Respecting Users' Individual Privacy Constraints in Web Personalization[1]

Yang Wang and Alfred Kobsa

Donald Bren School of Information and Computer Sciences
University of California, Irvine, U.S.A.
{yangwang, kobsa}@ics.uci.edu

Abstract. Web personalization has demonstrated to be advantageous for both online customers and vendors. However, its benefits may be severely counteracted by privacy constraints. Personalized systems need to take users' privacy concerns into account, as well as privacy laws and industry self-regulation that may be in effect. In this paper, we first discuss how these constraints may affect web-based personalized systems. We then explain in what way current approaches to this problem fall short of their aims, specifically regarding the need to tailor privacy to the constraints of each individual user. We present a dynamic privacy-enhancing user modeling framework as a superior alternative, which is based on a software product line architecture. Our system dynamically selects personalization methods during runtime that respect users' current privacy concerns as well as the privacy laws and regulations that apply to them.

1 Introduction

Numerous consumer studies and lab experiments (see [17, 22] for an overview) suggest that privacy concerns may prompt people to withhold information about themselves when interacting with personalized systems, thereby preventing users to fully benefit from the potential of personalization. These studies also show that people's privacy preferences differ to some extent. Since personalized websites collect personal data, they are also subject to prevailing privacy laws and regulations if the respective individuals are in principle identifiable. As we will show below, such laws often not only affect the data that are collected by the website, but also the personalization methods that may be used for processing them.

In this paper, we will investigate how personalized web-based systems can be compliant with the privacy constraints that are currently in effect for each individual user (namely privacy laws, industry and company regulations, and privacy preferences of every user). We propose a novel approach based on software product lines that allow the configuration of the employed personalization methods to be tailored to each user's privacy constraints. We will first analyze how such privacy constraints may affect the admissibility of personalization methods, both with regard to individual privacy concerns and privacy laws. We then review existing approaches

[1] This research has been supported through NSF grant IIS 0308277. We would like to thank André van der Hoek, Eric Dashofy and the UM07 reviewers for their helpful comments.

for handling the differences in privacy constraints that apply to different users, and analyze their shortcomings. Thereafter we present our software product line approach in Section 4, an illustrative example for its operation in Section 5, and conclusions and future work in Section 6.

2 Impacts of Privacy Constraints on Web Personalization

2.1 Impacts of Users' Privacy Concerns

Numerous opinion polls and empirical studies have revealed that Internet users harbor considerable privacy concerns regarding the disclosure of their personal data to websites, and the monitoring of their Internet activities. These studies were primarily conducted between 1998 and 2003, mostly in the United States. In the following, we summarize a number of important findings (the percentage figures indicate the ratio of respondents from multiple studies who endorsed the respective view). For more detailed discussions we refer to [17, 22].

Personal data.

1. Internet users who are concerned about the privacy or security of their personal information online: 70% - 89.5%;
2. People who have refused to give personal information to a web site at one time or another: 82% - 95%;
3. Internet users who would never provide personal information to a web site: 27%;
4. Internet users who supplied false or fictitious information to a web site when asked to register: 6% - 40% always, 7% often, 17% sometimes;
5. People who are concerned if a business shares their data for a different than the original purpose: 89% - 90%.

Significant concern over the use of personal data is visible in these results, which may cause problems for all personalized systems that depend on users disclosing data about themselves. False or fictitious entries when asked to register at a website make all personalization based on such data dubious, and may also jeopardize cross-session identification of users as well as all personalization based thereon. The fact that 80-90% of respondents are concerned if a business shares their information for a different than the original purpose may have impacts on central user modeling servers (UMSs) [16] that collect data from, and share them with, different user-adaptive applications.

User tracking and cookies.

1. People concerned about being tracked on the Internet: 54% - 63%;
2. People concerned that someone might know their browsing history: 31%;
3. Users who feel uncomfortable being tracked across multiple web sites: 91%;
4. Internet users who generally accept cookies: 62%;
5. Internet users who set their computers to reject cookies: 10% - 25%;
6. Internet users who delete cookies periodically: 53%.

These results reveal significant user concerns about tracking and cookies, which may have effects on the acceptance of personalization that is based on usage logs. Observations 4–6 directly affect machine-learning methods that operate on user log data since without cookies or registration, different sessions of the same user can no longer be linked. Observation 3 may again affect the acceptance of the central user modeling systems which collect user information from several websites.

Kobsa [17] suggests that developers of personalized system should however not feel discouraged by the abundance of stated privacy concerns and their potential adverse impact on personalized systems. Rather, they should incorporate a number of mitigating factors into their designs that have been shown to encourage users' disclosure of personal data. Such factors include perceived value of personalization, previous positive experience, the presence of a privacy seal, catering to individuals' privacy concern, etc. The approach proposed here addresses this last factor.

2.2 Impacts of Privacy Laws and Regulations

Privacy Laws. Legal privacy requirements lay out organizational and technical requirements for information systems that store and/or process personal data, in order to ensure the protection of these data. Those requirements include proper data acquisition, notification about the purpose of use, permissible data transfer (e.g., to third parties and/or across national borders) and permissible data processing (e.g., organization, modification and destruction). Other provisions specify user opt-ins (e.g., asking for their consent before collecting their data), opt-out, users' rights (e.g., regarding the disclosure of the processed data), adequate security mechanisms (e.g., access control), and the supervision and audit of personal data processing.

Our review of over 40 international privacy laws [24] shows that if such laws apply to a personalized website, they often not only affect the data that is collected by the website and the way in which data is transferred, but also the personalization methods that may be used for processing them. The following are some example codes:

1. *Value-added* (e.g. personalized) *services based on traffic or location data require the anonymization of such data or the user's consent* [9]. This clause clearly requires the user's consent for any personalization based on interaction logs if the user can be identified.
2. *The service provider must inform the user of the type of data which will be processed, of the purposes and duration of the processing and whether the data will be transmitted to a third party, prior to obtaining her consent* [9]. It is sometimes fairly difficult for personalized service providers to specify beforehand the particular personalized services that an individual user would receive. The common practice is to collect as much data about the user as possible, to lay them in stock, and then to apply those personalization methods that "fire" based on the existing data.
3. *Users must be able to withdraw their consent to the processing of traffic and location data at any time* [9]. In a strict interpretation, this stipulation requires personalized systems to terminate all traffic or location based personalization

immediately when asked, i.e. even during the current service. A case can probably be made that users should not only be able to make all-or-none decisions, but also decisions on individual aspects of traffic or location based personalization.

4. *Personal data must be collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes* [8]. This limitation would impact central UMSs, which store user information from, and supply the data to, different personalized applications. A UMS must not supply data to personalized applications if they intend to use those data for different purposes than the one for which the data was originally collected.
5. *Usage data must be erased immediately after each session* (except for very limited purposes) [7]. This provision could affect the use of machine learning methods when the learning takes place over several sessions.

Company and Industry Regulations. Many companies have internal guidelines in place for dealing with personal data. There also exist a number of voluntary privacy standards to which companies can subject themselves (e.g., of the Direct Marketing Association, the Online Privacy Alliance, the U.S. Network Advertising Initiative, the Personalization Consortium, and increasingly the TRUSTe privacy seal program).

3 Existing Approaches to Address the Variability of Privacy Constraints

No systematic approach has so far existed for building websites that cater to the different privacy constraints of different users. Sites that aimed at addressing this problem had to use simple escape strategies, which we list below.

Pseudonymous Personalization. Basically, this approach allows users to remain anonymous with regard to the personalized system and the whole network infrastructure, whilst enabling the system to still recognize the same user in different sessions and cater to her individually [19]. At first sight, this seems to be a panacea because in most cases privacy laws do not apply any more when the interaction is anonymous. However, anonymity is currently difficult and/or tedious to preserve when payments, physical goods and non-electronic services are being exchanged. It harbors the risk of misuse, and it hinders vendors from cross-channel marketing (e.g. sending a product catalog to a web customer by mail). Moreover, users may still have additional privacy preferences such as not wanting to be profiled even when this is done merely pseudonymously, to which personalized systems need to adjust.

Largest Permissible Dominator. Ideally, this approach means that only those personalization methods that meet all privacy laws and regulations of all website visitors are used. The Disney website, for instance, meets the European Union Data Protection Directive [8] as well as the U.S. Children's Online Privacy Protection Act (COPPA) [1]. This solution is likely to run into problems if more than a very few jurisdictions are involved, since the largest permissible denominator may then become very small. Individual user privacy concerns are also not taken into account.

Different Country/Region Versions. In this approach, personalized systems have different country versions, with personalization methods only that are admissible in the respective country. If countries have similar privacy laws, combined versions can be built for them using the above-described largest permissible denominator approach. For example, IBM's German-language pages meet the privacy laws of Germany, Austria and Switzerland [2], while IBM's U.S. site meets the legal constraints of the U.S. only. This approach is also likely to become infeasible as soon as the number of countries/regions, and hence the number of different versions of the personalized system, increases. Individual user privacy concerns are also not taken into account.

P3P. The Platform for Privacy Preferences (P3P) [25] enables websites to express their privacy policies in a standard machine-readable format that can be retrieved automatically and interpreted by user agents. Client-side agents can then inform users about the sites' privacy policies and warn them when those deviate from previously-specified preferences. P3P does not enforce privacy policies nor does it support different policies for different users. By itself, it is therefore not an answer to the need for privacy tailored to different user constraints. However, several proposals for individual negotiation of P3P policies have been made [5, 20]. The results of such negotiations could become the input to our own approach.

4 A Dynamic Privacy-Enhancing User Modeling Framework

User Modeling Servers (UMSs) store and represent user characteristics and behavior, integrate external user-related information, apply user modeling methods to derive additional assumptions about the user, and allow multiple external user adaptive applications to retrieve user information from the server in parallel [16]. UMSs are widely used for supporting user-adaptive applications. Our solution enhances a regular UMS by a new dimension of personalization, namely adaptation to each user's potentially different privacy constraints.

For many personalization goals, more than one method can often be used that differ in their data and privacy requirements and their anticipated accuracy and reliability. For example, a personalized website could use incremental machine learning to provide personalization to visitors from Germany (where user logs must be discarded at the end of a session to comply with Code 5 in Section 2.2), while it can use possibly better one-time machine learning with user data from several sessions to provide personalization to web visitors from the U.S. who are not subject to this constraint.

We propose a software architecture that encapsulates different personalization methods in individual components and, at any point during runtime, ascertains that only those components can be operational that are in compliance with the currently prevailing privacy constraints. Moreover, the architecture can also dynamically select the component with the optimal anticipated personalization effects among those that are currently permissible [15]. To implement this design, we utilize a product line approach from software architecture research and, simplistically speaking, give every user their own UMS instance which incorporates those user modeling methods only that meet the user's current privacy constraints [23].

Product Line Architectures (PLAs) have been successfully used in industrial software development [4]. A PLA represents the architectural structure for a set of related products by defining core elements that are present in all product architectures, and variation points where differences between individual product architectures may occur. Each variation point is guarded with a Boolean expression that represents the conditions under which an optional component should be included in a particular product instance. A product instance can be selected out of a product line architecture by resolving the Boolean guards of each variation point at design-time, invocation-time or run-time [13].

Figure 1 shows an overview of our PLA-based user modeling framework. It consists of external user-adaptive applications, the Selector, and the LDAP-based UMS of Kobsa and Fink [18] which includes the Directory Component and a pool of user modeling components (UMCs). External personalized applications can query the UMS for existing user information, so as to provide personalized services to their end users, and can supply additional user information to the UMS. The Directory Component is essentially a repository of user models, each of which stores and represents not only users' characteristics, behavior and inferences, but also their potentially different individual privacy constraints. The UMC Pool contains a set of UMCs, each of which encapsulates one or more user modeling methods (e.g., collaborative filtering) that make inferences about users based on existing user data.

The novel privacy enhancement consists in every user having their own instance of the UMC Pool, each containing only those user modeling components that meet the privacy requirements for the respective user (users with identical UMC Pool instances

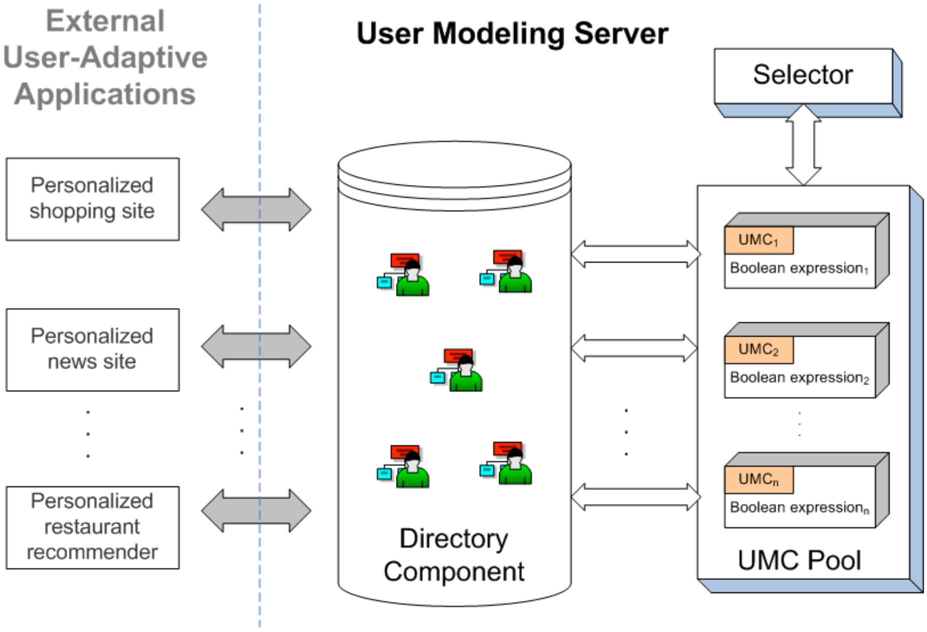

Fig. 1. A Dynamic Privacy-Enabling User Modeling Framework

share the same instance). To realize this, the above framework has been implemented as a PLA, with the UMCs as optional elements [14] guarded by a Boolean expression that represents privacy conditions under which the respective UMC may operate (e.g. "`(CombineProfile == allowed) && (TrackUser == allowed)`").

At the beginning of the interaction with a user, the Selector verifies for every UMC whether it may operate under the privacy constraints that apply to the specific user, and creates an architectural instance with these permissible UMCs (or lets the user share this instance if one already exists). Moreover, in order to maximize the benefits of personalization, the Selector can further select the UMCs with the optimal anticipated personalization effects among those that are currently permissible based on a designer-specified preference order. The PLA management environment that we employ [3] supports dynamic runtime (re-)configuration, which allows the Selector to react immediately, e.g., users change their privacy preferences during the current session. The framework therefore allows a personalized website to adjust its data practices to the user's preferences in a nuanced and highly dynamic manner. The fact that if two or more users have the same set of privacy constraints they will share a single personalization architecture is key to the scalability of our solution.

5 An Illustrative Example

Assume that MyTaste is a mobile web service that provides restaurant recommendations worldwide based on customers' current location (collected from their GPS-embedded mobile devices), their food preferences and demographics as well as the proximity of nearby restaurants and their ratings by other customers. Upon registration, users will be asked to disclose their identities and optionally disclose some information about themselves (e.g., their food preferences). The system will then automatically retrieve their demographics from commercial databases or credit bureaus. The system also encourages users to rate places they have patronized, by offering discounts for restaurants that will be recommended to them in the future. The processing of all personal data is described in a privacy statement, i.e. the disclosure duties of Code 2 in Section 2.2 are being met.

The MyTaste web server relies on our privacy-enabling user modeling framework to infer information about users to provide recommendations. Table 1 summarizes the usage of data and inference methods for each user modeling component. For example, UMC_1 can recommend restaurants based on ratings of people in the same nationality cluster. If a user indicates a high interest in a specific type of food, UMC_2 can recommend nearby restaurants that have good ratings in this category.

We have three hypothetical adult users, Alice from Germany, Cheng from China, and Bob from the U.S. Cheng dislikes being tracked online, while Alice and Bob do not express any privacy preferences. MyTaste.com can tailor its provided personalization to the different privacy constraints of these users in the following manner:

1. When users log into the website, the system gathers their current privacy constraints, namely those imposed by privacy laws and regulations as well as their personal privacy preferences. Users can specify their privacy preferences and change them anytime during the interaction with the personalized system.

Table 1. The UMC pool of MyTaste

UMC	Data used	Method used
UMC_1	– Demographic (such as age, gender, profession, nationality)	– Clustering techniques
UMC_2	– Food preferences	– Rule-based reasoning
UMC_3	– Demographic – Food preferences	– Rule-based reasoning
UMC_4	– Food preferences – Current session log (MyTaste pages that the user visited in the current session)	– Incremental machine learning
UMC_5	– Food preferences – Last n session log (MyTaste pages that the user visited across sessions)	– One-time machine learning across several sessions
UMC_6	– Demographic – Food preferences – Location data – Last n session log	– One-time machine learning across several sessions

2. Our framework determines which UMCs may operate for each user given their privacy constraints. For example, the German Teleservices Data Protection Act [7] and the EU Directive on Electronic Communications [9] apply to Alice, with the following consequences:
   - In the light of Code 4 in Section 2.2, UMC_1, UMC_3 and UMC_6 are illegal without Alice's consent because the demographic data that the website retrieves from commercial databases and credit bureaus had not been originally collected for personalization or recommendation purposes.
   - In the light of Code 5, UMC_5 and UMC_6 are illegal because they both use cross-session log data.
   - In the light of Code 1, UMC_6 is illegal without Alice's consent because it uses location data without anonymizing it.

   Hence UMC_1, UMC_3, UMC_5 and UMC_6 cannot be used for Alice without her explicit consent.
3. With similar analyses, the system can determine that UMC_4, UMC_5 and UMC_6 cannot be used for Cheng who does not like to be logged. No privacy restrictions apply to Bob.
4. The system will thus instantiate three different UMCs pools for these three users, i.e. each user will have his own instance of the personalized system that meets her current privacy constraints.

6 Conclusions and Future Work

Privacy constraints in the domain of web personalization derive from users' personal privacy preferences, privacy laws and regulations. These privacy constraints have substantial impacts on the ways in which web-based personalized systems may operate internally, and indirectly on how much personalization they are consequently able to provide. Existing approaches fall short of a flexible, systematic and scalable solution to respecting privacy constraints that may differ among users. Our proposed PLA-based user modeling framework allows personalized websites to address the combinatorial complexity of privacy constraints in a systematic and flexible manner, which builds on state-of-the-art industry practice for managing software variants at runtime. It should however not be misunderstood as a complete solution to all privacy issues in personalized web-based systems. Our approach focuses on the architectural aspects of user-tailored privacy provisioning but does not control (let alone enforce) what and how user data are or may be collected.

While we currently use Boolean variables to express identified privacy constraints [23], ultimately these constraints should be expressed in a privacy constraint specification language such as APPEL [6] or EPAL [21], or with semantic web technologies [10]. Unfortunately though none of these proposals has gained much impact so far. Future plans of P3P [25] include the support of privacy policy negotiation, whose results can be used as privacy constraints input to our system.

Conditions on the applicability of our constraints (e.g. the user's country) are currently fully "factored out", and nuances in the meanings of the same constraints in different contexts (e.g. countries) are currently represented by using different Boolean variables. It would be worthwhile to study the applicability of conditional constraints [11] and context-sensitive constraints [12], which allow for more compact representations and are also closer to the original legal phrasing.

Performance and scalability are of critical interest in practice, specifically if systems are expected to provide personalization services to hundreds of thousands of users from all over the world. We ran some basic performance experiments based on our current prototype [23]. The performance results imply that the overhead incurred by product line architecture is not negligible. We are currently experimenting with different ways of optimizing the architectural selection process. Fortunately though, since the number of privacy jurisdictions is limited (currently to about 40 countries and 100 states), we assume that many of our users will share the same architecture. The resource-intensive architecture selection and instantiation process is therefore likely not to be invoked too often. This reusability is key to performance and scalability, but its effects will need to be more thoroughly tested.

References

1. Personal Communication, Chief Privacy Officer, Disney Corporation. (2002).
2. Personal Communication, Chief Privacy Officer, IBM Zurich. (2003).
3. ArchStudio: ArchStudio 3.0. (2005). http://www.isr.uci.edu/projects/archstudio/.
4. Bosch, J.: Design and Use of Software Architectures: Adopting and Evolving a Product-Line Approach. New York: Addison-Wesley (2000).
5. Buffett, S., Jia, K., Liu, S., Spencer, B., and Wang, F.: Negotiating Exchanges of P3P-Labeled Information for Compensation. Computational Intelligence 20, (2004) 663-677.
6. Cranor, L., Langheinrich, M., and Marchiori, M.: A P3P Preference Exchange Language 1.0 (APPEL1.0): W3C Working Draft 15 April 2002.
7. German Teleservices Data Protection Act 1997, as amended on 14 Dec. 2001.

8. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of such Data. Official Journal of the European Communities, (1995) 31ff.
9. Directive 2002/58/EC of the European Parliament and of the Council Concerning the Processing of Personal Data and the Protection of Privacy in the Electronic Communications Sector. (2002)
10. Gandon, F. L. and Sadeh, N. M.: Semantic Web Technologies to Reconcile Privacy and Context Awareness. Journal of Web Semantics 1, (2004) 241-260.
11. Gelle, E. and Sabin, M.: Solving Methods for Conditional Constraint Satisfaction. The 8th International Joint Conference on Artificial Intelligence (IJCAI-03) (2003).
12. Gray, J., Bapty, T., Neema, S., and Tuck, J.: Handling crosscutting constraints in domain-specific modeling. Communications of the ACM 44, (2001) 87 - 93.
13. Hoek, A. v. d.: Design-Time Product Line Architectures for Any-Time Variability. Science of Computer Programming, special issue on Software Variability Management 53, (2004) 285-304.
14. Hoek, A. v. d., Mikic-Rakic, M., Roshandel, R., and Medvidovic, N.: Taming Architectural Evolution. The Sixth European Software Engineering Conference (ESEC) and the Ninth ACM SIGSOFT Symposium on the Foundations of Software Engineering (FSE-9), Vienna, Austria (2001) 1-10.
15. Kobsa, A.: A Component Architecture for Dynamically Managing Privacy in Personalized Web-based Systems. Privacy Enhancing Technologies: Third International Workshop, Dresden, Germany (2003) 177-188.
16. Kobsa, A.: Generic User Modeling Systems. In: The Adaptive Web: Methods and Strategies of Web Personalization, Brusilovsky, P., Kobsa, A., and Nejdl, W., Eds.: Heidelberg, Germany: Springer-Verlag (2007).
17. Kobsa, A.: Privacy-Enhanced Web Personalization. In: The Adaptive Web: Methods and Strategies of Web Personalization, Brusilovsky, P., Kobsa, A., and Nejdl, W., Eds.: Heidelberg, Germany: Springer-Verlag (2007).
18. Kobsa, A. and Fink, J.: An LDAP-Based User Modeling Server and its Evaluation. . User Modeling and User-Adapted Interaction: The Journal of Personalization Research 16, (2006) 129 - 169.
19. Kobsa, A. and Schreck, J.: Privacy through Pseudonymity in User-Adaptive Systems. ACM Transactions on Internet Technology 3, (2003) 149-183.
20. Preibusch, S.: Personalized Services with Negotiable Privacy Policies. PEP06, CHI 2006 Workshop on Privacy-Enhanced Personalization, Montreal, Canada (2006) 29-38.
21. Schunter, M. and Powers, C.: The Enterprise Privacy Authorization Language (EPAL 1.1): Reader's Guide to the Documentation. IBM Research Laboratory (2003).
22. Teltzrow, M. and Kobsa, A.: Impacts of User Privacy Preferences on Personalized Systems: a Comparative Study. In: Designing Personalized User Experiences for eCommerce, Karat, C.-M., Blom, J., and Karat, J., Eds. Dordrecht, Netherlands: Kluwer Academic Publishers (2004) 315-332.
23. Wang, Y., Kobsa, A., van der Hoek, A., and White, J.: PLA-based Runtime Dynamism in Support of Privacy-Enhanced Web Personalization. The 10th International Software Product Line Conference, Baltimore, MD (2006) 151-162.
24. Wang, Y., Zhaoqi, C., and Kobsa, A.: A Collection and Systematization of International Privacy Laws, with Special Consideration of Internationally Operating Personalized Websites. (2006). http://www.ics.uci.edu/~kobsa/privacy.
25. Wenning, R. and M. Schunter, Eds.: The Platform for Privacy Preferences 1.1 (P3P1.1) Specification: W3C Working Group Note (2006).