Expert Systems with Applications 40 (2013) 5228–5239

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Web mining and privacy concerns: Some important legal issues to be consider before applying any data and information extraction technique in web-based environments

Juan D. Velásquez

Web Intelligence Consortium Chile Research Centre, Department of Industrial Engineering, Universidad de Chile, Av. República 701, P.O. Box: 8370439, Chile

ARTICLE INFO

Keywords:
Web mining
Privacy
Regulation
Personalization

ABSTRACT

Web mining is a concept that gathers all techniques, methods and algorithms used to extract information and knowledge from data originating on the web (web data). A part of this technique aims to analyze the behavior of users in order to continuously improve both the structure and content of visited web sites. Behind this quite altruistic belief – namely, to help the user feel comfortable when they visit a site through a personalization process – there underlie a series of processing methodologies which operate at least arguably from the point of view of the users' privacy.

Thus, an important question arises; to what extent may the desire to improve the services offered through a web site infringe upon the privacy of those who visit it? The use of powerful processing tools such as those provided by web mining may threaten users' privacy.

Current legal scholarship on privacy issues suggests a flexible approach that enables the determination, within each particular context, of those behaviors that can threaten individual privacy. However, it has been observed that TIC professionals, with the purpose of formulating practical rules on this matter, have a very narrow-minded concept of privacy, primarily centered on the dichotomy between personal identifiable information (PII) and anonymous data.

The aim of this paper is to adopt an integrative approach based on the distinctive attributes of web mining in order to determine which techniques and uses are harmful.

1. Introduction

"Web Personalization" is the research branch of "Web Intelligence" dedicated to helping users find exactly what they are looking for in a web site, which can often be problematic. To tackle this issue certain systems have recently been developed in order to help the user by means of both improving their navigation methods and providing them advice related to their searching requirements. Furthermore, these systems yield valuable information to site owners and web administrators, which allows the execution of changes in the structure and content of a site, taking into consideration the idea of improving the user's experience and thus making the user feel more comfortable visiting the site. In order to carry out the above, multiple approaches have been developed which aim to extract information from the web data generated by each visit to a web site. In other words, as Ashworth and Free (2006) state "*the widespread acceptance of the Internet as a platform for commerce has made it possible for organizations to gather a wide range of consumer information including browsing patterns, items purchased, profitability, dates and times of activities and keystroke behaviour which uses browsing behaviour as a predictor of receptiveness to certain ad messages, has burgeoned*".

Although research in this field is motivated by altruistic beliefs, "the excess of help" may lead to impinging on the user's privacy, particularly through the persistent requirement of personal data. Various studies have shown that sites which include user-tailored contents can establish a loyal relationship with their visitors (Kobsa, 2001), but questions have arisen concerning what the trade-off of that is (Cavoukian, 2008). In the course of this research, the issues surrounding the effects of web personalization, a clear step toward a semantic web, ought to be tackled in an effort to go a step forward with technological development. However, it must be noted that when both ethical and legal issues associated to technological innovation, are not adequately addressed, individuals end up bearing the changes without any real possibility of avoiding them. As Marcella and Stucki (2003) said "*a dilemma arises when a society is forced to choose between increasing its quality of life and maintaining the privacy of that life. As technology has enabled individuals to enjoy an easier life than their ancestors, it has also opened their lives to examination and scrutiny by others*". In this sense, if we count cookies and IP addresses as personal

E-mail address: jvelasqu@dii.uchile.cl
URL: http://wi.dii.uchile.cl/

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.03.008

information, Internet users have left behind personally identifiable information everywhere they have been (Lundevall-Unger and Tranvik, 2011).

Even though the relationship between privacy and technology has been addressed a number of times, the predominant approach of the IT industry has been to assume a narrowed concept of privacy that denotes basically the dichotomy between public and private spheres based on the nature of data-namely, if these are personal identifiable data or anonymous data. Although the reason for doing that is not arbitrary, since the purpose is to have a useful concept which will enable the development of practical rules, significant negative consequences can be expected.

The consequences of maintaining a narrow-minded view of the issues concerning privacy would be to embrace a concept in disarray (Solove, 2006) and which needs to be reinvented (Gutwirth, 2009), or even more, that privacy itself in a technological environment is totally dead. Despite the fact that this way of understanding the nature of privacy can be useful, it does nothing but hide the real complexity of the phenomenon. That is the reason why most innovative legal theories concerning privacy suggest the building of a dynamic and flexible concept taking into account distinctive contexts (Craig, 2012).

Our proposal on this matter is to take information privacy one step further, by including within this concept both privacy issues concerning usage of personal data and those threats which arise when private parties make decisions about others solely from computer modeling outputs as if they were personal data.

This paper is organized into four sections, the first of which outlines the main characteristics of web mining and web data. The second section attempts to briefly portray the treatment of privacy in the U.S., highlighting some relevant federal laws applicable to privacy concerns on the internet. The third section gives a summarized description of most of the innovative legal theories concerning a flexible and context-reactive concept of privacy, and in turn deals with the issue of determining how and to what extent web mining tools threaten users' privacy. Finally, the fourth section highlights some recommendations to IT professionals for maintaining the personalization of the web through adopting a privacy-friendly approach.

2. Technical background

2.1. What is web mining?

Web mining is a concept that gathers all techniques, methods and algorithms used to extract information and knowledge from data originating on the web (web data). Thus, it could be said that this is an application of data mining theory onto web data.

Web mining has enabled the growing amount of data available on the web to be analyzed, and furthermore has demonstrated that conventional and classic statistical approaches tend to be inefficient in carrying out this task (Markov and Larose, 2007). The importance of having clean and consolidated data lies in the quality and utility of the patterns discovered by these tools, being directly dependent on the data which will be used. Because of the latter, web mining tools can yield wrong or misleading information, unlike statistical tools, due to these techniques being at the disposal of people who know little about this field.

According to the current stage of development of web mining tools, the following are among the most used data processing techniques (Velásquez and Palade, 2008):

1. Association rules: Are geared to finding relationships among data sets under a frequency ratio (confidence) of a proportion of total data (support). For instance: Bread; Cheese (support = 5%, confidence 42%). This means that 42% of people who bought cheese did so with bread as well in 5% of all transactions. Moreover the above association might be extended to a multidimensional array of meanings adding further attributes, for example: bread, butter, and cheese.
2. Classification: Is to place a series of logs into certain previously defined categories. In order to enable this, a knowledge process is often adopted in which the algorithm is applied to pre-classified data in order to determine under which values of the other attributes of the log one category or another belongs. Once the learning is achieved, one can proceed to evaluate the registers that were not used as a learning tool.
3. Clustering: Is to gather objects which have similar characteristics. Unlike the above technique, it is not known *a priori* which categories will be created. In fact, through this process one expects to find them. To do this, similarity measures are used amongst registers which enable them to be split according to their differences. There are three main clustering techniques.
   a. *Partitional clustering*: Under this technique *a priori* N clusters are defined, under which the records will be evaluated.
   b. *Hierarchical clustering*: This technique refers to the building of clusters through a hierarchical de-composition which can be categorized into two types: (1) agglomerative, which is initiated with one cluster per register, and which begin to be gathered according to a similarity quantification until a pre-defined terminal condition is obtained; and (2) divisive, which initiates as a single cluster which includes all records, then proceeds to split according to a similarity quantification until a terminal condition is established.
   c. *Density-based clustering*: Borrowing the definition of density given by physics consists of defining a density threshold which is no more than a predefined cardinality for each cluster, and a radius which consists of a pre-defined distance, so that clusters are formed by records at a distance from the centroid of the cluster with the lower radius. Centroids are defined for each iteration and algorithm when all cardinalities are lower than the value of the predefined density threshold.

2.2. Nature of web data

The information created on the web is characterized both by significantly increasing each day, and by consisting of highly varied data, especially when we take into account the mass of services of the so-called "Web 2.0", via which users assume a leading role in creating contents. However, the information which flows on the internet does not only refer to data contributed directly by private individuals. In many cases, it includes traces of interactions done by either a private individual or by the computer itself, which enables it to access the network in order to display certain contents.

For web miners, this attribute of web data is one of the main problems to consider when desiring a tailoring process to be performed. This implies that the data must be conceptually classified in advance so that, based on those categories, a mining technique can be selected. According to the most widespread theory, the data originating from web sites or web data may be classified under three different sources (Cooley et al., 1999):

1. *Contents*: Refers to the objects that are available within web pages, for instance, images, free text, sound, etc.
2. *Structure*: Refers to the hyperlinks structure that a web page contains. Mining the structure has the aim of analyzing the way in which different web documents are linked together.

3. *Usage*: Refers to all transaction data which is logged in a web server. It commonly corresponds to web logs which contain data on the entire history of interaction between users and the web site. It should be highlighted that this type of data often corresponds to what is known as "clickstream data", that may be defined as "*a record of a user's activity on the Internet, including every web site and every page of every web site that the user visits, how long the user was on a page or site, in what order the pages were isit, among others things*" (Garrie and Wong, 2007).

Some scholars argue that user-provided data such as name, age and sex should be considered as part of web data (Eirinaki and Vazirgiannis, 2003). However, this type of data, strictly speaking, should not be processed in a web mining project; hence this will not being considered in the current work.

Web data must be pre-processed before being entered into a web mining process, *ergo*, they are transformed into characteristic vectors which contains intrinsic information.

Even though the entire quantity of web data is important, special attention is paid to web logs, as these store the data concerning the history of interaction between the user and the web site, data regarding the user content preferences, and in short, that of most of the user's behavior. Due to this factor, we aim to particularly focus on web logs as the most controversial source with respect to the user's behavior analysis.

2.3. Web logs

Most interaction on a web site is recorded in files known as "web logs". Basically, through a web log it is possible to estimate the objects which were requested by the users, then rebuild their sessions and thus make an accurate tracing of their navigation activities.

Fig. 1 shows both the structure and conventional contents of a web log.

Therefore, the standard structure of a web log can be defined as follows (Velásquez and Palade, 2008):

1. IP address: Is the address of the internet host, namely the computer's identifier through which users access a web site.
2. ID: Refers to identification information provided by users (optional).
3. Access: Is also called "authuser", which is used when SSL (Secure Socket Layer) protocol is turned on. Through this field it is possible to both receive and send confidential information.
4. Time: This indicates both the date (DD/MM/YYYY) and hour (HH:MM:SS) when a web object has been solicited.
5. Request: This represents the object requested by the browser, specifying the return method (GET or POST), URL address and protocol used.
6. Status: Is a whole number which indicates the state of the request to the server. A typical status is the message "Error 404/Server not found" which expresses that the information was not found despite the server being located.
7. Bytes: Shows the amount of bytes in the petition.
8. Referrer: Is a text sent by the client's computer which indicates the original source of a request.
9. Agent: This field displays the browser name and O.S. version.

2.4. Sessionization

As outlined earlier, web mining tools use as a data input the web data preprocessed in the shape of characteristic vectors. Web logs are the type of data which provide the largest amount of information for doing a user's behavioral analysis on a web site (Cooley et al., 1999). The next stage is to develop a personalization process which consists of rebuilding the user session that stems from data present in the web logs. This process is known as "sessionization" and can be summarized as follows:

1. Cleaning records stored in web logs, leaving only those related to web page requests, and removing those which indicate requests of objects contained in those pages.
2. Identifying records which have requests made by web crawlers (i.e. Google bots.) In order to achieve this, official and unofficial lists of crawlers must exist on the web. Those might be identified through the field "agent" or failing that, by mean of an IP address.
3. Gathering records by IP address and agent. It should be noted that it is assumed that no further data should exist concerning those users who visit the site.
4. Sorting records from lower to higher timestamp, so that records appear chronologically.
5. Identifying navigation sessions, for which there are two options.
   a. Using statistical criteria, namely to assume that in general the real user sessions do not last longer than thirty minutes.
   b. Assuming that no pages are visited twice per session.
6. Finally, rebuilding the session using hyperlink structure, adding those pages which were not registered due to the fact that cache memory from the web browser or the corporative cache of a web server was used.

IP	URL	Time	Size (bytes)	Referring URL	User agent
200.89.69.174	/cpanel	5/31/11 2:26 PM	26		Mozilla/5.0 (Windows NT 6.1
31.184.238.12	/	5/31/11 2:05 PM	56901	http://www.jacerda.com/	Opera/9.0 (Windows NT 5.1;
31.184.238.12	/	5/31/11 2:05 PM	0	http://jacerda.com/	Opera/9.0 (Windows NT 5.1;
31.184.238.12	/foro/	5/31/11 2:05 PM	28321	http://www.jacerda.com/foro/	Opera/9.0 (Windows NT 5.1;
31.184.238.12	/foro/index.php	5/31/11 2:05 PM	0	http://jacerda.com/foro/index	Opera/9.0 (Windows NT 5.1;
208.115.111.7	/robots.txt	5/31/11 1:56 PM	24		Mozilla/5.0 (compatible; Ezo
208.115.111.7	/robots.txt	5/31/11 1:56 PM	24		Mozilla/5.0 (compatible; Ezo
124.115.0.141	/wp-includes/js/jque	5/31/11 1:26 PM	4041	http://www.jacerda.com/	Sosospider+(+http://help.so

Fig. 1. Web log structure.

In turn, depending on both the tools and mechanism used, two strategies to carry out the sessionization process are widely recognized (Spiliopoulou et al., 2003):

1. *Proactive strategy*: Consists of using invasive tools to identify users. This usually implies the use of cookies or spy-bots, which enable the identification of a user through a single identifier, in order for it to be possible to monitor their visiting frequency and to verify what extent their behavior has varied over time. The effectiveness of these techniques is not overly clear, especially since there are some pieces of software specializing in either deleting spy-bots and cookies, or stopping their function (Spiliopoulou et al., 2003).
2. *Reactive strategy*: Consists of using web logs as the sole data sources to rebuild sessions. Through this strategy the risk to a user's privacy may be lessened owing to that it is not necessary to ask for further user information. Despite the fact that this strategy provides less information due to the particular user not being thoroughly identified, this can be deployed onto any web site without extra costs.
   It should be noted that certain web sites have modified their structure with the aim of identifying their visitors. One strategy is to implement an ID system, and to encourage users to register on the site usually in exchange for new services. It is only possible to rebuild flawless sessions with registered users, given that non-registered users still remain somewhat anonymous.

A further strategy is by using dynamic web pages through which each visited page creates a single identifier per user (Eirinaki and Vazirgiannis, 2003). However, it compels one to rebuild the site and also creates a series of complexities in identifying what content is viewed by the users, considering that the URL addresses are dynamically generated.

IT professionals try to carry out their goals through tools that avoid linking a human persona to a web user as much as possible (Velásquez and Palade, 2008). In this sense, the process aims to ascertain user groups with similar navigation and content preferences without identifying the individual behind the session. Nevertheless, the extraction of both navigation patterns and user preferences can always be used as an indirect way of extrapolating the web visitor's behavior and therefore ascribing to him the characteristic of a group.

Finally, the last step consists of establishing how the extracted patterns should be used. From a strictly IT point of view, this "how" becomes "if-then-else" rules which, together with the patterns, comprise the knowledge which has been extracted from the web data. Once the knowledge has been obtained, a recommendation to the user should only be made once there is something to recommend.

3. Privacy in U.S. law

As most privacy scholars suggest, the starting point of the treatment of privacy in U.S. law was marked by the famous Warren and Brandeis article "Right to Privacy" which has survived more than 110 years since its first publication in the Harvard Law Review (Warren and Brandeis, 1890). In this article, Samuel Warren and Louis Brandeis, who was later to become a Supreme Court justice, confronted a not much different scenario than this paper seeks to address; the introduction of a previously unknown technology and the possibility of its misuse affecting an individual in some way.

In their case, the novelty was presented by the invention of the Snap Camera by EastMan Kodak in 1888, which was seen as an intrusive tool that could be used by journalists of the "yellow press" to take photographs of individuals in their homes without their awareness. The manner in which those authors framed privacy focused primarily on the existence of a "right to be alone", allowing individuals to prevent third parties from interfering in their private lives. Thus a need for legal protection of thoughts and emotions similar to that afforded to physical integrity was recognized.

In the 1960s, the prize-winning tort scholar William Prosser proposed a new privacy approach, which for the first time systematically considered several different ways in which privacy could be affected, and therefore, the issues that the concept of privacy implies. Thus, Prosser acknowledged not a single tort, "*but a complex of four different interests . . . tied together by the common name, but otherwise [with] nothing in common*" (Wacks, 2009): (1) intrusion upon the individual's seclusion; (2) public disclosure of embarrassing private facts; (3) publicity in a false light; and (4) appropriation of an individual's attribute (such as name or likeness) (Prosser, 1960).

Although Prosser's influence has been widely acknowledged, one example being the fact that the American Law Institute includes his fourfold classification in the second version of Restatement of Torts (Titus, 1977), some scholars in turn point out the negative side of his approach. On one hand it has been said that Prosser blurred a relevant part of the moral substrate that Warren and Brandeis sought to give to the treatment of privacy. On the other hand, it has been suggested that, by virtue of the prestige of his theoretical construction, the understanding of privacy in U.S. law has tended to stagnate. As Kalven wisely predicts in 60s ("*Given the legal mind's weakness for neat labels and categories and given the deserved Prosser prestige, it is a safe prediction that the fourfold view will come to dominate whatever thinking is done about the right of privacy in the future*".) (Kalven, 1966).

3.1. U.S. legal framework

Throughout the diversity of legal sources which comprise the U.S. legal system, certain rights, or better said, certain individual prerogatives related to privacy have been recognized, either by the federal Constitution, federal laws, state laws or other guidelines proposed by governmental agencies.

3.1.1. A constitutional right

Although a right of privacy has never been explicitly established, high courts have said that the Bill of Rights allows the securing of certain things under the concept of privacy,mainly by means of the interpretation of the First, Fourth and Fifth amendments (Volokh, 2000). Nonetheless, as a commentator suggests, "*[those opinions related to privacy have] little to do with any right of information privacy*". The main test used to determine what exactly the sphere of privacy is which would be worthy of protection under the Constitution has been to analyze when or where an individual has a "reasonable expectation of privacy". This approach has allowed judges both to recognize and assess privacy issues on a number of topics such as "matters relating to marriage, procreation, contraception, family relationships, child rearing, and education". As can be seen, this legal test can not by itself address the problem of what the value of privacy is which would deserve protection under law, but rather it reformulates concerns in terms of expectations that one can have in a particular context.

Another relevant aspect of the Constitutional treatment of privacy is that it focuses primarily on issues related to the possibility of the state, through the exercise of governmental powers, interfering or impinging in some manner upon an individual's personal life. Thus, little could be read from the Constitution on privacy issues in relations between individuals, as Volokh (2000) has pointed out: "*the Constitution says little about what private persons or*

businesses may or may not do; recall that the Bill of Rights starts with 'Congress shall make no law. . .' and the Fourteenth Amendment, which applies most of the Bill of Rights to state and local governments, starts with 'No state shall . . .' Whatever rights we might have against our business partners, none of these rights flow from the federal Constitution".

3.1.2. Federal laws

The U.S. has enacted quite a number of federal laws related to privacy issues, following a case by case basis (the so-called "sectoral law focus") (Marcella and Stucki, 2003), which are intended to tackle various particular aspects of privacy matters, unlike the comprehensive mechanism used elsewhere such as in the case of the EU community (Marcella and Stucki, 2003). As has been pointed out several times, the main weakness of such an approach is the high likelihood that the enacted laws would be quickly superceded by the introduction of new technologies. Another common criticism is the lack of a centralized privacy agency (Schwartz, 2004) (as some of the countries of the EU have). Although the amount of data that private parties can harvest on the Internet has steadily increased, legislative bodies have not paid enough attention to this situation, and thus the lack of legal bounds increases uncertainty. However, though there is no law which broadly regulates information privacy, at least two statutes can be mentioned that have been vital in protecting users' data processing. On one hand we have the Children's Online Privacy Protection Act (hereinafter COPPA) which provides safeguards to protect children's privacy on the Internet by regulating the collection of personal information from children under the age of 13. This law, departing from the general rule, includes a detailed definition of personal data, which we will analyze in detail in Part III. On the other hand, though it was primarily designed to prevent unauthorized government access to private communications, the Electronic Communication Privacy Act (hereinafter ECPA), under Title I and Title II, has allowed the application of federal law, in an enforceable manner, to private actors who fall under its scope.

According to the ECPA, in principle all private parties' interception of electronic communications is unlawful. Despite the above fact, Title II of the ECPA (which includes the Stored Communication Act) provides an exception to civil liability if there is prior consent from at least one party. Thus, in cases in which somebody interferes with a communication authorized either by "the person or entity providing a wire or communications service" or by "a user of that service with respect to a communication of or intended for that user", shall not be deserving of civil liability under the ECPA. That exception has proved to be relevant in finding or ruling out guilt under the ECPA of web advertising companies, as can be seen in such landmark cases as In re Doubleclick and In re Pharmatrack.

As we can see, the legal framework applicable to informational privacy issues between private parties is quite exiguous. One of the relevant things that can be derived is that, as Volokh (2000) has observed, contracts are the main legal tool to protect users' privacy, therefore affirming informed consent as being the key concept in deploying a suitable privacy policy.

3.1.3. Guidelines

Guidelines are compilations of best practices, frequently done by an advocacy agency, in order that actors within the industry assume both the principles and recommendations which these embody.

3.1.4. FIPPs

The FTC has collected through FIPP (Fair Information Privacy Principles) some of the according to them most broadly accepted principles regarding personal data processing in electronic markets: notice/awareness, choice/consent, access/participation, integrity/security, enforcement/redress. Those principles encompass a series of recommendations of good practices with special regard to privacy polices drafted by ITC businesses. For instance, with regard to the proper application of notice/awareness principles, it recommends describing an entity's information practices on a company's site on the web in a readily accessible, clear and conspicuous and above all understandable fashion. Another suggestion covered by the FIPP which is worth pointing out concerning the manner in which the information is acquired, said process could involve: (1) passive collection by electronic monitoring, as would be the case of web logs recording web usage data, or (2) actively asking the user to provide the information, as would be the case of information collected from web forms filled out by users.

FIPP principles have been strongly criticized both by privacy advocates and ITC sector businessmen. In fact, it has been claimed that these are less comprehensive than other guidelines such as the proposal of the Organization for Economic Cooperation and Development (OCDE, 2002) (which recognizes eight principles in lieu of only five developed by FIPP) and the Safe Harbor principles. Moreover, an ITC professional once said that rather than protect privacy, FIPP principles function as an encouragement to the movement of personal data (Bonner and Chiasson, 2005). It has also been said that the users' behavior assumptions on which FIPP rely heavily are erroneous, due to the impossibility of relevant numbers of users being able to assess risks and benefits in order to either give or deny their consent (Gross and Acquisti, 2005).

Finally, from the industry's point of view, FIPP recommendations are seen as too expensive to be readily deployable (Bonner and Chiasson, 2005).

To sum up, the only thing to be drawn from FIPP implementation is that neither its penetration nor its updates have worked as expected, even when it has been through three revisions since its first draft in 1977. The failure to consider the major concern about treatment of data independent of its current nature is an issue to bear in mind in future developments.

3.2. Personal identifiable information

A fundamental part of informational privacy is to determinate whether private data qualify as personal data or personally identifiable data. In general terms, this concept implies any information referring to an individual or allowing in some manner a connection with or identification of a particular person. Notwithstanding that in the context of U.S law there are a variety of definitions about what is and what could qualify as PII, not necessarily by just using the tag "personal information" (Bonner and Chiasson, 2005), that we suffer from lack of a generally applicable terminology to all interactions upon the web (cf. Edwards and Waelde, 2009).

However, some collections of personal data which flow on the internet are covered by COPPA, by which the notice and consent of parents are required in order to allow an operator to lawfully process the personal data of a child, unless certain exceptions provided by the law itself apply. It is worthwhile to note that the parental consent on collection, use and disclosure of a child's personal data must be verifiable, allowing for this purpose a reasonable effort by the operator given the available technology, to inform and request authorization from the parents or legal representatives.[1]

In COPPA, personal information means individually identifiable information about a child collected online, including:

[1] http://www.occ.gov/static/news-issuances/bulletins/rescinded/bulletin-2002-31.pdf.

1. a first and last name;
2. a home or other physical address including street name and name of a city or town;
3. e-mail address;
4. a telephone number;
5. a Social Security number;
6. any other identifier that the Commission (Federal Trade Commission) determines permits the physical or online contacting of a specific individual; or
7. information concerning the child or the parents of that child that the web site collects online from the child

Despite the fact that the definition given by COPPA only applies to collection of the personal information of children under age 13, various court decisions have used this definition of PII as a model to determine expectations of privacy in other contexts. Furthermore, it should also be noted that personal data or likeness concepts are commonly used in a large number of privacy policies by private parties, and in turn are a key point in some of the more relevant foreign legal systems. For instance, the main regulation in the EU community, the directive "EC 94/46 on the protection of individuals with regard to the processing of personal data and on the free movement of such data" (hereinafter Personal Data Directive or DPD) provides in Article 2a. a quite similar definition of PII.

When data protection laws are constructed around the concept of PII, it also implies identifying the relevant actors of the process in order to assign responsibilities and duties. First we must identify a "data subject" (the individual), who in turn has or should have a series of data control prerogatives, an "operator" or "data controller" (who decides which data will be processed) and a "data process" (the data treatment itself). Finally, there is a "data processor", who is the organization or individual that processes the data on behalf of another, the most common situation of a web miner.

3.3. A particular concern: are IP addresses PII?

As we have mentioned above, the web miner, who performs his activity in order to personalize web sites, has web logs as the main mining source. Inside web logs, inter alia, the IP addresses of visitors have been recorded. During the last few years there has been an intensive discussion about whether IP addresses qualify as personal data. Discerning the nature of IP addresses is relevant in order to apply data protection policies. This question is worthy of attention even when in the U.S there is no law applicable to all cases of IP address processing. This is based both on the fact that many companies, most of them industry leaders, include in their privacy policies a definition of personal data, and in turn the possibility that a foreign law may affect the operation of companies established in the U.S., as indeed might be the case regarding the European Union Directives. Positions on this matter could be readily grouped into two opposing views. The first argues that it is impossible to consider an IP address as personal data because it is not possible to identify with accuracy a single individual. This point of view is mainly expressed by those who defend the current treatment of this data by search engines and other related web services. Under this approach, we include the considerations expressed by Alan Davison, the privacy chancellor of Google Inc. For Davison, who said that a trace of information can be considered as PII, in the sense of producing a "reasonable link to an individual", it depends on whether the following conditions are met: "*Whether information can be 'reasonably linked' to an identifiable individual turns on (i) what the data itself is? And in particular how frequently it accurately and reliably describes an individual; (ii) what kind of additional information is needed to identify the specific person to whom that data relates; (iii) who has access to the additional data needed; and (iv) the circumstances under which the additional data will be made available to others*" (Lah, 2008).

Therefore, to companies such as Google, it is possible to state that an IP address is or not PII by virtue of the business model of who processes this information. As a result, if a private party does not have access to the complementary information which allows linking an IP to a real individual (as an ISP could do), the logs of IP addresses that he preserves do not qualify as PII.

Another nuance that could be emphasized is that, under Google's meaning of "reasonably linked", two factual circumstances can reinforce the idea that IP addresses are not PII: when the users have dynamic addresses and when the treatment is related to a non-authenticated user. In this sense, within the above-cited Davison letter, the following excerpt is illustrative:

"When an individual is not authenticated, we do not consider an IP address to be personally identifiable because we would need to get specific data from an ISP about which of its customers was using a particular IP address at a particular time on a particular day in order to link it to an individual. Even then, you could not say which member of a household was online at a particular time" (Bygrave, 2002).

On the contrary, privacy advocates suggest that IP addresses should be gathered as PII because they allow the identification of a single physical individual or at least allow a narrowing of the possibilities of doing so. The discussion on this issue has been enormously boosted mainly by one of the last decisions taken by the European Community advisory group on those matters created by DPD Directive, the Article 29 Working Party (Schreurs, 2008), and the subsequent reaction of the relevant industry actors. According to Opinion 1/2008 (WP 148, 2008), referring to data protection issues related to search engines, the Article 29 Working Party has said that "*unless an ISP is absolutely certain that the data corresponding to a user cannot be identified, all IP addresses should be treated as personal data to be on the 'safe side'*" (Marcella and Stucki, 2003). In order to reach their conclusion the advocacy group has taken special consideration of how search engines collect and conserve IP addresses, in order to finally determine what the possibility is that these can identify a particular individual (or a small group, but still identifiable people). The organism encourages search engines to adopt a series of recommendations which include measures such as records elimination, anonymization of gathered info (if it is not possible to destroy it completely), and setting expiration limits on the information collected of no more than 6 months.

The real impact of this opinion is still unknown, because despite the importance of this advocacy body, the EU countries are free to establish their own interpretations regarding these matters. Thereby, as a commentator has pointed out, "*perhaps the question should not be what will happen when the E.U. Member States adopt the Working Party's finding, but rather if they will adopt them*" (Lah, 2008).

Based on the most relevant topics pointed out in both positions, we consider that the analysis should be guided by the following points. We note that despite the fact that we borrow some ideas from the above opinions, our conclusions tend to be different.

1. A datum can be considered PII if one can at least significantly reduce the quantity of identifiable individuals.
   Personal data must lead us to a single individual. However, data protection should consider, in a preventive sense, all data that allows us to significantly reduce the quantity of potential persons to identify, for instance, in a case of the number of possible users being limited to inhabitants of a single house. It is not significantly different from the situation of other data such as telephone number, physical address and the like.

2. Consider what type of additional data must be related to the IP address in order to identify an individual.
   In order to protect IP addresses as pieces of information able to be linked to an individual, we must establish what kind of complementary data is needed to do so. In this case, it can be billing information, geolocalization records (often implemented by social networks) and some public records.
3. Consider who can access that complementary data and under what conditions.
   It is not only relevant to identify which complementary data is needed so that IP addresses can be used as ID systems, in fact some of these data can only be collected by a few limited agents, as in the case of billing information recorded by ISPs. Another factor to consider is how expensive it is for a private party to collect the information.
4. Consider which is the technological background: IPv4 or IPv6?
   A correct analysis of the nature of IP addresses should consider the relevant technical aspects of such data. The technical specifications of IP addresses are defined by the IP Protocol. This protocol has undergone a series of versions throughout its existence, version 4 (hereinafter IPv4) being the most widespread and thus the main basis of what we know as the Internet. Despite the success of this version in carrying out most of the requirements of such a complex web as the Internet, with the passage of time its inadequacy to meet the needs imposed by technological development has been noted. One of the most dramatic problems is the coming shortage of IP addresses. One solution to address that was to determine when a global identifier would be needed. The answer was simple whenever a package is delivered outside a local network. Thus, several devices within a local network could have their own private IP addresses, which do not exist on the internet, and share one public address through a router. Although efforts were made, it was always known that the above solution was temporary and that a new protocol should be developed. This change came with IPv6, which in reality is more an evolution than a substitution of the currently widespread version of the protocol. For example, it retains the idea of splitting information into datagrams which flow independently of each other, and it in turn solves the shortage of addresses by providing an extension of 128 bits namely 2 128 combinations. Also there are some characteristics of IPv6 which improve the performance between networks: e.g., a simpler header, space reduction and avoidance of intermediate fragmentation, among others. With respect to security features, IPv6 has a new authentication header and an improved Encapsulating Security Payload (ESP). The first provides both a correct authentication and an integrity system for datagrams, because it ensures that a package came from the "real" sender (declared by datagram source address). The second provides both integrity mechanisms and confidentiality through special fields where encryption keys are sent in order to establish a secured connection. To sum up, it could be said that overall the advantages of IPv6 are a greater number of addresses, improved security, better adaptability to new protocols and larger space to convey information (Kobsa, 2001). On the other hand, the disadvantages of IPv6 are fundamentally related to possible breaches of privacy. The new protocol by virtue of enhanced security standards implies new possibilities of identifying an individual with better accuracy, due to the fact that IP addresses include the information of a device identifier also know as a MAC address. This number is by definition unique in the world, despite being clonable or replaceable in some situations. Concerns increase if we consider that mobile devices or such like are often used by only one individual (Narten et al., 2013). Therefore, the chances of linking individuals through IP addresses are different when one compares IPv4 to IPv6. In the first case, we would need more complementary information, and often the integration of other mechanisms such as persistent cookies. On the contrary, under IPv6, IP addresses tend to resemble PII more closely without the need of a lot of complementary information to identify and link a person by means of an IP number.

4. A pragmatic informational privacy approach

Although one could consider that the rationale that privacy regulation is both diffused and insufficient results from the reluctance of decision makers to decidedly carry out an integrated discussion about privacy as a whole, often the problem lies mainly in how privacy is understood. As one commentator said, we cannot properly understand what privacy is without bearing in mind society and its current bounds. There is no chance to build protections surrounding our private life if we do not realize the changing nature of our social relations (cf. Young and Quan-Haase, 2009). While during former times the physical disruptions of our domestic life consumed the attention of individuals, at present the use of complex computer models in order to carry out a tailored profile of our behaviors is one of the most feared possibilities. One of the most famous pragmatic approaches is suggested by Daniel Solove. Solove, borrowing the terminology from Wittgenstein's language theory, states that privacy "*is an umbrella term that refers to a wide and disparate group of related things*" (Solove, 2006). Thus, the meaning of privacy is determined by a bottom-top system from facts to main related ideas. Under this approach, Solove proposes both a taxonomy of four different group of activities that involve an invasion of privacy and a non-closed list of relevant torts (Solove, 2006): Information Collection (such as surveillance and interrogation); Information Processing (which includes aggregation/insecurity/secondary use/exclusion); Information Dissemination (breach of confidentiality, disclosure, exposure, distortion); Invasion (intrusion and decisional interference).

A classification like the previous one allows us to identify for each human activity various issues related to privacy, as well as more readily detect what should be the set of possible solutions and stances to be taken by both individuals and decision makers.

In terms of informational privacy, the above taxonomy allows us to distinguish a number of problems that are generated along the vital circle of data, whether it is about personal data or not. In turn this implies the ability to dissociate each stage of the data processing process from the various actors, which is the tone of the current technological scenario.

For instance, consider the tort of informational aggregation and the tort of identification. In the author's words, informational aggregation means "*the gathering together of information about a person*" (Solove, 2006). In turn, identification can be understood as the process which "connects information to individuals" (Solove, 2006). In the digital era, both issues in a large percentage of cases are highly interrelated. A common internet user scatters a significant amount of data among various web sites. This information, as the mentioned author pointed out, creates an extension of personality on the internet, or better said a "digital person" which corresponds to "*a portrait composed of information fragments combined together*" (Solove, 2006). By an identification process, these profiles rich in texts, images and user preferences could be linked to a specific physical person (Solove, 2006). Nevertheless, it must be emphasized that the identification and aggregation in most cases are done by different agents. In this case, a question is raised about which of them must assume the main role in order to avoid connecting a physical individual with a particular digital profile. One might think that the better option would be to forbid both the gathering of information without consent and the possibility of its being linked without any awareness of it. But unfortunately

it is often not possible to reach such a level of control, especially if the aggregation either is not related to or is not triggered by the gathering of personal information, as would be the case of the application of a matching profile system built through, for instance, a web mining process.

An absolute prohibition is also an unrecommended solution, due to the need to protect other constitutional values such as the free speech of content providers. Thus, in these cases the relevant thing is to establish what circumstances result in a harmful output. Solove calls these problems "architectural problems", because "*They involve less the overt insult or reputational harm to a person and more the creation of the risk that a person might be harmed in the future*" (Solove, 2006). Architectural problems regarding privacy have a close resemblance to concerns regarding pollution (Solove, 2006) or the risks associated with tobacco consumption (Garrie and Wong, 2007). Taking into account the experiences that could result from both types of regulations, our approach must be focused on establishing which are the risks surrounding web mining and how these must be communicated to individuals in order that they can assess the benefits of an activity.

4.1. Profiling and tailoring

The mere fact of considering the possibility that the data which flows on the net can be "mined" by specific agents in order to extract behavior patterns will lead us straightaway to imagine a numberless amount of misuses, or at least marginally legitimate uses, such as indiscriminate governmental surveillance of our cyber life, disclosure of embarrassing habits based upon consumers' choices and the use of our opinions expressed in web forums to determinate our "suitability factor" for a specific job.

Thus, profiling processes have as their first generic objection the question of their main purpose. As they are known to be more than a collection of information, profiling systems seek to create large bases of knowledge, while their content and usefulness are unknown in advance in order to apply further commercial strategies. In this context, and in order that the knowledge produced be as faithful to reality as possible, the user is monitored through the least intrusive tools so that his normal activities are not interrupted. In other words, profiling and tailoring processes in a great number of cases are based, as Hildebrandt states, on the invisible visibility of users (Hildebrandt, 2009).

Another relevant topic is that often the users are unable to check on either the validity of the process or the congruency of the data used by the agent. This is enhanced by the fact that users do not have a recognizable link to the miner. Namely, there is no identification process, jeopardizing the views on this matter which emphasize a user-centric model, by which users should have enough power to control the use of their information.

Even though the greater part of these apprehensions have been validated by both actual evidence and the current state of information technology development, the technological phenomenon should not be analyzed solely by means of abstract insight without regarding that the relevant issue is determined by what behaviors are actually attempted against the individuals' rights. In fact, web tailoring may not only affect individuals in cases of either uninformed and unwanted use of personal information. Problems could involve minutiae, such as in the case of a Mozart music lover being erroneously recommended a Beethoven collection on a music web site, a situation in which the miner will have an incentive to quickly fix the problem in order to justify continuing the funding of mining processes. On the other hand terrible undesired situations might arise, such as the inclusion of someone in a classification of potential terrorists. However, the above must not lead us to state that certain informational technology should be prohibited, since as Brankovic and Estivill-Castro noted, the important thing is to determine what kind of decisions could be made as a result (Brankovic and Estivill-Castro, 1999): "*The issue of how data analysis may result in interpretations that may create or reinforce stereotypes is more an issue of the application of the technology than a problem of the technology. Is similar to attributing the impact on large audiences with political propaganda and biased media as a disadvantage of TV. This is not a disadvantage of the technology, but more a very complex issue of the social structures that allow the production and broadcasting of materials, later distributed through the powerful technology (...) [Thus] when does finding two groups justify creating different policies is not an issue of the technology that discovered the two clusters*".

Therefore, in discussing the effects of web mining tools upon privacy, we must have at least a basic archetypal concern about the context and the aims sought in their use. As a result, our scope will be focused specifically upon the implications of web personalization within the context of private relations, usually characterized as the interaction between web site owners and their users. Hence, we exclude from our focus the use of web mining tools by governmental agencies in order to effect the purposes that legislation may require, as well as the interaction which occurs within a labor context between employers and employees.

Therefore, we should be discerning about which risks are associated with commercial web mining. Whether or not we see the issue as a matter of level of harmfulness, it must be said that there is clearly a point at which web mining is dramatically opposed to individual interests. This is the case when private parties mine the web in order to build a system of "suitability factors" which enable discrimination against individuals based on a group which their profiles match.

Consider as an example a system to rule out certain people based on their hostile or unfriendly comments scattered among forum topics on a web. As Hindebrant observed in the context of the EU "*group profiles that are applied to me have been inferred from masses of (personal) data that are not mine (hiding my data will not stop the process of group profiling); second, they are applied because my (personal) data match the profile, which does not imply that the profile actually applies (the problem of non-distributive profiles); third, sophisticated profiling technologies like e.g., behavioural biometric profiling (BBP) do not require identification at all, thus falling outside the application of the directive [DPD directive]*" (Hildebrandt, 2009). Notice that if the information was considered to be personal data, individuals would receive better protection because the principles of no disclosure without consent, no secondary use, and the obligation to inform when information is gathered would be applicable. Thus, the user would be empowered enough to control to what extent he is willing to be the object of profiling.

Usually web sites include at the bottom of their pages a privacy policy which states what kind of data are collected, by what method and what uses will be made of it, including a brief description of the possibility of transferring them to third parties. Many times these privacy policies are drafted full of legalese, with both abstract and imprecise descriptions of purposes, and often include a changeability clause at the web site owner's discretion.

Another problem which has been posed is how sites acquire a user's consent in order for these privacy policies to constitute lawful contracts between site owners and users. Because of this, the industry often defends the existence of a browsewrap contract, namely, a contract that is perfected by the mere fact of user navigation through a web site. The opposite form is the clickwrap contract which depends on a sequence of clicks in order to give the user's acceptance to the terms and conditions of service. These contracts are the basis of the End User License Agreements (EULA) used by software companies.

Whether or not we understand, as we have described earlier, that a web mining process could have different harmfulness levels,

it is possible to establish different types of privacy policies by virtue of which kind of information is gathered and what the purpose is of doing so. In cases where the web mining process is focused on weblogs to improve the site content and the ease of navigation, a clear privacy policy which describes what information is inside logs, and in turn avoids the creation of a permanent link to users. In other words avoids establishing an ID system, could seek consent through a browsewrap contract? It should be noted that this would be despite the fact that the information is gathered automatically due to both the security and data integrity of the server (Dinant, 2009). On the other hand, whenever the mining process involves the possibility of an individual being the object of a suitability test or something similar, it is more advisable to establish a clickwrap license of use which allows the explicit authentication of the user so that he can control the supplied information and the subsequent uses of it.

4.2. Trade-off of privacy

According to Oxford Dictionary of Economics trade-off is "*the requeriment that some of one good or one objective has to be given up to obtain more of another*" (Black et al., 2009). Privacy, as well as most interests that deserve legal protection, are not absolute concepts. When compared with other values such as national security (Solove, 2008), informational utility (Clifton et al., 2002; Dutta et al., 2011), and commercial benefits (Cohen, 1999), it must be weighted in order to establish equilibrium among different factors. Both the ITC industry and the on-line advertisement business have put emphasis on establishing the optimal level of privacy for a given situation through an analysis of the type of "cost-benefit". Under this criteria, a comparison of the personal benefit of privacy with the benefits for society of the availability of information is intended (Wilson, 2006). In the case of using data mining tools, it has been said by some ITC professionals that "*even modest privacy gains require almost complete destruction of the data-mining utility*" (Brickell and Shmatikov, 2008).

Sharing the idea that privacy must be weighted in each context against other values (such as access to cheaper goods by means of having more information available), we believe that the contrast must be based upon comparable criteria, so that none of those values will deserve *a priori* more protection, in effect rendering the assessment exercise irrelevant.

To Solove, the understanding of privacy as a private right as such is a flawed option: "*Current approaches that emphasize privacy as a personal right fail to capture its importance because they obscure society's interest in privacy protections*" (Solove, 2006). This implies that at the moment of assessing privacy requirements, this should be seen as a value which is of interest to society as a whole, as well as having safety, informational utility and other values.

Otherwise, another ITC scholar, acknowledging the problem of how to assess privacy requirements (Li and Li, 2009), has stated that: "*It is inappropriate to directly compare privacy with utility, because of several reasons, including both technical and philosophical ones. The most important reason is that privacy is an individual concept, and utility is an aggregate concept. The anonymized dataset is safe to be published only when privacy for each individual is protected; on the other hand, utility gain adds up when multiple pieces of knowledge are learned*" (Li and Li, 2009).

Even though at first the two opinions could be seen as opposing, at their root they pretend to tackle the same situation with a shared common idea? That the current widespread comparison between privacy and other values is both flawed and biased. Moreover, we would say that in this case, the Li and Li stance ends up being a practical application of Solove's understanding of privacy, because although the former put emphasis on the individual aspect of privacy, it is possible to conclude that in his opinion privacy is totally protected only if the privacy of each individual is ensured, namely, that privacy is completely meaningful only at a societal level.

When it comes to distinguishing privacy we also must take into account the dualism of two effects, the individual and the aggregate. When we consider the social benefits of ready access to more goods, we tend to rely on general statements often based on the underlying assumptions of a free market economy. However, little effort is invested in clarifying just what are the specific benefits to individuals of a particular technology. This also leads to a user's unwillingness to hand over personal information, because they see no short term benefits in doing so.

Even more, the web user sees in the use of mining tools a kind of surveillance without any distinction between private parties or government agencies. The common image of our technological scenario is something similar to an Orwellian paradox but, as Solove subtly appreciated, rather than the existence of a big brother watching you all the time, we are immersed in a Kafkaesque process, one in which we lack sufficient awareness of what is happening around us Bélanger and Crossler (2011). But this understanding of reality could be changed if contents providers assumed the task of explaining to users.

Hence, for the semantic web to have active user participation, the ITC business should consider privacy-protection as a fundamental element of its strategy to reach more users, because in this era of digital risk, the providers which offer better warranties will end up receiving greater community acceptance.

4.3. Bounds of data protection laws

As we have already briefly reviewed, when it regards data related to our person, data protection laws, although never without pitfalls, try to both avoid the misuse of our data and give us the power of control over how it is used and transferred. Although this approach to carrying out issues surrounding data which flows between individuals seems to be suitable, it has, by its own definition, a stumbling block; often the data must qualify as personal data or personal identifiable data in order to fall under the scope of the law. But some readers could be thinking, why is it a problem? What is the reason for not considering data protection to be a sufficient legal implementation of informational privacy, without suggesting that this is in some manner a flawed option? In order to answer that question one idea must first be emphasized. We believe that data protection, especially when it is correctly implemented, is an adequate option to redress some threats related to informational privacy. However, when certain concerns fall outside of its scope, whether or not we have a context-reactive understanding of privacy, we will be unable to correctly assess the risks and benefits of a technological paradigm. In order to see the problem, take a look at the following excerpt of an article posted by Alma Whitten,[2] a Google Inc. software engineer, on the company's public policy blog: "*To protect privacy, you first have to identify what data is personal. That's why there has been a lot of discussion in recent months around the world to try to define "personal data" (as it is referred to in Europe), or 'personally identifiable information' (as it is called in the U.S.)*".

This view entails serious risks to users, as based on an all or nothing criterion, privacy on the net would be defined only by the existence of the treatment of personal data. One can suggest that to fix the possible narrowed scope of privacy laws, it is sufficient to have a definition of PII flexible and modular enough to allow it to be applied to as many matters as possible. We believe

[2] http://googlepublicpolicy.blogspot.com/2008/02/are-ip-addresses-personal.html.

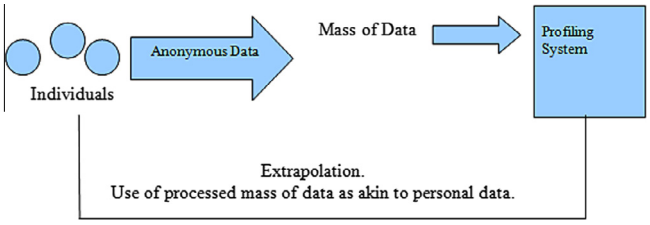

Fig. 2. Double dimension of informational privacy. Personal data and the use of processed data as if those were personal data.

that this proposal should be rejected for a number of reasons both technical and legal, with some of these highlighted below.

1. Data protection laws are focused only on individuals, i.e. rely on identifying who is the data subject and who is a data controller. If a group of people are completely identified but not limited enough, with a high likelihood we are beyond the scope of these laws.
2. Data protection laws rely on prior and informed consent as the main mechanism of liability exception and lawful authorization in order to process personal data. However, this does not solve the problem of the use of recommendations given by a system regarding an individual due to their rank within a particular profile. Even if the individual avoided using their personal data the profiling process would not stop.
3. Data protection laws impose certain restrictions on the data controller. Therefore, those data controllers should avoid the use of personal data as much as possible. But if most of the information qualifies as PII, business development tends to be excessively expensive and consequently research into those matters tends to be restricted. As we have noted, a large number of issues regarding web personalization are based on architectural problems, i.e. problems which enhance risk in the future and in turn are not easy to quantify at present. The regulatory experience has shown us that command and control regulations are mainly ineffective in controlling industries in continual growth, and usually end up being quickly outdated (Hirsch, 2006).

For these reasons, in Fig. 2 we propose that informational privacy must be composed of two groups of questions of a quite different nature but with common elements. On one side data protection laws which govern personal data treatment, giving to users the power to control the data which flows from them. On the other side the use by others of data on individuals, whether these sets of information are personal data or not. This is the case of profiling systems based on masses of collected data. The For

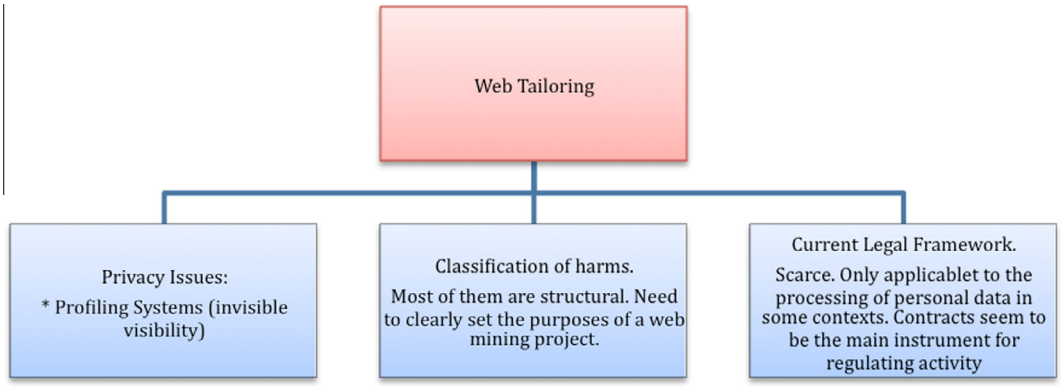

Fig. 3. Nature of privacy concerns regarding web mining.

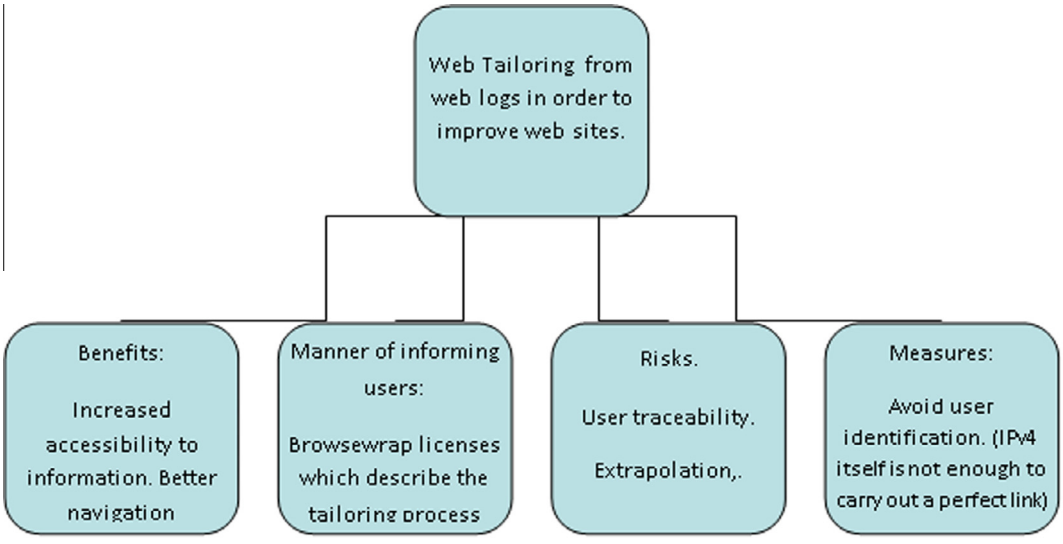

Fig. 4. Web mining model 1: tailoring the web as from web logs in order to improve web sites.

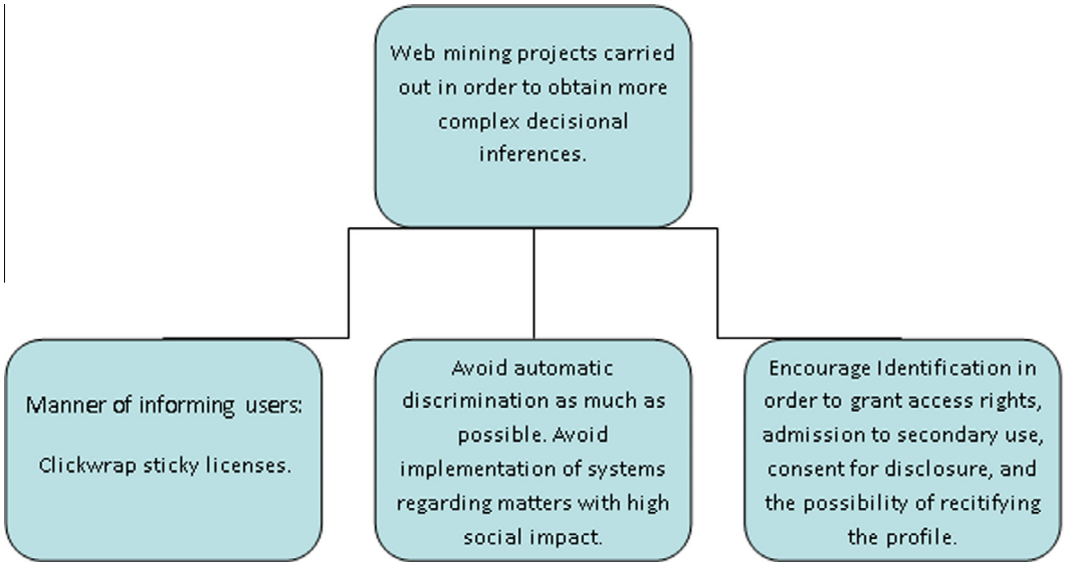

Fig. 5. Web mining model 2: web mining projects related to decisional inferencing systems.

these reasons, we propose that informational privacy must be composed of two groups of questions of a quite different nature but with common elements. On one side data protection laws which govern personal data treatment, giving to users the power to control the data which flows from them. On the other side the use by others of data on individuals, whether these sets of information were personal data or not. This is the case of profiling systems based on masses of collected data. The regulation of this activity should be focused on limiting the kind of decisions that by those means could be made by someone else.

5. Conclusion: our proposal

Taking into account the elements described throughout this work, at this point we can outline which are the most relevant privacy issues regarding web personalization from a user perspective.

As we mentioned above, a web mining process considers ups rebuilding an approximately a user session. This information, often anonymous, either by its nature or as a consequence of an anonymization process, can be used as an input for developing profiling systems, an issue which could be perceived by individuals as an undercover surveillance mechanism. The result of those profiling systems, with the help of complementary user-tracking technologies, can result in the persistent identification of a user on the web, who certainly will not be able to repudiate the qualities that are ascribed him because he lacks knowledge of both the information and mechanisms used. This is the issue that we have described as the problem of using anonymous or group data as if they were personal data, which brings us to the apparent paradox that the user would be more protected if he voluntarily gave personal information.

Those risks to users' privacy are so integral to the development of a web mining process that they deserve to be considered as structural problems. In this case, the core of the question lies in the level of awareness that a user has or should have about the existence, purposes and data collection systems used in a certain process, as is shown in Fig. 3.

Considering the current legal framework applicable to privacy affairs between private parties, the most adequate method of protecting users would be by means of contractual remedies that could be provided. The contract structure must be suited to the specific characteristics of the mining project, taking care that this does not involve excessive costs for the miner, whose activity, carried out correctly, generates important social benefits.

In order to graphically illustrate the distinct alternatives, we shall define two basic categories of web mining projects: (1) Those projects based on the mining of web logs with the intention of improving the navigation experience within a certain web site; and (2) the use of mining tools upon web data in order to make more complex inferences about an individual's attributes (see Figs. 4 and 5).

In the case of (1) the publication of a clear privacy policy which details both the purposes and the pattern extraction techniques would seem sufficient. Consent would be established by means of a browse wrap contract. It must be noted however, that the data miner should take care to not introduce any unnoticed ID mechanisms, and at the same time establish an anonymization policy and a preset expiration date.

On the other hand, in the case of (2), the recommendations are substantially different. In those cases, the data miners should attempt to get user consent through more complex forms, such as the use of clickwrap contracts. Due to the use the data will be given, it is more advisable to encourage rather than avoid user registration. This is because a unique ID will give users the chance to opt to be part of the process or simply stop using the service.

Finally, regarding the aims of a project, the web miner should take care to not use available technology to obtain automated decisions about individuals regarding topics with a high social impact, as in the development of a labor discrimination system by suitability factors.

Acknowledgement

This work was supported partially by the FONDEF Project D10I-1198 entitled, WHALE: Web Hypermedia Analysis Latent Environment and the Millennium Institute on Complex Engineering Systems (ICM: P-05-004-F, CONICYT: FBO16).

References

Ashworth, L., & Free, C. (2006). Marketing dataveillance and digital privacy: using theories of justice to understand consumers' online privacy concerns. *Journal of Business Ethics, 67*, 107–123.

Bélanger, F., & Crossler, R. E. (2011). Privacy in the digital age: a review of information privacy research in information systems. *MIS Q, 35*, 1017–1042.

Black, J., Hashimzade, N., & Myles, G. (2009). *A dictionary of economics*. USA: Oxford University Press.
Bonner, W., & Chiasson, M. (2005). If fair information principles are the answer, what was the question? An actor-network theory investigation of the modern constitution of privacy. *Information and Organization, 15*, 267–293.
Brankovic, L., & Estivill-Castro, V. (1999). Privacy issues in knowledge discovery and data mining. In *Proceedings of Australian Institute of Computer Ethics Conference (AICEC99)*.
Brickell, J., & Shmatikov, V. (2008). The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 70–78). ACM.
Bygrave, L. (2002). *Data protection law: approaching its rationale. Logic and limits*. The Netherlands: Kluwer Law International.
Cavoukian, A. (2008). Privacy in the clouds. *Identity in the Information Society, 1*, 89–108.
Clifton, C., Kantarcioglu, M., & Vaidya, J. (2002). Defining privacy for data mining. In *Proceedings of the national science foundation workshop on next generation data mining* (pp. 126–133).
Cohen, J. (1999). Examined lives: informational privacy and the subject as object. *Stanford Law Review, 52*, 1373–1417.
Cooley, R., Mobasher, B., Srivastava, J., et al. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems, 1*, 5–32.
Craig, B. (2012). *Cyberlaw: the law of the internet and information technology*. Pearson.
Dinant, J. (2009). The concepts of identity and identifiablity: legal and technical deadlocks for protecting human beings in the information society? *Reinventing data protection?* (pp. 111–122).
Dutta, S., Dutton, W., & Law, G. (2011). The new internet world: a global perspective on freedom of expression, privacy, trust and security online. <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1810005>.
Edwards, L., & Waelde, C. (2009). *Law and the internet*. UK: Hart Publishing.
Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for web personalization. *ACM Transactions on Internet Technology (TOIT), 3*, 1–27.
Garrie, D., & Wong, R. (2007). The future of consumer web data: a european/us perspective. *International Journal of Law and Information Technology, 15*, 129–152.
Gross, R., & Acquisti, A. (2005). Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on privacy in the electronic society WPES '05* (pp. 71–80). New York, NY, USA: ACM.
Gutwirth, S. (2009). *Reinventing data protection?* Springer Verlag.
Hildebrandt, M. (2009). Who is profiling who? Invisible visibility. *Reinventing data protection?* (pp. 239–252).
Hirsch, D. D. (2006). Protecting the inner environment: what privacy regulation can learn from environmental law. *Georgia Law Review, 41*, 1–63.
Kalven, H. (1966). Privacy in tort law: were warren and brandeis wrong? *Law and Contemporary Problems, 31*, 326–341.
Kobsa, A. (2001). Generic user modeling systems. *User Modeling and User-Adapted Interaction, 11*, 49–63.
Lah, F. (2008). Are ip addresses personally identifiable information? *ISJLP, 4*, 681–693.
Li, T., & Li, N. (2009). On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 517–526). ACM.
Lundevall-Unger, P., & Tranvik, T. (2011). IP addresses – just a number? *International Journal of Law and Information Technology, 19*, 53–73.
Marcella, A., & Stucki, C. (2003). *Privacy handbook: guidelines, exposures, policy implementation, and international issues*. Willey.
Markov, Z., & Larose, D. (2007). *Data mining the web: uncovering patterns in web content, structure, and usage*. Willey-Interscience.
Narten, T., Draves, R., & Krishnan, S. (2013). *Privacy extensions for stateless address autoconfiguration in ipv6*. <http://tools.ietf.org/html/rfc4941.html> Accessed 01.01.13.
OCDE (2002). Guidelines on the protection of privacy and transborder flows of personal data. Organisation for Economic Co-operation and Development.
Prosser, W. L. (1960). Privacy. *California Law Review, 48*, 383–423.
Schreurs, W. e. a. (2008). Cogitas, ergo sum. The role of data protection law and non-discrimination law in group profiling in the private sector. In S. Gutwirth & M. Hildebrandt (Eds.), *Profiling the european citizen. Cross-disciplinary perspectives* (pp. 241–264). Dordrecht: Springer.
Schwartz, P. M. (2004). Property, privacy, and personal data. *Harvard Law Review, 117*, 2055–2128.
Solove, D. (2008). Data mining and the security-liberty debate. *The University of Chicago Law Review*, 343–362.
Solove, D. J. (2006). A taxonomy of privacy. *University of Pennsylvania Law Review, 154*, 477–560.
Spiliopoulou, M., Mobasher, B., Berendt, B., & Nakagawa, M. (2003). A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *INFORMS Journal on Computing, 15*, 171–190.
Titus, H. (1977). Restatement (second) of torts section 402a and the uniform commercial code. *Stanford Law Review, 22*, 713.
Velásquez, J. D., & Palade, V. (2008). *Adaptive web sites: a knowledge extraction from web data approach*. IOS Press.
Volokh, E. (2000). Personalization and privacy. *Communications of the ACM, 3*, 84–88.
Wacks, A. (2009). *Privacy: a very short introduction*. OUP Oxford.
Warren, S., & Brandeis, L. (1890). The right to privacy. *Harvard Law Review, 4*, 193–220.
Wilson, J. (2006). Health insurance portability and accountability act privacy rule causes ongoing concerns among clinicians and researchers. *Annals of Internal Medicine, 145*, 313–316.
Young, A., & Quan-Haase, A. (2009). Information revelation and internet privacy concerns on social network sites: a case study of facebook. In *Proceedings of the fourth international conference on communities and technologies* (pp. 265–274). ACM.