unhiding, disabling and enabling, removal/addition. Adaptations of structure are widely used for producing adaptive recommendations (for products, information or navigation), as well as constructing personal views and spaces.
- **Presentation and media format**: in this type of personalized output the informational content ideally stays the same, but its format and layout changes (for example from images to text, form text to audio, from video to still images). This type of adaptations is widely used for web access through PDAs or mobile phones, or in websites that cater to handicapped persons.

2.2 Threats to privacy

Personalized websites need by definition a way of identifying users in repeating visits so as to keep user profiles updated and consistent and maintain user history records. To accomplish this objective the typical approach is a login process with the user being associated with his profile data upon entering username and password. Another solution used is cookie files but they do not suffice when users enter the website form different machines (which today is quite trivial, i.e. from internet cafes). The first user privacy threat connected with personalization stems from the requirement of identifying the user and thus denying the right to (or wish for) anonymity that most web users value.

After having reviewed the major personalization modules, one easily concludes that threats to privacy come from the first module, namely data acquisition for user profiling purposes. And

while the more "raw" forms of personalization (i.e. check box customisation) are constrained to collecting only a limited and well-defined set of data that are explicitly inputted by the user (static user profile), moving on to more intelligent and dynamic personalization scenarios (with dynamic user profiling) brings in more hazards and threats for user privacy. In other words, privacy is not threatened if a website asks directly the user for the personal information it needs and the user inputs the information if he chooses so. In this case, the user is in control of the whole data disclosure process and knows exactly what information the website keeps on him. Privacy is further assured in this case if the user is also allowed to access at any time in the future his data and make any change he wishes. Finally, it is important for the user to be reassured that none of that information is revealed, or sold to third parties (it is a usual case for today's web to receive a huge amount of trash e-mail as a result of a website disclosing our e-mail to others that use it for reaching us with various motives).

Even if the situation described so far can be controlled so as not to present real threats to privacy –at least no additional threats as comprised to trivial non-personalised navigation– the real problem arises in the cases where the user is monitored and recorded on a click-by-click basis, so that his behaviour may be analysed for producing more "sophisticated" personalization. In this case, on one hand the user is not intruded when navigating the site and on the other, the site has a rich data repository that can be used for discovering new knowledge about the users and adjusting itself in a way that better suits them individually.

Both arguments come from the site owner point of view in an attempt to favour or justify dynamic personalization scenarios where the site changes dynamically as a result of user monitoring and internal –transparent and unknown– processing and reasoning. The user in many such cases does not know neither that he is being watched, nor on what assumptions the website decides to base the personalized experience that will be served to him. The website may know when the certain user entered a page, from which URL he entered it, which IP he used, how long he stayed in it, what link(s) he activated, whether he scrolled or not, and the complete history of each visit to the site. A very common policy is to use the personal history for identifying groups of like-minded users and use this integrated behaviour pattern to determine the adaptations the site will take up in order to personalize the group's web experience.

Another potential privacy hazard derives form some forms of personalized output such as product recommendations where the user receives personalized promotion material that depends on his buying history and/or product correlations. This push marketing philosophy may threaten the user privacy since in many cases he has not been informed about or consented to it.

3. Technological solutions

The need for user privacy has created a new market of products designed to protect his personal information. These technological solutions, varying from managing cookies to generating and analyzing online privacy policies, can help users to recapture control over their sensitive data. The following table summarizes the most well-known of these tools [6], [2].

Tools for managing cookies:	Burnt Cookies, Buzof, Complete Cleanup, Cookie Cutter, Cookie Crusher, Cookie Editor, Cookie Master, Cookie Monster 3.00, Cookie Pal, Cookie Terminator, Cookie Webkit, Crystal Clean, IEClean & NSClean,

	MagicCookie Monster, NoCookie, PGPcookie.cutter, Spy Blocker.
Tools for surfing anonymously through proxy servers:	Anonymity 4 Proxy, Anonymizer, Basic SignupShield, Bypass Proxy Client, GetAnonymous, GhostSurf Pro, Internet Junkbuster Proxy, Naviscope, Primedius Total Privacy & Security, PrivadaProxy, ProxyChecker, ProxyMate.
Tools for encrypting e-mail:	Authora Zendit, Disappearing E-mail, HushMail, Magic Mail Maker, PEM, PGP, Privateimail, SecExMail Secure Email, ZipLip Mail.
Tools for blocking unwanted files:	AdSubtract SE, IDcide Privacy Companion.
Tools for cleaning residual files:	Internet Guard Dog, Window Washer.
Tools for managing user's identity:	Digitalme, Freedom, PersonalChild Persona.
Tools for purchasing anonymously:	iPrivacy, ZixCharge.
Tools for maintaining user's firewall:	3B Personal Firewall Pro, Agnitum Outpost Firewall Pro, AntiFirewall Anonymizer, Armor2net Personal Firewall, BitGuard Firewall, Fireball CyberProtection Suite, GoldTach Pro, HTTP-Tunnel, McAfee Firewall, My Firewall Plus, Norton Internet Security 2004, Personal Firewall, Sygate Personal Firewall, ZoneAlarm Pro 4.
Tools for searching the Internet privately:	TopClick Private Web Search.
Tools for generating and analyzing privacy policies:	DMA's Privacy Policy Generator, Enonymous Advisor, Privacy Wizard, OECD Privacy Policy Generator, P3Pwriter Privacy Policy Editor, Policy Editor.
Organizations for privacy seals:	BBBonline, CPA WebTrust, TRUSTe

Table 1: Technological solutions for protecting user's privacy.

Especially P3P is a significant attempt to this direction and was developed by the World Wide Web Consortium (W3C) in 1999 [13]. P3P is a standard, which provides a simple and automated way for users to gain more control over their personal information when visiting websites. At its most basic level, P3P is a standardized set of multiple-choice questions, covering all the major aspects of a website's privacy policies. Taken together, they present a clear snapshot of how a site handles personal information about its users. P3P-enabled websites make this information available in a standard, machine-readable format that P3P-enabled browsers can "read" automatically and compare it to the user's own set of privacy preferences. P3P enhances user control by putting privacy policies where users can find them, in a form users can understand, and, most importantly, enables users to act on what they see.

Many websites provide a privacy statement or a P3P policy that the user can view with a browser. P3P helps protect the privacy of user's personal information on the Internet by simplifying the process for deciding whether and under what circumstances personal information is disclosed to websites. However, while P3P provides a standard mechanism for describing privacy practices and ensuring that users can be informed about privacy policies before they give personal information, it does not set a privacy standard which websites must follow. This means that it does not ensure that the websites will act according to their policies.

In Internet Explorer for instance the user can define his/her privacy preferences for handling cookies. So, when he browses to websites, Internet Explorer determines whether the sites provide P3P privacy information. For sites that provide this information, the browser compares user's privacy preferences to the site's privacy policy information. In this manner, Internet Explorer decides whether to allow cookies or restrict them. As an example, the user can choose to block cookies which use personally-identifiable information without his/her clear consent. A P3P-compliant website must provide a clear definition of its privacy policies.

Consequently, although the new technologies and products for protecting user's privacy on computers and networks are becoming increasingly popular, none can guarantee absolutely secure communications. Electronic privacy issues in the foreseeable future will become highly crucial and intense.

4. Conclusions

On the road to enhance a website's user experience and treat each user individually, personalization plays a central role. The benefits for both users and website owners are significant when personalization really works. However, understanding the online user is not an easy or straightforward process [9]. Personalization identifies the user, records the user's online behaviour in as much detail as possible and extracts needs and preferences in a way that the user cannot notice, understand or control. The big challenge though remains the lack of trust on the user side and the lack of user data in the website side to base the personalization decisions upon.

Online user behaviour recording and prediction is a very hard task. As web surfers are increasingly extending their online experience, they become suspicious to any transparent logging or recording processes in progress. On the other hand, they become more demanding on special services focused on their individual needs. Research in the area of mining web usage data [19] is accompanied by security preservation methods to increase users' confidence while interacting with an online application.

From the above, the arisen question is how to protect people from the misuse of personal information on the web. This need for web user's privacy has created a new market dedicated to the design and development of products for protecting information privacy. There are many products for this purpose as indicated in section 3. Especially P3P proposed by W3C is expected to have significant impact since it provides the framework on which to build privacy mechanisms, although it does not ensure yet that the website will act according to the stated privacy policy.

Another challenge of the area is to find exactly "what really do web users want?". Ease of use and no charge are their main requirements, since they do not want to have to pay for their privacy. However, in many cases they do not realise the threats and the potential dangers. This fact prevents them from taking the necessary precautions for their privacy and security protection.

Furthermore, some privacy advocates believe that the following practices should be followed:

- Websites that collect personal information must declare explicitly their practices to the users and also ensure that the data will not be used without authorization or for other purposes.
- Users must have the choice to accept or refuse the utilization of their personal information and they should also be able to access the collected information about them and check it for accuracy and correctness.

The future challenges and research in the direction of delivering web personalization without jeopardising –but in fact protecting- privacy relate to [7]: P3P support, intelligible disclosure of data, disclosure of methods, provision of organizational and technical means for users to modify their user model entries, user model servers that support a number of anonymization methods, and adapting user modelling methods to privacy preferences and legislation.

In conclusion, privacy is the key to deploy web personalization and allow it to reach its full potential. The threats for user privacy are so many that a single solution does not exist. The powerful and intelligent technologies that personalization and web mining have brought into the picture may well provide for protecting and reassuring privacy, rather than jeopardising it.

5. References

[1] Blom, J. (2000), Personalization - A Taxonomy, Proceedings of the CHI 2000 Workshop on Designing Interactive Systems for 1-to-1 Ecommerce, New York, N.Y.: ACM, online at http://www.zurich.ibm.com/~mrs/chi2000/, accesses 15.1.2004.

[2] Cavoukian, A. and Crompton, M. (2000), Web Seals: A Review of Online Privacy Programs, A Joint Project of The Office of the Information and Privacy Commissioner/Ontario and The Office of the Federal Privacy Commissioner of Australia, 22nd International Conference on Privacy and Personal Data Protection, Venice, Italy, online at http://www.privacy.gov.au/publications/seals.html, accessed 15.1.2004.

[3] Eirinaki, M. and Vazirgiannis, M. (2003), Web Mining for Web Personalization, ACM Transactions on Internet Technology (TOIT), Feb 2003, ACM Press New York, Vol. 3, No. 1, pp. 1-27.

[4] Furnell, S.M. and Karweni, T. (1999), Security Implications of Electronic Commerce: a Survey of Consumers and Businesses, Internet Research, Vol. 9, No. 5, pp. 372-382.

[5] GVU – Graphics, Visualization and Usability Center of the Georgia Institute of Technology (1996), 6th WWW User Survey, online at http://www.gvu.gatech.edu/user_surveys/survey-10-1996/#highsum, accessed 15.1.2004.

[6] Information Technology Industry Council (2000), Personal Privacy Solutions, A Look at Privacy Enhancing Technologies Available to Consumers, online at http://www.itic.org/digital_frontier/consumer/intro.html, accessed 15.12.2003.

[7] Kobsa, A. (2001), Tailoring Privacy to Users' Needs (Invited Keynote), In M. Bauer, P. J. Gmytrasiewicz and J. Vassileva, eds., User Modeling 2001, 8th International Conference, Berlin - Heidelberg: Springer Verlag, pp. 303-313, online at http://www.ics.uci.edu/~kobsa/papers/2001-UM01-kobsa.pdf, accessed 15.1.2004.

[8] Liu, C. and Arnett, K. (2002), An Examination of Privacy Policies in Fortune 500 Web Sites, Mid-American Journal of Business, Vol. 17, No. 1, pp.13.

[9] Markellos, K., Markellou, P., Rigou, M., Sirmakessis, S. and Tsakalidis, A. (2002), Who is Today's E-Customer? A Description of his Behavioral Model, Proceedings of the eBusiness and eWork 2002 Conference and Exhibition, Prague, Czech Republic.

[10] Markellou, P., Rigou, M. and Sirmakessis, S. (2004), Mining for Web Personalization, In Anthony Scime (Ed.), Web Mining: Applications and Techniques, Idea Group Publishing Inc., in press.

[11] Mobasher, B. and Dai, H. (2003), A Road Map to More Effective Web Personalization: Integrating Domain Knowledge with Web Usage Mining, Proceedings of the International Conference on Internet Computing 2003 (IC'03), Las Vegas, Nevada.

[12] Mulvenna, M., Anand, S. and Bchner, A. (2000), Personalization on the Net Using Web Mining, Communications of the ACM, Vol. 43, No. 8, pp. 122-125.

[13] P3P, Platform for Privacy Preferences Project, online at http://www.w3.org/P3P, accessed 15.1.2004.

[14] Perner, P. and Fiss, G. (2002), Intelligent E-marketing with Web Mining, Personalization, and User-Adapted Interfaces, P. Perner (Ed.): Advances in Data Mining 2002, LNAI 2394, 37-52, 2002, Springer-Verlag Berlin Heidelberg 2002.

[15] Personalization Consortium, online at http://www.personalization.org, accessed 15.1.2004.

[16] Privacilla.org, Privacilla's Two-Part Definition of Privacy, online at http://www.privacilla.org/fundamentals/privacydefinition.html, accessed 15.1.2004.

[17] Riecken, D. (2000), Personalized Views of Personalization, Communications of the ACM, August 2000, Vol. 43, No. 8, pp. 27-28.

[18] Schafer, J., Konstan, J. and Riedl, J. (2001), E-commerce Recommendation Applications, Data Mining and Knowledge Discovery, Vol. 5, No. 1, pp. 115-153.

[19] Sirmakessis, S. (Ed.) (2004), Text Mining and its Applications, Studies in Fuzziness and Soft Computing, Springer Verlag.

[20] Teltzrow, M. and Kobsa, A. (2004), Impacts of User Privacy Preferences on Personalized Systems: a Comparative Study, In C.-M. Karat, J. Blom and J. Karat, eds, Designing Personalized User Experiences for eCommerce, Dordrecht, Netherlands, Kluwer Academic Publishers, online at http://www.ics.uci.edu/~kobsa/papers/2004-PersUXinECom-kobsa.pdf, accessed 15.1.2004.

[21] Vassiliou, C., Stamoulis, D. and Martakos, D. (2002), The Process of Personalizing Web Content: Techniques, Workflow and Evaluation, Proceedings of the International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet January 21 2002, L'Aquila, Italy, online at http://www.ssgrr.it/en/ssgrr2002s/papers.htm, accessed 15.1.2004.

[22] Volokh, E. (2000), Personalization and Privacy, Communications of the ACM, Vol. 43, No. 8, pp. 84-88, August 2000.