*Ethics and Information Technology* **6:** 129–140, 2004.


# Ethical issues in web data mining

Lita van Wel and Lambèr Royakkers*
*Department of Philosophy and Ethics of Technology (*Author for correspondence: Department of Philosophy and Ethics, Faculty of Technology Management, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands; Fax: +31-40-2444602; E-mail: L.M.M.Royakkers@tm.tue.nl)*

**Abstract.** Web mining refers to the whole of data mining and related techniques that are used to automatically discover and extract information from web documents and services. When used in a business context and applied to some type of personal data, it helps companies to build detailed customer profiles, and gain marketing intelligence. Web mining does, however, pose a threat to some important ethical values like privacy and individuality. Web mining makes it difficult for an individual to autonomously control the unveiling and dissemination of data about his/her private life. To study these threats, we distinguish between 'content and structure mining' and 'usage mining.' Web content and structure mining is a cause for concern when data published on the web in a certain context is mined and combined with other data for use in a totally different context. Web usage mining raises privacy concerns when web users are traced, and their actions are analysed without their knowledge. Furthermore, both types of web mining are often used to create customer files with a strong tendency of judging and treating people on the basis of group characteristics instead of on their own individual characteristics and merits (referred to as de-individualisation). Although there are a variety of solutions to privacy-problems, none of these solutions offers sufficient protection. Only a combined solution package consisting of solutions at an individual as well as a collective level can contribute to release some of the tension between the advantages and the disadvantages of web mining. The values of privacy and individuality should be respected and protected to make sure that people are judged and treated fairly. People should be aware of the seriousness of the dangers and continuously discuss these ethical issues. This should be a joint responsibility shared by web miners (both adopters and developers), web users, and governments.

**Key words:** ethics, individuality, KDD, privacy, web data mining

## Introduction

The World Wide Web can be seen as the largest database in the world. This huge, and ever-growing amount of data is a fertile area for data mining research. Data mining[1] is the process of extracting previously unknown information from (usually large quantities of) data, which can, in the right context, lead to knowledge.[2] When data mining techniques are applied to web data, we speak of web-data mining or web mining. In accordance with Kosala, Blockeel and Neven (2002), we define web mining as *the whole of data mining and related techniques that are used to automatically discover and extract information from web documents and services*, based on the definition of Etzioni (1996).

The important ethical issue with data mining is that, if someone is not aware that the information/knowledge is being collected or of how it will be used, he/she has no opportunity to consent or withhold consent for its collection and use. This *invisible information gathering* is common on the Web. Knowledge discovered whilst mining the web could pose a threat to people, when, for instance, personal data is misused, or is used for a purpose other than the one for which it is supplied (*secondary use*). This same knowledge, however, can bring lots of advantages. Knowledge discovered through data mining is important for all sorts of applications involving planning and control. There are some specific benefits of web-data mining like improving the intelligence of search engines. Web-data mining can also be used for marketing intelligence by analysing a web user's on-line behaviour, and turning this information into marketing knowledge.

[1] In this study, the term 'data mining' refers to the entire Knowledge Discovery in Databases process (KDD). In KDD jargon, data mining is just one step in the entire process. The term 'data mining,' however, is often used to refer to the entire process, and, as there is no common use of a term like 'Knowledge Discovery in the Web' (as a special database), we shall use the more commonly used terms 'data mining' and 'web-data mining.'

[2] More detailed descriptions of data-mining techniques can be found in Berry and Linoff (2002).

It should be noted that ethical issues can arise from mining web data that do not involve personal data at all, such as technical data on cars, or data on different kinds of animals. This paper, however, is limited to web-data mining that does, in some way, involve personal data. We shall only look at the possible harm that can be done to people, which means that harm done to organisations, animals, or other subjects of any kind fall beyond the scope of this study. Since most web-data mining applications are currently found in the private sector, this will be our main domain of interest. So, web-data mining involving personal data will be viewed from an ethical perspective in a business context. We clearly recognise that web-data mining is a technique with a large number of good qualities and potential. Web-data mining is attractive for companies because of several reasons. For instance, to determine who might be a new customer by analysing consumer data, government records, and any other useful information. In the most general sense, it can contribute to the increase of profits be it by actually selling more products or services, or by minimising costs. In order to do this, marketing intelligence is required. This intelligence can focus on marketing strategies and competitive analyses, or on the relationship with customers.[3] The different kinds of web data that are somehow related to customers will then be categorised and clustered to build detailed customer profiles. This not only helps companies to retain current customers by being able to provide more personalised services, but it also contributes to the search for potential customers.[4] So, it is beyond dispute that web-data mining can be quite beneficial to businesses. To make sure that this technique will be further developed in a properly thought-out way, however, we shall focus on the possible objections to it. Awareness of all the possible dangers is of great importance for a well-guided development and a well-considered application. Dangers of web-data mining lie in the different ways in which privacy is threatened.

To structurally analyse the many different ways to mine the web, we need to be able to distinguish between these different forms of web-data mining. We can distinguish between actual data on web pages, web structure data regarding the hyperlink structure within and across web documents, and web log data about regarding the users who browsed web pages. Therefore, in accordance with Madria et al. (1999), we shall divide web-data mining into three categories.

1. The category of *content mining*, which is to analyse the content data available in web documents. This can include images, audio file, etc. In this study, however, content mining will only refer to mining text.
2. The category of *structure mining*, which focuses on link information. It aims to analyse the way in which different web documents are linked together.
3. The category of *usage mining*. Usage mining analyses the transaction data, which is logged when users interact with the web.[5] Usage mining is sometimes referred to as 'log mining,' because it involves mining the web server logs.

Structure mining is often more valuable when it is combined with content mining of some kind to interpret the hyperlinks' contents. As content and structure mining also share most of the same advantages and disadvantages (as we shall see below), we shall discuss them together, and consider them as one category. Web usage mining is quite distinct in its application. As it has also different advantages, and threatens values in a different way, we shall discuss it separately.

## Privacy threatened by web-data mining

In this section, we shall point out that web-data mining, which involves the use of personal data of some kind, can lead to the disruption of some important normative values. One of the most obvious ethical objections lies in the possible violation of peoples' (informational) privacy. Protecting the privacy of users of the Internet is an important issue. Our understanding of privacy, however, is conceptually fragile. The term 'privacy' is used to refer to a wide range of social practices and domains (cf. Johnson 2001). In this article, we will not discuss the philosophical and legal discussions on privacy. Here, we will restrict ourselves with an informal (and common) definition of informational privacy. Informational privacy mainly concerns the control of information about oneself. It refers to the ability of the individual to protect information about himself. The privacy can be violated when information concerning an individual is obtained, used, or disseminated, especially if this occurs without their knowledge

[3] See e.g., Philips, Vriens and Kienhorst (1999) and Fordham, Riordan and Riordan (2002).

[4] Custers (2002) gives an enumerative description about the advantages of group profiles: group profiling can be a very useful method for (Internet) companies to find target groups, when using group profiles there are more possibilities to identify potential customers than when using individual profiles only, and group profiling is – in most cases – more useful than not profiling at all.

[5] For specific advantages of usage mining we refer to Srivastava, Cooley, Deshpande and Tan (2000).

or consent. Privacy issues due to web mining often fall within this category of informational privacy. Therefore, we will be focusing on this category. In the remainder of this article the term privacy will be used as referring to informational privacy. When the judgement and treatment of people is based on patterns resulting from web-data mining, however, the value 'individualism' could also be violated. Let us first clarify the relationship between privacy and individualism. When information is discovered through web-data mining, someone's privacy might be directly violated in the process. When the information is classified and clustered into profiles, and then used for decision-making, people could feel violated in their privacy. When the data is made anonymous before producing the profile, however, the discovered information no longer links to individual persons, and there is no direct sense of privacy violation because the profiles do not contain 'real' personal data (cf. Vedder 2001). These group profiles can, however, be used as if they were personal data, which leads to the unfair judgement of people – referred to as de-individualisation (see following section). Therefore, privacy can be seen as a stepping-stone for other basic values. Or, as Vedder (2000: 452) puts it: "... privacy is a servant of many master values." A solution would be to extend the definition of privacy by referring to categorical privacy, which would allow group characteristics that are applied as if they were individual characteristics to be seen as personal data. Tavani (1999b), however, believes that such a solution would not be suitable as it might lead to the need for new privacy categories with the introduction of new technologies.

There are some differences between privacy issues related to traditional information retrieval techniques, and the ones resulting from data mining. The technical distinction between data mining and traditional information retrieval techniques does have consequences for the privacy problems evolving from the application of such techniques (cf. Tavani 1999a). While in traditional information retrieval techniques one has to 'talk' to a database by specifically querying for information, data mining makes it possible to 'listen' to a database (cf. Holsheimer 1999). A system of algorithms searches the database for relevant patterns by formulating thousands of hypotheses on its own. In this way, interesting patterns can be discovered in huge amounts of data. Tavani (1999a) argues that it is this very nature of data mining techniques that conflicts with some of the current privacy guidelines as formulated by the OECD.[6] These guidelines correspond to the European Directive 95/46/EC of the European Parliament and the Council of 24 October 1995. Every European Union member state has to implement the Directive in their national laws. The work – the implementation of the Directive's definitions and principles brings with it – should not be underestimated. Principles like the "use limitation principle" and "the purpose specification principle" state that information cannot be used for other purposes than the one it was originally retrieved for, and that this purpose has to be clearly explained to the person whose data is being mined before the actual collection. One of the features of data mining, however, is the fact that it is not clear what kind of patterns will be revealed. This makes it impossible to clearly specify the exact purpose and notify the data-subjects in advance. Besides, data mining is often performed on historical data sets. This, too, makes it rather difficult to comply with these guidelines.

*Content and structure mining*

When we look at the way in which the different types of web data can be collected and analysed by data mining tools, we have to conclude that it is difficult to comply with the current guidelines too. In addition, another problem arises: content and structure data are publicly available. People might have placed certain bits of information on their homepage for certain purposes and within a certain context. When web data is mined, however, it can be used for totally different purposes, taking it completely out of context. Because most web data has been made public by web-users' own leave, it is debatable whether this kind of information deserves to be protected at all. Nissenbaum (1997) argues that it is wrong to assume that an aggregation of information cannot violate privacy if its parts, taken individually, do not. This is based on the assumption that a single fact can take on a new dimension when it is combined with another fact, or when it is compared with similar facts. Even non-identifiable data can become identifiable when merged. Certain bits of data that are not considered to be privacy violating, can be collected and assembled with other data, leading to information that can be regarded as harmful (cf. Fulda 2001).

*Usage mining*

With usage mining, the public debate has to take on an extra dimension. Web log data are not publicly available. Yet, the data do represent someone's actions

[6] In 1980 the OECD (Organization for Economic Cooperation Development) formulated some internationally accepted principles regarding the collection, use, and unveiling of personal data.

within a public environment. Let us compare it to a shopping street. Web users are the people who move around in this shopping street. When entering a store, a person will still (find himself to) be in a public area because everybody can freely enter the store. The area, however, is confined to a single store instead of an entire shopping street. So, only people who are inside the store can see and watch this person. Let us imagine that there is nobody else in the store except for the people who work there. Those people are the only ones that can observe him/her. A web site could be seen as such a store, with only one visitor and the employees inside. Once a web user enters a web site, the people who manage and own the site are able to observe his/her steps. When there are video cameras in every corner of the store, a visitor loses his privacy. When the store only uses cameras to protect themselves against burglars this is not considered to be a violation. Consider the possibility of the store using the cameras to precisely record what products each customer looks at, and for how long. An owner of such a store could also decide to try and relate that information to some more personal details the owner has recorded, like the gender of the visitor, the clothes he/she is wearing, the kind of haircut he/she has, or the colour of his/her skin. Once a person is at a web site, certain data are logged. This data, however, is not the personal kind that can be recorded with a video camera. Web log data do not actually identify a person, but they do identify a web user: a user who has characteristics like an IP-address, date and time of entering and leaving the site, path of followed hyperlinks (click stream), type of browser used, and so on. So they do not know his name or what he looks like. But, the next time he visits the web site, he will be recognised as a regular visitor through the use of cookies.[7]

ISPs[8] can directly link all web surfers' behaviour to their personal data. When a web user starts his surfing session, an ISP knows who this user is because they are the ones that provide him/her with access to the Internet. In order to obtain an access account, the web user had to provide the ISP with some of his/her personal data. Every time he/she uses the web, he/she browses via the ISP's server. In this manner, they can monitor all the moves he makes. The Dutch law for lawful interception obliges ISPs to keep logs of all those user transactions in case the government will need them for criminal investigations. However, law also restricts ISPs. In the Netherlands, ISPs are legally regarded as telecommunications service providers, and they have to obey the new Dutch Telecommunications Law[9] (since 1998). This law stipulates that telecommunication service providers are obliged to erase, or make anonymous, all traffic data relating to subscribers upon termination of the call. The only exceptions to this rule are for traffic data that are necessary for the purpose of subscriber billing, marketing purposes (provided that the subscriber has given his consent), control of telecommunications traffic, supply of information to the customer, the settling of billing disputes, detection of fraud, or otherwise allowed (or even obliged) by legislation. Note that mining for marketing purposes is only allowed with a customer's consent. An ISP can specifically ask for this consent when a new customer subscribes. However, not all users are able to foresee all the consequences of agreeing to the use of their data for marketing goals. And, as explained before, it is difficult to clearly state the purpose of web-data mining analyses due to their exploratory nature. This could lead to potentially privacy-violating situations. Most ISPs, however, have privacy guidelines as part of their code of conduct that they have to honour. They can indeed connect personal data to their web log data, but they are limited in their freedom to analyse the data and act on it.

There is another type of web usage mining: the use of forms on the web. Lately there has been a tendency to trade information *quid pro quo* (see Custers 2000). Web users often have to fill out a form to simply gain access to a web site. Or, there are fields to be filled in on online ordering forms, which are of no relevance to the purchase. When a user fills in an online form of any kind, the data he/she shares can be used for customer profiling. By sending back the form, the web log also registers the current IP address of the web user, and his/her personal data can, therefore, be linked to his/her browsing behaviour on that particular web site. Although a user decides for himself/herself whether

[7] Cookies are small files that are placed on the hard disk of the web user during his browser session. The cookie will make sure that the web user's computer will be recognised and identified the next time the same web site is visited. Therefore, cookies can be used to track a user online, and enable the creation of a profile without him/her realising it. Some site owners allow advertisers to place banners referring to their own (ad)server on a web site. By loading the banner ad, a cookie is placed on the web user's computer. In this way, online advertising companies are also able to track (part of) a user's movements on the web.

[8] An Internet Access Provider (IAP) provides Internet Access (e.g., PPP, SLIP, shell, or UUCP accounts) and, possibly, other services. An Internet Service Provider (ISP) provides one or more basic Internet services such as Internet access, web site hosting, or DNS support for domains. In casual conversation, IAP and ISP are interchangeable terms. As the term ISP is more commonly used, we shall speak of ISPs, even though in a strict sense we actually refer to IAPs.

[9] The Dutch Telecommunications Act came into force in 1998. The sections (in English) can be found at http://www.minvenw.nl/dgtp/home/wetsite/engels/index.html.

or not he/she will fill in a form, the way in which the data is used after collection might still violate his/her privacy. This is especially the case when he/she is not aware of the fact that his/her personal data is being classified and clustered into profiles. Moreover, it is often unclear to a web user how some apparently trivial piece of data might result in non-trivial patterns.

## Individuality

Individuality can be described as the quality of being an individual – a human being regarded as a unique personality. Individuality is one of the strongest Western values. In most Western countries people share a core set of values maintaining that it is good to be an individual and to express oneself as that individual. Profiling through web-data mining can, however, lead to de-individualisation, which can be defined as *a tendency of judging and treating people on the basis of group characteristics instead of on their own individual characteristics and merits*.[10]

When group profiles are used as a basis for decision-making and formulating policies, or if profiles somehow become public knowledge, the individuality of people is threatened. People will be judged and treated as group members rather than individuals. This is especially harmful when profiles contain data of a sensitive nature and are, for instance, used for selections in certain allocation procedures. People could be discriminated against, or become stigmatised simply by being labelled as a member of a group or by being labelled as an individual with certain characteristics. Some criteria, like race and religion, can be inappropriate and discriminatory to use in decision-making.

An even more distressing threat is posed by the use of *non-distributive group profiles* where not every characteristic of the group is shared by every individual member. In non-distributive group profiles, personal data are framed in terms of probabilities, averages and so on, and, therefore, often made anonymous. Once the data is anonymised, it can no longer be regarded as *personal data*, which is commonly defined as data and information relating to or traceable to an individual person.[11] Vedder (1999) states that the current privacy regulations, as well as current ethical theory concerning privacy, start form too narrow a definition of personal data to capture the problems with group profiles. This also holds for the basic principles of the European Directive like the "collection limitation principle" and the "openness principle" stating that personal data should be obtained by fair means, with the knowledge or consent of the data subject, and that a subject has the right to know what personal data is being stored. The "individual participation principle" even gives a subject the right to object to the processing of his/her data. All of those principles heavily depend on the assumption that there is some kind of direct connection between a person and his/her data. In anonymous profiles, however, this direct connection has been erased. Nevertheless, the profiles are often used as if they were personal data. This sometimes makes the impact on individual persons stronger than with the use of 'real' personal data. The information cannot be traced back to individual persons. Therefore, individuals can no longer protect themselves with traditional privacy rules and people could be judged, treated, and possibly discriminated against or stigmatised based on characteristics they do not even possess. Apparently, web-data mining jeopardises different values like privacy and individuality, and other underlying values like non-discrimination, fair judgement and fair treatment.

## Arguments in defence of web-data mining

All the benefits obviously show that web-data mining is a highly valuable technique, which is being developed and applied on a large and growing scale. However, the threats to some important values tend to be rather serious, and will create tension in the web-data mining field. Unfortunately, many professionals applying web-data mining in a business context do not foresee any moral dangers in web-data mining. To gain some insight into current web-data mining practices and the attitude of web data miners to the ethical issues involved, twenty of these professionals were interviewed. These interviews combined with a literature study teach us that people prefer to focus on the advantages of web-data mining instead of discussing the possible dangers. Moreover, they revealed several different arguments to support the view that web-data mining does not really pose a threat to privacy and related values. The arguments given in their defence of almost unlimited use of data mining can be sorted into six arguments, and are enlightening. We shall discuss these arguments briefly to show that these arguments do not justify unlimited use of data mining.

[10] Cf. Vedder (1999: 275).

[11] It should be noted that not everybody agrees on this point. Evidently, group profiles cannot be considered to be personal data. But, the application of group profiles could be considered as "processing of personal data," and could, therefore, be protected by legal regulations. Although this debate on whether or not group data are equally protected as personal data has not yet been settled, it is clear that law does not fully grasp the ethical issues concerning web-data mining.

| | Possible arguments against the danger of web-data mining |
|---|---|
| 1 | *Web-data mining itself does not give rise to new ethical issues.* |
| 2 | *There are laws to protect private information, and online privacy statements guarantee privacy.* |
| 3 | *Many individuals have simply chosen to give up their privacy, and why not use this public data.* |
| 4 | *Most collected data is not of a personal nature, or is used for anonymous profiles.* |
| 5 | *Web-data mining leads to less unsolicited marketing approaches.* |
| 6 | *Personalisation leads to individualisation instead of de-individualisation.* |

If all, or most, of these arguments can be refuted, we would have to conclude that there definitely is a substantial amount of tension in the field of web-data mining.

### 1st Argument

*What is new about this? Web-data mining itself does not give rise to new ethical issues.*

This argument corresponds with the traditionalist account: all that is necessary is to take traditional moral norms and the principles on which they are based, and apply them to the new situations created by computer and information technology. Johnson (2001) observes that the traditionalist account has a serious problem. The account over-simplifies the task of computer ethics insofar as it suggests that extending old norms to new situations is a somewhat mechanical or routine process. One first has to clear up the conceptual muddles and uncertainties, which have to do with understanding what data mining is or should be and what kinds of situations it creates. Most of the possible dangers come from group profiling (in particular non-distributive group profiling), and since group profiling has been done before data mining techniques were known, the issues could be considered to be old news. It is, however, important to realise that data mining does significantly enlarge the scale on which profiling can be practised. A lot more data can be collected and analysed, and (as explained) new patterns can be found without asking for them through specific queries and hypotheses. Using these data mining techniques on web data creates another level of decision-making in which companies can use large amounts of detailed profiles based on the behaviour and characteristics of web users. So, although the ethical issues involved need not actually be, new technologies (computers, software, Internet) make it possible for individuals to mine data in new ways. We should, thus, think about what these new capabilities mean for our moral values, and our principles.

### 2nd Argument

*There are laws to protect private information. Besides, privacy policies found on many web sites guarantee privacy. So, why worry?*

This argument is not convincing, as the law is not, and never will be fully sufficient with respect to privacy problems. For instance, the fact that current privacy laws only offer protection for the misuse of identifiable personal data shows us that there is no legal protection for the misuse of anonymised data used as if it were personal data. Besides that, actions can be unethical independent of their legal status.

The growing number of online privacy policies is an example of self-regulating efforts. Such policies, however, are not found on every site. Thus, there are still a lot of sites that a person, who is concerned about his online privacy should not visit. In addition, it is not always an easy task for a web user to search for, and thoroughly read the privacy statements on every site he/she visits, or to check whether the policy has changed every time he/she visits the same site again.

Furthermore, the policies are often vague and ambiguous about certain points. Take, for instance, sentences like “will not share information with third parties.” Privacy declarations promising not to sell or give information to third parties are often not clear about who those third parties are, and, more importantly, who they are not. Stakeholders and co-operating businesses should also be regarded as third parties, but often are not. Big businesses that consist of many smaller constituent companies could exchange lots of private information if they do not regard every sub-company as a third party. Thus, in large companies information could be shared widely; a lot more widely than most people would think. Another example is the description of the use of collected data. If a goal is just marked as ‘marketing goal,’ then how can a user know what specific kind of marketing to expect? In short, privacy policies are often difficult to understand, hard to find, take a long time to read, and can be changed without notice. Clearly, regulation and self-regulation efforts do not offer sufficient protection for web users’ privacy.

**3rd Argument**

*As people can refuse to give out information about themselves, they do have some power to control their relationship with private and public organisations. Many individuals simply choose to give up their privacy. And, what can be wrong with collecting this public (content and structure) data from the web that is voluntarily given? It is there for the taking.*

For most people it depends on the situation whether or not they are willing to lose some privacy. One should, however, notice that the price of privacy on the web has become rather high. Of course, one can choose not to have an account with an Internet provider, or if one does have such an account, one can choose never to access information on the web so that there are no records of what has been viewed. These choices reduce the amount of information, which organisations on the web have about this person. However, he/she will have to give up a great deal. Not using the web does not seem to be a fair option – it could be a high price to pay for his/her privacy. Furthermore, most people who use the web are not aware of the ways in which their web data can be analysed. Is it fair to say that people choose to give up their privacy when they are not fully aware of the consequences of their actions? Can we expect people to be fully aware of those consequences? There are of course situations in which people are made aware and can then make informed choices about whether or not to give up their privacy. In the next section we shall discuss some possible solutions.

But, even though people might have some control over whether or not to give up their privacy in certain situations, it is rather difficult, if not impossible, for an individual to protect himself against the threat of de-individualisation due to the use of non-distributive group profiles. Even if an individual would be able to refuse his data to be used for profiling, that would not prevent profiling itself, and profiles could still be projected on this individual. So, he/she could still be judged and treated according to group characteristics.

In Section 3, we have already mentioned the privacy concerning web content (and structure) mining. Even if data is publicly available on the web (for instance, on someone's homepage), it can still be morally wrong to mine it, and use it for certain purposes without the publisher's knowledge and consent. Furthermore, when different bits of information are assembled, aggregated and intelligently analysed, they can invade someone's privacy, even if the parts individually taken do not (cf. Nissenbaum 1997).

Furthermore, one should be aware of poor data quality of content data on the web. As there is no overall control on web content (and structure) data, and there are no rules (except for existing laws on the freedom of speech and publication, and its boundaries), anybody can place any information he/she wants on the web. In other words, reliability of content data on the web is quite poor.

**4th Argument**

*The data collected is not of a personal nature, and most web-data mining applications result in anonymous profiles. So, why should there be a (privacy) problem? An argument often heard is: "Our software is used to identify 'crowd' behaviour of visitors to web sites. Therefore, if we don't know/care who you are, how can we be invading your privacy?"*

In the previous section, we stated that anonymous profiling could be harmful, and lead to possible discrimination and de-individualisation. This projecting of group profiles on individuals is often done in order to analyse crowd behaviour on web sites: to create clusters and categorise site visitors according to the clusters, for instance, in order to personalise the web site. It is not necessary to actually know who this person is (or to know his/her name, address, and so on) as long as this particular web user can be identified as being part of a certain cluster.

Furthermore, the assembly of certain bits of non-privacy-violating information might after mining data imply new and potentially privacy violating information. So, in the context of web content mining, certain bits of content data that are not considered to be privacy violating can be collected and assembled with other data resulting in some form of information that can be regarded as harmful. Especially, when the data is connected to external databases, for instance, demographic databases.[12]

**5th Argument**

*Data mining techniques will provide more accurate and more detailed information, which can lead to better and fairer judgements. So, web-data mining leads to less unwanted marketing approaches. Therefore, why would people complain?*

Some interviewees even foresee a possible occasional alliance between privacy protectors and data miners because of their mutual goal: less unsolicited

[12] Which is, for instance, done by Webminer. See: http://www.webminer.com/.

marketing contacts. Web-data mining could definitely be beneficial to individuals if it would lead us to less unwanted marketing approaches due to better profiling. It could, however, just as well be the other way around. Web-data mining techniques might lead to more profiles than ever before, and the probability that people will be part of a larger number of profiles could grow. In spite of greater accuracy in approaching people, the amount of (unsolicited) 'special' offers could get larger. And, although web-data mining makes it possible to provide businesses with more accurate profiles, businesses might still approach as many people as possible as the more people are approached, the higher the probability of reaching the right person(s) is. Besides, even if marketing approaches match their interests, people will still not have asked to be approached. Mulvenna et al. (2000) note that web users find online solicitations from web sites a hindrance rather than helpful. Furthermore, as Clarke (1994) states, "*at some point, selectivity in advertising crosses a boundary to become consumer manipulation*."

Data mining can lead to discrimination, and can violate the norm of equality, which contradicts a 'better and fairer judgement.' On the basis of group profiles, individuals can be included or excluded from targets groups. This can lead to poignant situations, especially, within non-distributed group profiles. Pierik (2001), for instance, shows how applying non-distributed group profiling can discriminate women in the labour market, since, on the basis of certain group profiles, many employers assume that, on average, women are less productive than men.

### 6th Argument

*Web-data mining is often used for personalisation. This leads to individualisation instead of de-individualisation. Most customers like to be recognised, and treated as a special customer. So it is not considered a violation of privacy to analyse usage interaction.*

One of the main goals of web-data mining, as used in e-commerce, is indeed personalisation: customising web sites and services to the individual wishes of each visitor. So, it does appear to promote individuality instead of threatening it. If, however, the personalisation is done by creating non-distributive group profiles, it can lead to de-individualisation and discrimination.

There is a fine line between personalisation and personal intrusion. This argument refers back to the privacy in a public domain, already mentioned. When the man behind the counter recognises someone as a regular customer and greets him/her in a friendly manner, this loss of privacy is not considered to be a violation. Knowledge about this association, however, is only in the heads of people. The data stored in a person's brain is limited to the capacity of the brain, and is not accessible to others, unless the person decides to share it with someone. Data that is digitally stored does not have to be limited by the size of the database: the database can just be upgraded. Digital data can easily be distributed and is accessible by anyone who wants to use it and is granted access to the database. Intelligent data analyses are possible, and data can easily be connected and compared to other data stored in the database. More importantly, many, if not most, of the associations between different pieces of data stored in someone's head are forgotten. Data mining not only finds unexpected associations, but the associations can also be stored for future use and reference. Note that this is not necessarily morally wrong, but there certainly is a problem. Work is required to articulate what is acceptable and what is unacceptable usage mining.

Furthermore, one should be aware of the limited data quality. The possible changes in IP addresses, and the simple fact that even when it can be established that it is the same computer (by the use of cookies), some other person might be using it, show us that recognition is definitely not flawless.

The above is not purporting to be an exhaustive list of all the possible arguments and counter-arguments in their defence of almost unlimited use of data mining. And, evidently not all arguments have been refuted: and some have just been placed in a different perspective. It has been made debatable that the dangers do pose a serious threat. One thing is clear, even though web-data mining is still in an early stage of development, there is no way back. It would not be an option to just declare that the disadvantages are so strong that web-data mining should be abolished. Nor would it be an option to choose in favour of the advantages and just ignore all the disadvantages and pretend there are no dangers. We shall have to take a middle road that leads us right through the web of tension between the advantages and the disadvantages. This is not an easy task. For, there is no marked path. We should strife to benefit from web-data mining as much as possible while causing the least possible harm.

## Possible solutions

There are means to solve some problems with respect to privacy in the ethical context of web-data mining. We can distinguish solutions at an individual and at a collective level. With solutions at an *individual*

*level*, we refer to actions an individual can take in order to protect himself/herself against possible harms. For example, using privacy enhancing technologies (PETs),[13] being cautious when providing (personal) information on-line, and checking privacy policies on web sites. The solutions at a *collective level* refer to things that could be done by society (government, businesses, or other organisations) to prevent web-data mining from causing any harm. For example, further development of PETs, publishing privacy policies, web quality seals,[14] monitoring web mining activities, legal measures, creating awareness amongst web users and web data miners, and debating the use of profiling. A mixture of technical and non-technical solutions at both the individual and the collective level is probably required to even begin solving some of the problems presented here. But, to what extent can the problems really be solved in both web-data mining categories?

*Content and structure mining*

Looking at the possible misuse of web content and structure data, and the described solutions, we have to conclude that there is little that can be done to limit the dangers. Legal measures could only provide a baseline level for better ways to handle the problems. Because mining personal data published on the web is not prevented by legal measures, monitoring web content mining activities seems to be a useful additional solution. This, however, is rather difficult to actually implement. The monitoring should be done by an impartial organisation, and it will only work if businesses co-operate and agree to give insight into their web-data mining activities. Without the miner's co-operation it would be rather difficult to monitor web-data mining activities because of the fact that it can be so easily concealed. Even with a well working monitoring 'system,' however, actual prevention of web-data mining harms still seems hard to achieve. Therefore a discussion on the social and economical circumstances that make group profiling attractive could be fruitful. The discussion on rethinking the use of profiling is, however, a difficult one. Aside from the practical difficulties to make sure that certain data or profiles will no longer be used, it is also rather complicated to determine what data or profiles are not to be used. Not to mention the unpredictable situations and patterns that can arise when applying web-data mining. Still the essence of this debate on the use of profiles is quite important. Not only should people continue to be critical about using profiles, it is also good to ask ourselves whether or not certain inclusion and exclusion mechanisms are right. Besides advising to for web users to be cautious when publishing personal information on the web, there is the collective 'obligation' to make people more aware of the dangers. This is not just about creating awareness amongst web users, but also about appealing to the moral sensitivity of web-data miners.

*Usage mining*

There are several tools that can help to protect a web user's privacy while surfing the web. A privacy-enhancing tool could enable the web user to make informed choices. It could help users to understand privacy policies, and it could provide the option of informed consent: businesses and organisations must inform customers about what information is being collected and how they will use it. However, such a tool alone would not support all basic provisions of the European Directive 95/46/EC, and there are quite a few sceptics who believe that many of these so-called privacy empowering tools are designed more to facilitate data sharing than to protect users (Scribbins 2001). In addition, seal programs and third party monitoring help to ensure that sites comply with their policies. There are some people who remain sceptical about quality seal programs because of the commercial interests. According to them, web quality seal programs have failed to achieve sufficiently high standards, or to censure companies when they violate privacy. Ultimately, it is up to individuals whether or not they trust the seals and privacy policies.

Individuals can, to some extent, protect themselves against web usage mining harms by being cautious when an online organisation asks for personal data, and by using privacy enhancing technologies. Anonymity tools reduce the amount of information revealed while browsing by, for instance, adding cookie-rejecting tools. Notice, however, that if all web users were to ban ads, a lot of online services would be lost. Many search engines, web portals, news sites, etc., obtain their profits from selling online advertising spaces. Clear privacy policies, seal programs, and privacy enhancing tools can help to win a user's trust and perhaps make a user decide not to reject the ads and cookies from the web sites he/she is visiting.

[13] The Electronic Privacy Center (EPIC) has published the "EPIC Online Guide to Practical Privacy Tools" on the web. It provides a list of privacy enhancing tools categorised by area of application. See http://www.epic.org/privacy/tools.html. For a discussion of PETs, we refer to Tavani (2000) and Tavani and Moor (2001).

[14] There are several web quality seal initiatives such as TRUSTe, the Better Business Bureau Online (BBBOnline), the Japanese Privacy Mark, and CPA Web trust. Given the nature of the World Wide Web, however, an international standard would be preferable.

As legal measures and ethical debates usually take some time before being effective, technical solutions seem to provide web users direct possibilities to actively protect themselves for possible harms. Openness about web-data mining activities is required. Consumers have to be explicitly informed that data mining techniques are being used by certain businesses, and that data about them is currently being mined in ways, which they probably have not explicitly authorised. Only then will they be able to make informed choices that will best contribute to their overall well-being. Therefore, consumers should be given three choices: (1) not having their data mined at all, (2) having their data mined to some extent (for instance, only within a certain company), and (3) having their data mined without limits (for instance, shared with third parties). A web user can use different kinds of PETs to protect himself/herself against being tracked online. However, technical measures concerning the protection, e.g., to prevent the misuse of content and structure mining, are not yet at hand. We suggest that creating a *disallow-mining* standard could be a possible solution.

*Disallowing-mining standard*

Openness about the mining of content and structure data would imply that an individual would be able to determine whether or not his/her homepage, for instance, may be mined, and, if so that he/she would be informed about it when this would happen.[15] A possible way in which to do this is by creating an automatic 'disallow-mining' standard. Evidently, asking each individual for his/her consent would be a very time consuming matter, and content miners will not do it. There might, however, be a way to automatically check for consent using techniques used by search engines. Search engines use web agents, also known as robots, to create the indexes for their databases searches. This process is called 'spidering.' Robots.txt is a file that web agents check for information on how the site is to be catalogued. It is a text file that defines what documents and/or directories are forbidden. Similarly, a 'mining.txt' file could be created which a content mining tool would check before mining the content of the site. A site owner could state in this file whether or not personal information contained within the site may be mined for certain purposes.

As the robots.txt file already provides users with the option not be indexed by search engines, it might be possible to add an extra section which states whether the site may be mined for other purposes. This solution would only work for people who have access to the document root of their web site as the robots.txt file sits in the document root of the server. In cases where people do not have root access, the site manager could, for instance, choose to alter the robots.txt file in case a user would have specific wishes.

There is also an option to direct the web agents per page. Every HTML document contains a heading section in which meta data about the document (like keywords, a description of the content, and so on) can be included. Such sections are called 'meta-tags.' Within the meta-tags of each HTML document one can specify whether or not a robot is allowed to index the page and submit it to a search engine. So, we suggest adding a 'disallow-mining' option to the HTML standard in such a way that an individual could specify his/her wishes in the meta-tags of his own homepage. Of course, such techniques would only work if they became standards (like robots.txt and meta-tags), and were widely accepted by web content and structure miners. A drawback of this suggestion is that some robots will simply ignore the meta-tags because of the fact that those tags are often misused by page owners who want to get a higher ranking in a search index. Robots may also ignore the robots.txt file or purposely load the documents that the file marks as disallowed. Therefore robot exclusion files and meta-tags should not be regarded as security measures. Misuse would still be possible, but organisations that have no intention of doing any harm would be able to respect the person's wishes without having to consult with every individual first.

None of the above-mentioned solutions addresses the protection of privacy once information is collected or the problem of group profiles (the fact that one individual has little or no influence on the entire profile). As mentioned before, even if an individual makes sure his/her data will not be used in data mining analyses, profiles derived from those analyses could still be projected to him/her. Therefore, individuals are limited in their possibilities to protect themselves. They could choose to combine forces, and make sure that personal data is systematically refused by large groups of people. A large-scale refusal of data would, however, also block all the possible advantages of group profiling. Clearly, in order to benefit from web-data mining as much as possible, while causing no, or as little as possible harm, a combined solution-package is needed.[16] Or, as Clarke (1998) states it: "*effective protection is dependent on a multi-partite, tiered*

[15] We are not taking into account, here, of the fact that a person is not always aware of all the data that is published about him on the web.

[16] Johnson (2001: 132, 135) refers to this as a 'many-pronged approach.'

*framework, in which layers of technology, organisational practices and law combine to ensure reasonable behaviour*."

## Closing remarks

To keep up with technical changes, ongoing debates about ethical issues are an essential part of this combined solution-package, and can help prevent possible future harms. As the World Wide Web becomes an increasingly important part of modern society, organisations of all kinds are engaged in efforts to use web-data mining technology for varied purposes. Many of those purposes are of a commercial nature as there is money to be made in the collection and the intelligent analysis of information about people. While most of the benefits go to the web miners, the web users are facing the dangers of web-data mining.

Although the impact of web-data mining should be a concern for every web user, there is no reason for people to panic. This technique is not yet being used to its full potential. Analysing user interaction data has been done for a long time. There are lots of different statistical tools available to analyse things like hits and page views on a web site. More intelligent usage mining tools are being developed in which data mining techniques are being used to learn more about online behaviour of customers, and to find more interesting patterns. There also are initiatives to build more intelligent and personalised search engines. Some companies focus on web content and structure. They, for instance, aim to automatically collect curriculum vitae which are published somewhere on the web, and combine and categorise them in one online database which can be used by companies in need of new employees. There are also tools available which analyse the content and structure of web sites from competing businesses in order to gain strategic insight in online management, or to spot possible brand misuse.

There is, however, no clear indication of web data being misused to an extent that people are actually hurt by it. So, it is by no means clear that companies are using unexpected and non-obvious associations, classifications, clusters, and profiles based on web data as grounds for decision-making. Indeed, we mentioned that one of the dangers lies in the hidden way in which web-data mining can be used. Companies can conceal their ultimate goals quite easily when obtaining certain bits of information. And, one could get somewhat worried when reading some of the web sites that offer web-data mining tools. Slogans like: "*The Internet is your Database*," "*Increasing the value of EVERY customer interaction*," "*Turning data chaos into profit*," and sentences like "*... to foster long-term customer relationships you need more than the analysis of log files and click stream behaviour. We believe e-retailers and content providers need to know what their customers' values are and how and where they live*." or "*There is real data on the Internet, including addresses, phone numbers, calendars, prices, job listings*" are found on web sites of companies that offer web-data mining services. Although this does indicate that the privacy of web users is threatened, there is still no reason to panic. Web-data mining is still in an early stage of development, which means that there are things that can be done to guide this technique in a socially acceptable direction. The solutions discussed in the previous section can contribute to the responsible and well-considered development and application of web-data mining. As ethical issues will grow as rapidly as the technology, ethical considerations should be an integrated and essential part of this development process instead of something at its side. It is probably not possible to develop comprehensive ethical guidelines covering every possible misuse. This is all the more reason to realise the seriousness of the dangers, and to continuously discuss these ethical issues. This is a joint responsibility of both web miners and web users.

## References

M.J.A. Berry and G.S. Linoff. *Mining the Web: Transforming Customer Data*. John Wiley & Sons, New York, 2002.

R. Clarke. 'Profiling' and Its Privacy Implications. *Privacy Law & Policy Reporter*, 1: 128, 1994.

R. Clarke. Platform for Privacy Preferences: A Critique. *Privacy Law & Policy Reporter*, 5(3): 46–48, 1998.

B. Custers. Data Mining and Group Profiling on the Internet. In A. Vedder, editor, *Ethics and the Internet*, pages 87–104. Intersentia, Antwerpen Groningen Oxford, 2001.

O. Etzioni. The World Wide Web: Quagmire or Gold Mine? *Communications of the ACM*, 39(11): 65–68, 1996.

D.R. Fordham, D.A. Riordan and M. Riordan. Business Intelligence. *Management Accounting*, 83(11): 24–29, 2002.

J.S. Fulda. Data Mining and Privacy. In R. Spinello and H. Tavani, editors, *Readings in CyberEthics*, pages 413–417. Jones and Bartlett, Sudbury MA, 2001.

D.G. Johnson. *Computer Ethics*, 3rd. edition. Prentice-Hall, Upper Saddle River New Jersey, 2001.

J.F. Johnson. Immunity from the Illegitimate Focused Attention of Others: An Explanation of our Thinking and Talking about Privacy. In A. Vedder, editor, *Ethics and the Internet*, pages 49–70. Intersentia, Antwerpen Groningen Oxford, 2001.

R. Kosala, H. Blockeel and F. Neven. An Overview of Web Mining. In J. Meij, editor, *Dealing with the Data Flood: Mining Data, Text and Multimedia*, pages 480–497. STT, Rotterdam, 2002.

S.K. Madria, S.S. Bhowmick, W.-K. Ng and E.P. Lim. Research Issues in Web-data Mining. *Lecture Notes in Computer Science*, 1676: 303–312, 1999.

B. Mobasher, R. Cooley and J. Srivastava. Automatic Personalization Based on Web Usage Mining. *Communications of the ACM*, 43(8): 142–151, 2000.

M.D. Mulvenna, S.S. Anand and A.G. Büchner. Personalization on the Net Using Web Mining. *Communications of the ACM*, 43(8): 123–125, 2000.

H. Nissenbaum. Toward an Approach to Privacy in Public: Challenges of Information Technology. *Ethics & Behavior*, 7(3): 207–220, 1997.

E. Philips, D. Vriens and G. Kienhorst, editors. *Business Intelligence*. Kluwer, Deventer, 1999.

R. Pierik. Group Profiles, Equality, and the Power of Numbers. In A. Vedder, editor, *Ethics and the Internet*, pages 105–123. Intersentia, Antwerpen Groningen Oxford, 2001.

K. Scribbins. *Privacy@net, An International Comparative Study of Consumer Privacy on the Internet*. Consumers International (http://www.consumersinternational.org/news/pressreleases/fprivreport.pdf), 2001.

J. Srivastava, R. Cooley, M. Deshpande and P.N. Tan. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *ACM SIGKDD*, 1(2): 12–23, 2000.

H.T. Tavani. Informational Privacy, Data Mining, and the Internet.' *Ethics and Information Technology*, 1: 137–145, 1999a.

H.T. Tavani. KDD, Data Mining, and the Challenge to Normative Privacy. *Ethics and Information Technology*, 1: 265–273, 1999a.

H.T. Tavani. Privacy Enhancing Technologies as a Panacea for Online Privacy Concerns: Some Ethical Considerations. *Ethics and Information Technology*, 9: 26–36, 2000.

H.T. Tavani and J. Moor. Privacy Protection, Control of Information, and Privacy-Enhancing Technologies. *Computers and Society*, 31: 6–11, 2001.

A. Vedder. KDD: The Challenge to Individualism. *Ethics and Information Technology*, 1: 275–281, 1999.

A. Vedder. Privacy and Confidentiality. Medical Data, New Information Technologies, and the Need for Normative Principles Other than Privacy Rules. *Law and Medicine*, 3: 441–459, 2000.

A. Vedder. KDD, Privacy, Individuality, and Fairness. In R. Spinello and H. Tavani, editors, *Readings in CyberEthics*, pages 404–412. Jones and Bartlett, Sudbury MA, 2001.