

On the Design of a Tool for Controlling Privacy in the WWW

Lucila Ishitani Gustavo M. C. Gama

Virgílio Augusto F. Almeida Dorgival Olavo Guedes Neto

Wagner Meira Jr.

Departamento de Ciência da Computação

Universidade Federal de Minas Gerais

31270-901 Belo Horizonte MG Brasil

{lucila, gmcgama, virgilio, dorgival, meira}@dcc.ufmg.br

Abstract

Concerns about users' privacy are currently attracting significant attention. Users usually do not approve of someone mining their actions and habits while they use the Web. On the other hand, having some information about the user's behavior is essential for providing personalized services. Thus, there is a clear demand for mechanisms that allow users to control their privacy without relying on the visited site's policies. In this paper we propose a tool that tells the user the level of privacy maintained throughout his/her interaction with a given site. By using our tool, the user can become aware of privacy violations and may use privacy preservation mechanisms.

1 Introduction

Nowadays, there is an increasing number of sites trying to adapt or personalize their services to the characteristics of the users. Personalization of services offers many advantages, like evaluation and improvement of a site design, or the possibility of cross-sales and up-sales in e-commerce. In this sense personalization is good to users, since they can achieve a higher level of satisfaction, and it is also good to service providers, since that satisfaction can lead to customer loyalty.

In order to provide personalized services, some information about the users must be stored and mined, and that is turning out to be a complex topic: are these information being used with the user's approval? If so, that is all right; if not, some kind of privacy invasion is taking place. Thus, we are facing a contradictory situation: how can web mining and privacy preservation coexist?

This paper proposes a tool that can be a solution for the conflict presented. It is structured as follows: Section 2 discusses the users privacy concerns and the technologies and mechanisms proposed to preserve privacy. Section 3 presents our solution. Section 4 presents some conclusions and outlines further work.

2 Privacy Concerns and Privacy-Preserving Mechanisms

A survey of the American public conducted by Louis Harris & Associates and Dr. Alan F. Westin of Columbia University, in 1998, found that 81% of the net users are concerned

about their personal privacy while online. Patrick Sullivan of Price Waterhouse stated: “The results of the survey, especially concerning meaningful, verifiable privacy policies, are made all the more important by the Federal Trade Commission’s recent report that only 14% of commercial websites in the U.S. tell consumers anything about the site’s information practices, and only about 2% have any clear privacy policies posted.” [5]

In another document [1] we can find the following: “A survey of privacy preferences shows that the vast majority of users are pragmatic in their privacy concerns and their opinions vary a great deal in what they consider to be important aspects of their privacy.”

David Sims [6] presents two interesting statements. The first one is from James Pitkow, of the Xerox Parc: “There’s a disparity between what people think can be logged, what can be logged, and what they think should be logged”. The second one is from Lori Fena of the Electronic Frontier Foundation, who said that “when users trust a site, they don’t often mind giving up their personal information — which they know is valuable to publishers and advertisers.”

From such statements, we can conclude that there are some complex questions related to privacy: what is it, what do users consider privacy invasion, and which mechanisms and technologies can be offered to preserve users’ privacy?

Wang *et al.* [2] present a privacy definition in the context of e-commerce: “privacy usually refers to personal information and the invasion of privacy is usually interpreted as the unauthorized collection, disclosure, or other use of personal information as a direct result of electronic commerce transactions.” In this definition, the use of the adjective *unauthorized* grabs our attention. In other words, the **authorized** collection, disclosure and use of the same set of personal information is not considered privacy violation.

In order to try to solve the problem of privacy preservation, the following mechanisms and technologies have been proposed and used:

1. **Encryption tools**

In general, such tools encrypt only the content of a message. So, there is not really complete privacy protection, as servers still have access to information like the IP address of the user, the time of interaction and the URL requested.

2. **Filters**

Filters cut out cookies and banner advertisements. By doing that they protect the user, but they do not prevent the use of other information such as that listed above.

3. **Tools for anonymity and pseudonymity**

In these tools users adopt a pseudonym, individual or collective, that will be used to interact with the sites. The idea is to protect the real identity of the user in order to prevent privacy invasion. In some implementations, pseudonyms are directly related to a unique user or group of users but, in general, this direct association is not provided. If that is the case, there is no reasonable condition for the site to do any personalization, even if the user so desires.

4. **Rules to control the use and collection of user information**

This is the basic idea of the Platform for Privacy Preference Project (P3P) of the World Wide Web Consortium (W3C) [4]. P3P allows web sites to negotiate with the user which information will be collected and how and for what it will be used. This negotiation can be done automatically by user agents, so that users will not have to perform it directly for every site they visit. Doubts can be solved through user interaction. This mechanism has many disadvantages:

- (a) There are many privacy laws and regulations all over the world and it will be very difficult to unify all these laws.
- (b) Although P3P provides a technical mechanism for ensuring that users are informed about privacy policies before they release information, it does not provide a technical mechanism for making sure that sites act according to their policies [4]. That means that users will have to trust the sites they negotiate with.
- (c) The purpose of any data collection will have to be clear to the users, and the data can be used only for that stated purpose. However, “since data mining is based on the extraction of unknown patterns from a database, data mining engines do not (cannot) know at the outset, what personal data will be of value or what relationships will emerge. Therefore, identifying a primary purpose at the beginning of the process, and restricting one’s use of the data to that purpose are the antithesis of a data mining exercise” [7].
- (d) Sometimes it is difficult for the user to understand which information could be collected before his/her privacy becomes an issue. So we can end up being faced with the extreme situation of having every user prohibiting the collection of any information. If this occurs, how can service personalization be offered?

Thus, there is a clear demand for mechanisms that allow users to control their privacy without relying on the visited site policies. With our tool we believe we start to address this issue and to offer a reasonable solution for the problem.

3 A Tool for Controlling Privacy

A tool for controlling privacy must consider the following problems:

1. Users do not like the idea of having their privacy violated, but they may agree to give some information if they trust the site they are visiting or if they conclude there are advantages in doing that.
2. Many users do not understand exactly which information is sensitive, nor how and for what goals it can be used, so they need a tool that can “take care” of them.
3. In order to offer personalized services a site’s server must have access to at least some information about the users.

The solution we propose is a tool that is able to discover when users’ privacy might be invaded, and at what level the invasion is occurring (from this point on, we will call this tool the Guardian). We can define the Guardian as a new entity which should be inserted between the user and the server, whose purpose is to verify all user transactions. By analyzing the amount and the type of information given by the user and the answers received, this tool will be able to dynamically determine the level of privacy protection achieved. Another activity the Guardian will perform is to monitor URLs, to verify if any personal information is appearing in URLs which are being transmitted; such kind of “information leak” has been discussed in the literature [3]. The tool will also tell the user the level of privacy during his/her interaction with a given site. Finally, at any moment, the user will be able to choose when privacy should be protected through traffic anonymization and when the interaction should continue without protection, but with some gains. For

example, the advantage of receiving a personalized service or receiving discounts and promotions offered by virtual stores which are only available if users continue an interaction without anonymization.

In order for this tool to become real the following issues must be addressed:

1. design an interface that is simple and easy to use and understand;
2. identify mechanisms to determine the level of personalization information available in any interaction over the Web.

3.1 Interface design

Our idea is to implement something like a speedometer or a thermometer to indicate the level of privacy invasion. Besides the graphic representation, the user will be able to select two buttons: EXPLAIN and PROTECT. The first one will have the function of giving some explanations about what is happening in case the tool considers that privacy is being violated. The second one will have the purpose of initiating a protection scheme, through anonymization of the client's traffic.

3.2 Mechanisms to identify when personalization occurs

To address this issue, the tool will have to be able to analyze whether the server has enough information to personalize the service. For example, the most simple task is to verify if cookies are being stored. Another task the tool will have to do is to use algorithms to predict user clicks. If the tool is able to succeed in this task, certainly the server will also be able, mainly because it has access to an even larger amount of information. Since the server has more information, it will be interesting for the tool to use other mechanisms to find whether personalization is taking place. The mechanism we propose is for the guardian to make the same requests as the user, but using anonymization. If the two answers received are different, so we can conclude that some personalization occurred.

To evaluate the possibility of predicting users clicks, we performed some experiments based on user interaction [8]. In the article, the authors presented four prediction models: Time Markov, Space Markov, Second-order Time Markov and Linked Space-Time Markov. The first two models consider only temporal information, in other words, the last one or two documents requested. The Space Markov model considers the structure of the site. The last model considers temporal and structural information. According to the authors the two best models (models that predict with the highest accuracy) are the Space Markov model and the Linked Space-Time Markov model. Unfortunately, we were not able to run tests using those two models, since they depend on the knowledge of the site's structure, and none of our logs had that information.

One of that article's failures is that it does not present a comparison of the models based on correct predictions. To address that we decided to perform some tests to have some real results. Since we are working in the e-commerce context, we chose to use logs from a virtual store. In this context, not only the url is important, but also the parameters used on searches, add-to-cart operations, and so on. The problem is that when we consider all that information the number of vertices in the Markov model can rapidly explode. Therefore, instead of using all of the 100,000 sessions we had, we selected the sessions in which we could find at least one of the most common search terms. As a result, we got a set of 10,426 sessions. Using this set, we ran experiments using the following models:

1. Time Markov model, considering only the last operation selected by the user.
2. Time Markov model, considering the operations' parameters.
3. Time Markov model, including the search parameters.
4. Second-Order Time Markov model, considering only the last two operations selected by the user.
5. Second-Order Time Markov model, considering the last two operations selected by the user, and their parameters.
6. Second-Order Time Markov model, considering the last two operations selected by the user, their parameters, and also the search parameters.

Model	Results			
	% right pred.	% wrong pred.	% unknown	% unpredictable
1	37.85	40.87	0.00	21.38
2	13.84	19.31	0.19	66.67
3	0.46	3.59	3.96	92.00
4	24.92	16.19	0.00	58.89
5	13.72	16.35	0.17	69.76
6	2.76	12.18	4.44	80.62

Table 1: Minimum prediction probability = 60%

Model	Results			
	% right pred.	% wrong pred.	% unknown	% unpredictable
1	22.41	23.57	0.00	54.03
2	11.31	18.21	0.19	70.29
3	0.25	2.86	3.96	92.93
4	18.04	9.69	0.00	72.27
5	8.20	12.45	0.17	79.18
6	1.01	9.47	4.44	85.07

Table 2: Minimum prediction probability = 80%

Each experiment was repeated 28 times, considering the amount of information used to construct the Markov model and the minimum prediction probability. Tables 1, 2 and 3 present some of our results. The result area of these tables contains four columns: % right pred. indicates the percentage of times the next operation/url was predicted correctly; % wrong pred. indicates the percentage of times the next operation/url was not predicted correctly; % unknown represents the percentage of operations/urls referenced in the test sessions, which had not been seen during the construction of the Markov model; % unpredictable represents the percentage of time the next operation/url could not be predicted with at least the minimum prediction probability. The results presented in the first two tables were obtained from each of the six Markov models constructed with 90% of the

% of used sessions	Time Markov Model Results			
	% right pred.	% wrong pred.	% unknown	% unpredictable
90	22.41	23.57	0.00	54.03
80	23.19	17.52	0.00	59.29
70	22.47	19.16	0.00	58.37
60	27.55	18.68	0.00	53.76
50	26.37	19.59	0.00	54.04
40	25.49	20.16	0.00	54.35
30	24.53	20.51	0.00	54.26
20	24.64	20.60	0.00	54.77

Table 3: Minimum prediction probability = 80%

sessions. The last table shows the influence of the number of sessions used to construct the model, considering only the Time Markov Model.

From the results, we conclude that:

- Models that use the context of the operations had the worst results. That is because the number of outgoing edges from each vertex is much higher than that of the models constructed without considering the context. Besides that, the processing time was higher, which is a result of great importance, since the prediction is proposed to be executed in real time.
- In all the models used, the minimum prediction probability do not affect the relation between the right and wrong predictions, but it does influence the number of unpredictable operations/urls.
- It is not necessary a large number of sessions to construct the Markov model (see Table 3). The variance of the results when we change the number of sessions used to construct the model — from 20% (2085 sessions) to 90% (9383 sessions) — is lower than 5%.

The last conclusion is specially important, since, in general, our proposed tool may not have much information available.

4 Conclusions and Future Work

In this paper we proposed a tool that gives the user more control of his/her privacy. This tool offers many advantages:

- it does not depend on other parties to run;
- it is easy to use;
- it allows the users to determine the privacy level they want, with flexibility;
- it can protect the user privacy through anonymity;
- users do not have to rely on the visited site policies.

One of the mechanisms proposed to measure privacy is the prediction of the next user's click.

As soon as we finish the implementation of this tool, we intend to extend its use by defining a distributed architecture that will exchange model information between the various Guardians in the Web to improve privacy.

References

- [1] Mark S. Ackerman and Lorrie Faith Cranor. Privacy critics - safeguarding users' personal data. *Web Techniques*, September 1999. <http://www.webtechniques.com/archives/1999/09/ackerman>.
- [2] Huaqing Wang, Matthew K. O. Lee, and Chen Wang. Consumer privacy concerns about internet marketing. *Communications of the ACM*, 41(3), March 1998.
- [3] David M. Martin Jr., Richard M. Smith, Michael Brittain, Ivan Fetch, and Hailin Wu. The privacy practices of web browser extensions. *Communications of the ACM*, 44(2), February 2001.
- [4] P3P. *Platform for privacy project*. <http://www.w3.org/P3P>.
- [5] Privacy and American Business. *E-Commerce & Privacy: What Net Users Want*. <http://www.pandab.org/ecommercesurvey.html>.
- [6] David Sims. Do you trust the web? *Webreview*, January 1997. <http://www.webreview.com/1997/01>.
- [7] Kurt Thearling. Data mining and privacy: A conflict in the making? *DS*, March 1998.
- [8] I. Zukerman, D. W. Albrecht, and A. E. Nicholson. Predicting users' requests on the www. In *UM99 Proceedings - the Seventh International Conference on User Modeling*, pages 275–284, Banff, Canada, 1999. Springer-Verlag.