P3P Semantic Checker of Site Behaviours

Robson Eduardo Grande[1] and Sérgio Donizetti Zorzo
Federal University of São Carlos, Computing Department
Zip Code 676 - 13565-905, São Carlos, Brazil
{robson_grande,zorzo}@dc.ufscar.br,
WWW home page: http://www.dc.ufscar.br

Abstract. The interactive use of the web between users and service providers introduces a privacy problem that involves the undesired disclosing of user personal information, mainly with the presence of personalization that needs this type of information. Also there are many manners to face it, but the Platform for Privacy Preferences (P3P) is one that provides a variable level of privacy for the user's browsing. However, the P3P only introduces a privacy contract between the site and the user, without guarantees that it will be obeyed by the site. Then a semantic checker can be added to the P3P architecture to compare the contract with the site attitude and to increase the trustworthiness on the P3P contract. Some experiments are accomplished and the results are displayed to show the present situation of the privacy policies of the sites, and we discuss what it implies in the data gathering and what is gained with the use of the semantic checker.

1 Introduction

The interface implementation in e-commerce applications must consider two aspects: the marketing necessities and the user privacy. For both aspects user's personal data has a great worth. On one side, providing competitive advantages in the market through the marketing and, on the other hand, presenting a bigger user confidence with regard to his disclosed data. This user confidence is recognized as an important item to bring to the success of an online marketing, and consequently to increase purchases and sales in the web and to improve the e-commerce markets.

The personalization application is one of the strongest competitive advantages in the market. It has become an indispensable instrument to the progress of the services and online businesses. The idea of receiving personalized services from visited web sites is sufficiently attractive, besides very well accepted by the users. In accordance with Kobsa [1], clients need to feel that possess a personal and unique relationship

[1] Robson Eduardo Grande receives financial support from CAPES.

with the enterprises, and to confirm this is presented a research that shows that sites offering personalized services achieved an increase of 47% in the number of new clients.

Meanwhile, so that personalization to be applied it is necessary the information collection originated in several provenance, which can be gotten explicitly or implicitly.

An individual that sends explicitly information is obviously aware of the sending of it. On the other hand, the implicit gathering of information is the resultant data acquisition from the user browsing observation, and he can be or not aware of its existence. Information obtained by this method contemplates the e-commerce interaction data and the clickstream [2] [3] data, which makes possible the creation of user profiles based in user interests, browsing patterns, preferences and others.

But, while the user data gathering can help in the marketing of e-commerce sites, it can prejudice the marketing too. Depending on the form that the gathering method and analysis process of user data are accomplished it can characterize a privacy invasion, whereas the user can loose the control of his personal information [4].

This privacy lack results to a user confidence loss, as from this he stops accessing certain services fearing that personal information, which has a considerable value for him, is disclosed or has a malicious use. This can be confirmed by Teltzrow [5] that says 64% of web users haven't accessed some time a web site, or they haven't bought something from it because they don't know how their information would be used. Also 53% of the users don't trust in commercial web sites that gather data, 66% of them don't register in online sites fearing that their information may be used inappropriately, and 40% of them falsify data when registering online [6].

Moreover, privacy is considered to be intrinsically related with the control that an individual has over determined information [4]. In this way, it must be inherent in trustworthy transactions, in another way a privacy lack will contribute to cause a fault of the business model of the electronic commerce.

An increase of the user control perception causes an increase of adoption of services and products based on the web. The user control perception is provided by bigger information disclosing of the collected data use that is made, and the receiving of something worthy in return stimulates the data disclosing by the user. Jutla [6] reports that 51% of web users desire to disclose personal data to receive something worthy in return and, a research shows that 90% of users want to be asked after permission before their information is used or gathered [5].

The question is to find an equilibrium point between personal information gathering and user privacy. Nevertheless, to find this equilibrium point becomes difficult because the privacy is subjective, in this case each one has its privacy discernment.

Several mechanisms utilized to guarantee the user privacy have consequences in their access form, producing degrees of reachable personalization. In this way they can denigrate the personalized service availableness.

The 3P Platform, that have been used in a pretty crescent and extensive way [7], is interesting to become possible the privacy level modulation in accordance with the user preferences, and in this way adapting better to the user characteristics. Also, to the use of this platform isn't necessary to make many modifications, since basically it includes an automatic method of privacy policy reading. Add to that, the privacy policies have a big user acceptance, 76% of users think privacy policies are very important and 55% of them believe that it turns the personal information disclosing more comfortable [5].

Notwithstanding, this platform presents a low level of trustworthiness in the semantic aspects of the manipulated data. In this manner, it is proposed a semantic checker looking at the increment of the trustworthiness degree in the P3P tool. The user trustworthiness is incremented by comparing the P3P privacy policies with the site's behavior. Tests are accomplished and their results show that some sites write P3P policies correctly, and the inclusion of the checker can obey the other sites to improve the construction of their policies.

This work introduces in the second section several access forms to the user data with privacy. In the third section we present the 3P platform mechanisms and its limitations. The fourth section presents the Semantic Analyzer. The fifth section shows the experiments and results obtained and finally the sixth section describes the conclusions and future works.

2 Privacy Mechanisms

As manners to face the problem of privacy in the web, many proposals exist that can be divided in two basic forms of approach. One of these forms aims at the introduction of architectures or mechanisms, tries to keep the anonymity of the user, or makes difficult the identification of him. The next mechanisms follow this line of approach.

Cookie crushers or cookie filters are the most common of them. They provide a way of controlling or not permitting the cookie existence in the user computer, avoiding that personal information can be stored to be recovered subsequently.

Theoretically cookie [8] is used to store in the user computer the estate of his browsing in a determined site. The cookie content is created by web server. The sites utilize this information piece to characterize the user profile, gathered through analysis methods of browsing as clickstream.

Clickstream, also known as clickpaths, is the route that the user chooses when he clicks or browses by a site. This information is used to determine user profile in his browsing.

The anonymity is wanted by several users that don't permit that any personal information is discovered, thus avoiding any privacy problem and identification of the user identity. Three mechanisms are presented as examples of this type of approach. Anonymizer [9] is a web proxy that forwards the user requisitions and applies certain methods to mask them as requisitions from the proxy.

Onion Routing [10] is constituted of one or more routers and each one works as a proxy that applies certain methods to improve the privacy and to forward randomly to the next router or to the destiny site in question. This routers net is built dynamically, it is fault tolerant and works to avoid eavesdroppers by making difficult to determine the user requisitions source.

In Crowds [11], each user contributes hiding the real origin of a requisition, with the member co-operation of a group. Randomly and dynamically a requisition can be forwarded to a member of a Crowd group or to the site destiny, and it isn't possible to make backtracking search to determine the origin user because each user requisition changes the routing process to create a different path.

Another approach to permit privacy during the user browsing is using pseudonyms. It consists basically in to create fictitious names to users to disguise the user identity permitting personalization, as soon as web sites are able to determine user profiles without link them to the user real identity. An example of this type of mechanism is the JPWA (Janus Personalized Web Anonymizer) [12]. It acts as an intermediary entity (proxy) between the user and the web site generating automatically nicknames when users want to access determined services, executing the authentication process. If the real identity of a pseudonym is discovered, all the user actions in the past will be automatically exposed.

Managing Anonymity while Sharing Knowledge to Servers (MASKS) [13] is another approach to protect the user privacy permitting personalization based in the pseudonyms idea. It hides the user identity under masks or pseudonyms. These pseudonyms are associated by some way to a group of similar requisitions. These groups are defined in accordance with user interests exhibited during the interaction with a web service by making requisitions in name of a group, contrary to an individual user, thus not disclosing the user identity. The user requisition is designed to a group and not to a user because the requisition represents the user interest in a specific moment. The MASKS doesn't provide privacy when users send explicitly their information to sites.

All the presented mechanisms show some limitations in preserving the privacy or permitting personalization. Those mechanisms that protect all the user privacy don't permit personalization, and those that permit personalization have faults to protect the user privacy, or they have some problems in the security aspects.

3 P3P

Another line of approach introduces the idea to police the sites or to inform the user about the privacy policies that are adopted by them, communicating the information that is gathered. According to this criterion are presented the following systems.
One of them is a tool that provides information related to context about privacy and personalization options [14]. It is a support system to the user navigation exhibiting a situational communication dialogue when the user information is gathered.

P3P [15][16] inserts a way to manage the user browsing through of a standardized method to disclose how the user information is gathered, how they will be utilized and the sharing of them to third parts. This management is made through the P3P privacy policy checking by a user agent.

P3P inserts a contract between sites and users, defining a protocol that permits the site administrators to publish a site privacy policy. Add to that, a user agent is defined by the platform that reads automatically this privacy police, verifying if it combines with the user privacy preferences or user security configurations, because the majority of the users pay attention poorly to the privacy policy readings [5].

Web sites are qualified to express their privacy actions by the 3P platform in a standard format. This standard format consists in indications made and based in the P3P vocabulary to express the privacy behavior of each web site. These indications are made by a XML codification with name spaces from P3P vocabulary to provide information that defines the site privacy policies, informing which information is obtaining and which form is using to obtain it, where and how long the information will be stored, who is the responsible and the information gathering purpose. The P3P vocabulary é planed to be descriptive of the site behavior, but not to be simply an obedience indicator to a particular law or a conduct code.

The privacy policies of a site must have a reference with their respective and specific particularities, permitting to determine the policy range in determined site region.

The policy reference is a codification in XML with name spaces that can specify the policy to an entire site, portions of a site or to a unique web document. It also links the site parts to its respective policies, shows the exact location of the file that contains the P3P privacy policies, defines the access methods to which the policy is applicable and the time period to which it claim that is considered to be valid.

By any means the P3P reference file location must be known to find the file, to begin the analysis, and to know the P3P privacy policy location. To his, four mechanisms can be used to indicate the policy reference file place: a well known location (/w3c/p3p.xml), HTTP headers can be used to point to a policy reference file through the creation of a new answer header, a HTML or SHTML link tag can be used to obtain the policy reference file.

The user preferences inform the way how the user agent must act when analyzing a privacy policy of a site.

The user P3P agent works as a tool that acts co-operating with the browser. It can be a plug-in added to the browser, a proxy server, or built in the web browsers.

The user agent looks for a reference of a policy reference file in the HTTP header, HTML or SHTMLK link tags and by the well known location. With the reference file the agent can obtain the privacy policy respective to the URL the user requested and can begin to analyze the policy by comparing it with the user privacy preferences. This analysis process results in an indication with symbols, sounds or generating alerts to the user.

P3P Limitations

According to the P3P specification there is one limitation. It's argued that the user should have control of his privacy instead of to trust completely in privacy policies of web sites [13]. Add to that, this is a great problem with the P3P because it can't guarantee sites will act following their policies, considering that P3P is only a document describing the privacy policy of a site.

In this way the site can be obtaining additional data that is specified by the privacy policy, and the user doesn't have guarantees of what information is collected. Add to that, this is passed in way that the user doesn't perceive, whereas he trusts on the signalizing of the user agent that makes the automatic policy analysis.

Laws and auto-regulatory programs that protect the user privacy can be verified during the P3P policy assimilation, and, thus imposing certain obedience to the policy correctness related to the site. Such laws are only a way to influence web sites to be sincere in the building of its policies. However, these laws don’t guarantee that every web site will apply it.

To guarantee that the P3P policies are obeyed is necessary to insert some checker that can’t be handled by the sites, it executes a verification of the privacy policy faithfulness considering the sites behavior, and returns a guarantee seal.

4 P3P Semantic Checker

The Semantic Checker presented in this work objectifies to extend the 3P Platform. It adds a bigger trustworthiness to the user browsing by the conference of the privacy policy correctness proposed by the 3P platform. The conference of the privacy policy correctness is made by inserting a semantic analyzer to the P3P agent user that makes a P3P privacy policy checking. This checking is made comparing the site behavior with it privacy policy, and this behavior is represented by the site source code that comes to the user browser.

It must be localized in the user computer, trying to increase the security of this checker proposed. With this guarantee that the site doesn’t have access to the checker and, thus the site can’t corrupt it to produce a false result to the user. If it produces a false result it will be deceiving the user, coming back to the original situation where there weren’t privacy guarantees.

After that the user information is collected there isn’t possibility of knowing the destiny that will be given to it. Thus, the privacy policy checking with the site behavior is only possible at the verification of user data gathering, and without this approach only rests the trustworthiness on the privacy policy as the unique way of to police the site attitudes with the collected information, or to belief on the law codes [16].

At the implicit gathering, the information collectors utilize ways that can’t be identified by some mechanic analysis method, or they don’t follow a pattern to be identified. The cookies can be identified, seeing that they are stored in the user computer through a http requisition. However, they aren’t a form of information gathering, but a form of storage of information gathered, and their content can’t be understood.

The analysis of explicit gathering is the way that better shows the site behavior at the user computer view. This type of gathering can be observed in the html page source code that the user accesses. The data requisition, which is made to the users, is accomplished by html elements of data entrance. Also, the html code is very used to the construction of Web pages, it is generated by the majority of languages of dynamic pages as ASP, PHP, JSP and others, and it is embedded in the JavaScript code.

To find the places where the user information is obtained is necessary to look for by the form tags in the page source code in question. Each data entrance field in each form in the site source code represents a data entrance that the user can enter his information to be sent to the site, and the set of these input fields results in the information set that the user send explicitly to the site. To accomplish the analysis is necessary to find each one of these data entrance fields e to compare with the respective P3P policy to verify if, through the page, the site is obtaining some information over that the policy specifies.

The checker can be added as a module to the user agent, accomplishing the checking every time that the user does a page requisition, which can be visualized by the figure 1. In the figure the first communication that is accomplished with the site server is to obtain the privacy policy, with it the agent can make the comparison with the user preferences. In the second communication the server sends the requested page, and the semantic checker accomplishes the analysis with the source code of this page and the P3P privacy policy.

The privacy policy validation can be made before or after that the agent makes the comparison between the policy and the user preferences, resulting in an additional signalization to the user or even to influence the result that the agent signalizes.

Therefore the semantic checker functioning is made of the following form, incorporated at the 3P platform, which can be visualized by the figure 1.

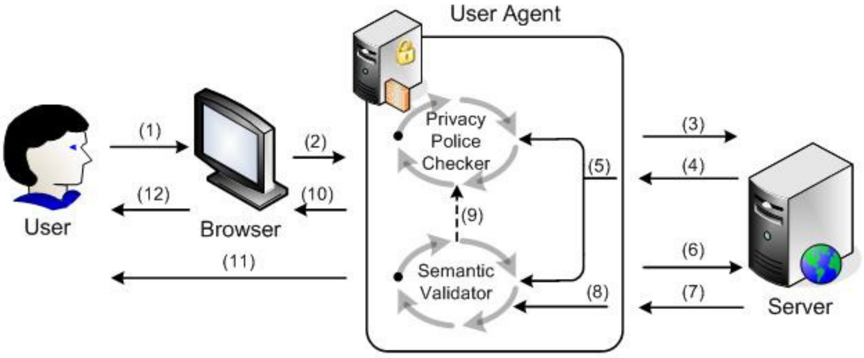

Fig. 1. A Diagram that represents the insertion of the semantic checker to the P3P Platform architecture.

In a page requisition made by the user in his browser (1) in an architecture that uses P3P, the user agent intercepts the requisition (2). Initially it looks for (3) (4) a P3P privacy policy to analyze the site policies with the user preferences (5). With the analysis made, the agent can signalize to the user positively (11), permitting that the user can access the page (6) (7) (10) (12), or negatively, letting to the user to take the decision in accessing or not the page.

The checker makes a semantic analysis with the privacy policy obtained (5) and the source code (8). By this analysis is created a signalization that can be incorporated to the signalization that the user agent produces (11), or can be refined by the agent as an additional criterion in its policy analysis process (9).

A functioning architecture of semantic checker is presented by the figure 2. At it, three entrances can be identified, html source code file and the respective URL (1), privacy policy reference file (2) and P3P privacy policy file (3).

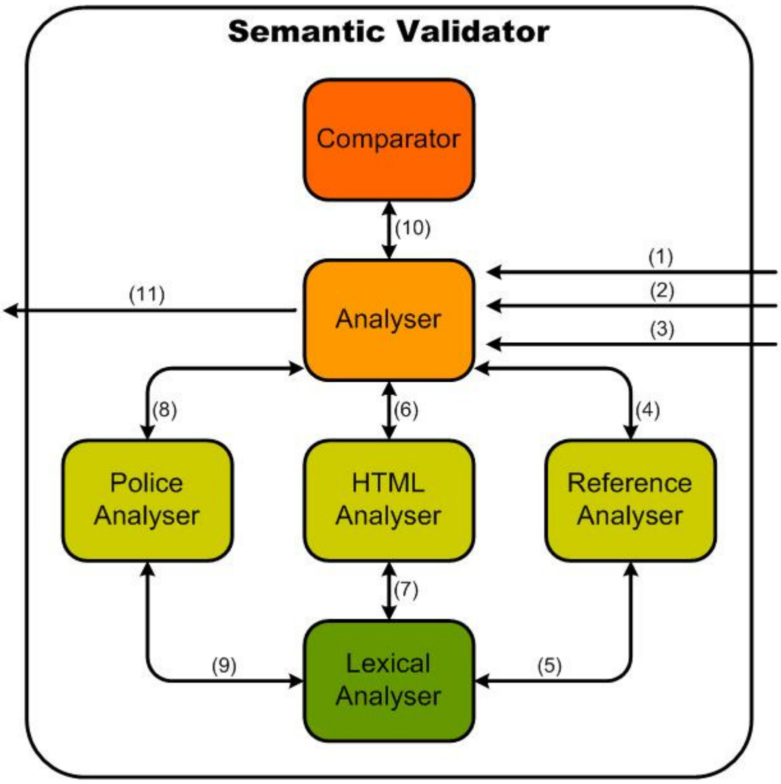

Fig. 2. Semantic Checker Architecture and functioning – through the html source code analysis of the site and of its privacy policy is generated a result to the policy approbation.

By the entrances at the figure 2 will be begun the policy analysis by the Analyzer. First is necessary to find the respective page privacy policy according to its URL (4), to this is utilized a lexicon analyzer to recognize the reference file elements (5). At the end of this process it is returned the policy name that must be used.

Determined the respective policy, the data entrance elements are obtained, which are identified by the *input* fields (6). Also it is used the lexicon analyzer to recognize the html elements and to be able to found them (7). The privacy policy data elements also will be obtained (8) using a lexicon analyzer to identify them by the *data* fields.

Obtained the both elements, is accomplished a comparison of each input field of the html source code with the privacy policy elements to investigate is the field is specified by the policy (10). Depending on this comparison made, a negative or positive result can be returned (11).

The figure 3 presents a sequence diagram of a user's page requisition. The sixth arrow represents the semantic checker action added to the P3P, and the others arrows represent the user requisitions and the P3P answers without the checker functioning inclusion.

Add to that, each information entrance field needs to be delimited by the site privacy policy, and the fields of data entrance in the site source code needs to have some binding with the respective elements of its P3P specification, which in the case of this initial implementation, the link is the html field to have a name equals to the P3P policy specification.

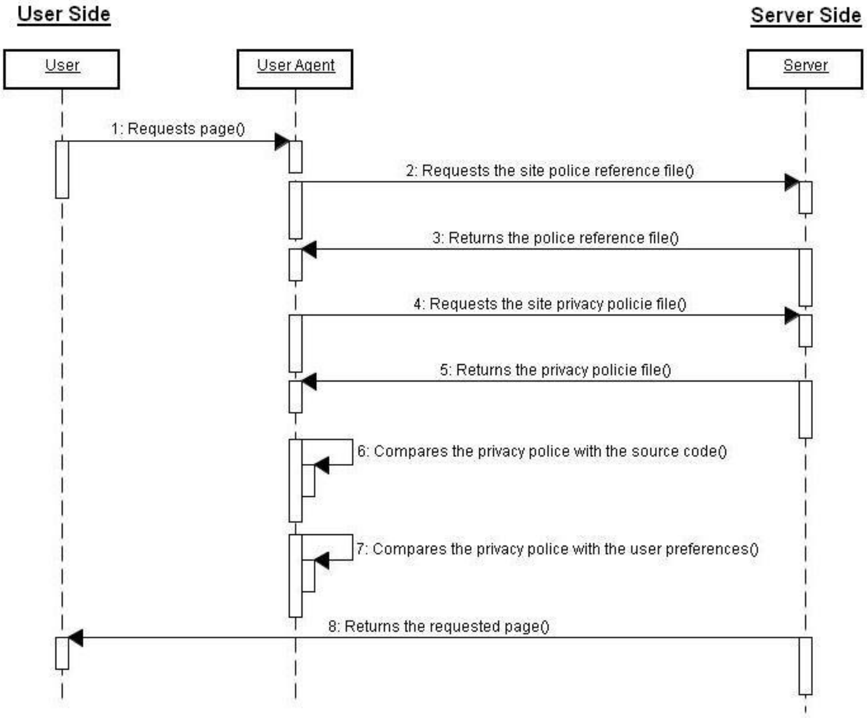
User Side
Server Side
User
User Agent
Server
1: Requests page()
2: Requests the site police reference file()
3: Returns the police reference file()
4: Requests the site privacy policie file()
5: Returns the privacy policie file()
6: Compares the privacy police with the source code()
7: Compares the privacy police with the user preferences()
8: Returns the requested page()

Fig. 3. Sequence diagram of an example of semantic checker execution in a page requisition.

However, this proposed link between names restrings the page construction. But, the proposal suggested is initial and utilized more to test. A proposal more suitable is to create an attribution file together with the P3P policy to be utilized to accomplish the verification of the input fields. This additional brings flexibility to the owners of the sites at the creation of entrance fields.

5 Experiments and Results

The experiments are accomplished trying to present a situation of how the sites and their respective P3P privacy policies are.

The research consisted in to make a process closed to the mechanism of the semantic checker, obtaining each source code of each page and its respective privacy policy. It was made a comparison between the data element names of the policy and the input element names of the page, signalizing if some element wasn't delimited by the policy.

The result of the automatic analysis was the same as the result that was hoped, no one page that had text entrance elements and was analyzed passed by the verification with positive signalization, whereas the analysis is based in the comparison of names.

But, in a manual analysis of the gathered data some evidences could be observed. This manual analysis consists in to observe the input elements obtained with their respective data elements. In this observation was approached also the comparison between the name means of the gathered input fields and the means of the data elements obtained, and thus to try understand how was constructed the privacy policy and the site page to relate them in some way.

Initially were accessed 100 compliant sites with the 3P Platform. The addresses of these sites were obtained from a listing in *http://www.w3.org/P3P/compliant_site*. In a sampling of 100 sites only 57 could be analyzed, the others 43 had some problem that made impossible their access or the use of their P3P privacy policy: the site was in construction, it wasn't found, the access was forbidden, there were problems in the syntactic construction of the P3P policies, the policies weren't found or the policy reference files weren't found, seeing that the reference file was looked for in the well known location.

With these 57 correct sites 120 pages were obtained so that their html source codes were analyzed. By the analysis, 33 pages of 120 didn't have any input element of data entrance, and thus they weren't utilized to the verification. Therefore, with a sampling of 87 pages was obtained the following graphic presented in figure 4.

The research only approaches the explicit information gathering, the dynamic data specification [16] wasn't considered, as *dynamic.http*, *dynamic.clickstream* or *dynamic.cookies*. The cookie content doesn't follow a construction pattern: each site builds it in a different way, needing knowledge of all site behavior and source code to predict its value.

The 87 pages were classified in six categories, in accordance with how they presented their privacy policies, and each category represents a percentage of the 87 pages.

The "A" category delimits the pages that have a P3P privacy police and doesn't describe any one of their input fields of their html source codes.

The “B” category delimits the pages that have policies specified in a generic way all or a big part of their input fields. It describes only what is made with any data that is collected by the page, or also it can be a generic specification as *user.business-info* or *user.home-info*.

The “C” category delimits the pages that have a privacy policy that specifies each input element of the form in their html codes.

The “D” category delimits the pages that have privacy policy that specifies only some input fields of the source code.

The “E” category delimits the pages that have names of input fields in the html code that their meaning can’t be understood or they have a very generic meaning, as loesung1, sp-q or word.

The “F” category delimits the pages that have some input field names in the html code that can be understood and they are specified by the privacy policy.

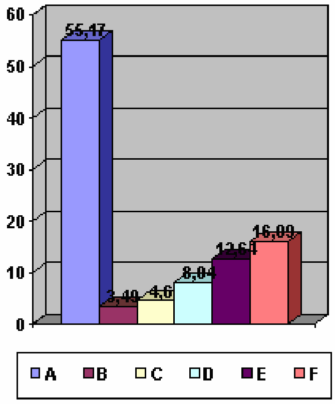

Fig. 4. Graphic that presents the experiment results classified in categories of P3P privacy policy situation.

Also it was observed that the majority of the pages, close to the totality, specified the dynamic data gathering. Even to those policies that informed nothing to the explicit data gathering. This evidences that the P3P compliant sites are aware in specify the implicit information gathering, but they specify poorly the input fields at html source code.

Therefore, in general terms can be concluded that 44.82 percent of the pages, almost half of the pages that utilize the 3P platform specify with more details in majority of their fields of explicit information gathering. This shows that the mechanism can be used as a tool to improve the user trustworthiness by verifying.

6 Conclusions

Among all the experiment results made, it can be observed that the totality of the sites compliant with the P3P policy don't satisfy the restrictions to the execution of the checker. It was also observed that more than half of the sampling of these 100 sites complaints with the 3P platform don't look for to specify in detail the data that is obtain from the users.

The insertion of this checker in the 3P platform increases the user trustworthiness guaranteeing that all the data that is gathered explicitly is specified by the P3P privacy policy. The checker would be an additional reinforcement to the sites that use P3P and detail their specification of data gathering, and would force the others to detail the elements of the information gathered. Thus, over increasing the trustworthiness in the privacy policies, also would improve the construction of the P3P policies, with more details in the site specifications.

The next step is to improve the process of checking by retiring the restriction that each html entrance field must have the same name as in its respective P3P privacy policy, this can be made by using a file stored together with the privacy policy that makes the linking between the names, as suggested before.

As evidenced, this proposal considers the explicit data gathering, which is analyzed in the site source code. However a cookie analysis can be added to the checking process, the cookie presence is identified by the HTTP communication, but its content can't be understood or analyzed.

References

1. A. Kobsa (2001). Tailoring privacy to user's needs. Proc. Of 8th International Conference on User Modeling. http://www.ics.uci.edu/~kobsa/papers/2001-UM01-kobsa.pdf.

2. R. E. Bucklin, J. M. Lattin, A. Ansari, D. Bell, E. Coupey, S. Gupta, J. D. C. Little, C. Mela, A. Montgomery, J. Steckel, "Choice and the Internet: from Clickstream to Research Stream", U.C. Berkeley 5th Invitational Choice Symposium, Mareting Letters, 13(3), 245 -258, Last Revised February 10, 2002.

3. A. L. Montgomery, S. Li, K. Srinivasan, and J. C. Liechty, (2004) "Modeling ONline Browsing and Path Analysis Using Clickstream Data", Marketing Science, Vol 23, No. 4, p579-595.

4. Privacilla, (October 11, 2005); http://www.privacilla.org.

5. M. Teltzrow and A. Kobsa (2004) "Communication of Privacy and Personalization in E-Business". Proceedings of the Workshop "WHOLES: A Multiple View of Individual Privacy in a Networked World", Stockholm, Sweden. http://www.ics.uci.edu/~kobsa/papers/2004-WHOLES-kobsa.pdf.

6. D. Jutla, and P. Bodorik, (2003) “A Client-Side Model for Electronic Privacy”. 16th Bled eCommerce Conference and Transformation. June 2003.

7. P. Kumaraguru, and P. Cranor, “Privacy in India: Attitudes and Awareness”. In Proceedings of the 2005 Workshop on Privacy Enhancing Technologies (PET2005), 30 May - 1 June 2005, Dubrovnik, Croatia.

8. D. Kristol, and L. Montulli, “HTTP State Management Mechanism”. Bell Laboratories, Lucent Technologies. Epinions.com, Inc. October 2000. RFC 2965. http://www.ietf.org/rfc/rfc2965.txt.

9. Anonymizer, Inc. (2004) “Anonymizer Enterprise Network Privacy/Security Appliance”. Technology Overview. www.anonymizer.com.

10. D. Goldschlag, M. Reedy, and P. Syversony, (1999) “Onion Routing for Anonymous and Private Internet Connectinos”. January 1999. www.onion-router.net/Publications/CACM-1999.pdf.

11. M. K. Reiter, and A. D. Rubin, (1997) “Crowds: Anonymity for Web Transactions”. AT&T Labs – Research. avirubin.com/crowds.pdf.

12. E. Gabber, P. E. Gibbons, Y. Matias, and A. Mayer (1997) “How to Make Personalized Web Browsing Simple, Secure, and Anonymous”. Bell Laboratiories, Lucent Technologies. http://www.bell-labs.com/project/lpwa/papers.html.

13. B. G. Rocha, V. A. F. Almeida, L. Ishitani, and W. Meira Jr., (2002) “Disclosing Users’ Data in an Environment that Preserves Privacy”. Workshop On Privacy In The Electronic Society.

14. A. Kobsa, and M. Teltzrow, (2005). “Contextualized Communication of Privacy Practices and Personalization Benefits: Impacts on Users’ Data Sharing and Purchase Behavior”. PET 2004.

15. L. F. Cranor, (2003) “’I Didn’t Buy it for Myself’ Privacy and Ecommerce Personalization”, WPES’03 (Washington DC, USA, October 30, 2003), AT&T Labs-Research.

16. Platform for Privacy Preferences Project, P3P Public Overview. http://www.w3.org/P3P/.