Mining for data and personal privacy: reflections on an impasse

July 2004
Ioannis-Evangelos Papasliotis

ABSTRACT: Data mining, also known as ‘knowledge discovery in databases’, is a technique that enables the user to find answers to questions he/she never thought of asking. It already has had applications in medical research, commerce and national security, amongst others. This is a challenge for both law and ethics, especially in the context of issues of privacy. To what extent are the current laws adequate to deal with the new technology? Do we need a new set of laws or principles? Although this article does not give direct answer to those questions, it tries to highlight the various problem areas and to suggest possible approaches towards a solution.

i) Introduction

Information Technology (Personal Computers, World Wide Web) will continue to overwhelm us with enormous amounts of widely distributed, interconnected and dynamic hypertext information. The abundant information held on various computers both on the Internet and on private networks is not stored in any systematically structured way. This poses a great challenge to those searching for high quality information (requires for purposes of commerce, scientific research, law enforcement, and security) and seeking to uncover knowledge buried among billions of gigabytes.

Today, technology offers powerful new tools for gleaning otherwise concealed knowledge from databases - the tools of data mining, also known as ‘knowledge discovery in databases’. In this paper I use my term mining for data as having a wider scope than data mining. It involves data mining but also includes the development and use of information technology in order to collect new data (including its automatic or continuous collection).

Increasingly, organizations are generating vast amounts of data as a result of running a variety of information systems. This data is normally used to record transactions and for status reporting purposes. What data mining does is to employ elements of statistical analysis, artificial intelligence, machine learning and advanced modelling techniques to predict, for example, customer behaviour patterns, or security threats, using large data warehouses and other forms of data resources. Typically, a data mining project draws on the information in more than one data base.

1 I am grateful to Ms Sivaramjani Thambisetty and Dr. Stephen Hailes for their helpful comments and suggestions on a draft of this article.
2 LLB (Hons), Maitrise en Droit, MSc in Computer Science, Advocate (Athens Bar)
3 Enables linking from one web page to another
4 Unit for measuring the amount of data held (1 gigabyte = 1024 megabytes = 181775 A4 pages approx.)
An example is useful. In 2000, the Government of Iceland gave a private company a 12-year licence to create and manage an electronic database of the country’s medical records, previously decentralised among a number of clinics. The company, DeCode, is also collecting genetic information on individual Icelanders. What is more important, DeCode’s licence from the Government allows it to cross-reference information in the medical-records database with information in the genetic database. Three years ago the company expressed the hope to be able to identify the function of genes or genetic variations that contribute to the onset of 35 illnesses, including Alzheimer’s and lung cancer. The company’s objective is to discover and market new drugs and diagnostic tests for the illnesses. On the other hand, the developments sketched above have aroused deep public concern about their possible impact on personal privacy and about the adequacy of statutory protections. In 2003 the Supreme Court of Iceland handed down a landmark decision. It was decided that a daughter might have an interest in preventing the transfer of health data about her deceased father into the medical-records database, “as information could be inferred from such data relating to the hereditary characteristics of the father which might also apply to herself”. The legal and ethical issues involved in the ongoing developments continue to be debated in Iceland and internationally (including in countries such as Estonia and the UK).

This article is primarily about the implications of the tools or techniques of data mining for the protection of personal privacy. It attempts to outline answers to two questions: Does mining for data create problems which are not adequately addressed by current laws and regulations for the protection of personal privacy? Are other approaches necessary in the light of new developing technologies such as data mining?

**ii) Mining for data, privacy and data protection rights**

The law on protection of privacy applies to the disclosure of information generated by mining for data methods. Because data mining tools can be used even from a remote location, the question as to what is the ambit of an individual’s private life becomes more relevant than ever. Based on Campbell v MGN Ltd, it could be argued that the test to be applied is whether in respect of the information discovered by mining for data the person whose data is being mined has a reasonable expectation of privacy.

When mining for data, there can be instances where data is collected and analysed without the data subject’s consent (which in any case would be very difficult to obtain in view of the number of subjects involved). Nevertheless, when a mining for data project leads to disclosure of new information about an individual, the use of that information could constitute an invasion of privacy. Furthermore, when using mining for data tools, personal space can be violated, this could also amount to an infringement of privacy. Lord Mustill said in R (on the application of BBC) v Broadcasting Standards Commission [2001] QB 885 “an infringement of privacy is an affront to the personality, which is damaged both by the violation and by the demonstration that the personal space is not inviolate”.

The “weight” of the law varies according to whether information is public or private. With mining for data it is not always clear in which area—public or private—the research takes

---

5 http://www.cio.com/archive/071501/dilemma.html

place. Especially in an environment like the Internet, the line that separates the two can be very thin.

When providing information to a doctor, a merchant, or a bank, individuals expect that those professionals or companies will collect the information they need to deliver a service, and will use it for that sole purpose. However, the ‘data miner’ may be enabled to come up with answers to questions that he (and the doctor) has never thought of asking, or information that he has never thought of looking for. Therefore information may not be used solely for the purpose for which it was collected. Individuals expect that they have the right to object to any further use of their personal information. Implementation of data protection / fair information practices (e.g. in the US) is key to preserving this personal autonomy by ensuring that an individual’s privacy interests in his or her personal information are protected.

Privacy today also refers to protection from government surveillance. New technologies such as data mining that enhance the ability of law enforcement agencies to compile an array of information about an individual test the limits of existing statutes and constitutions.

Interpretation of the scope of the right of privacy will clearly impact on how the process of mining for data is viewed. An absolutist view of privacy is based on the notion that an individual has the right to control the use of his information in all circumstances\(^7\). One solution is to require the anonymisation of personal data\(^8\). This has the effect of providing total privacy protection for all data subjects, regardless of any perception they may have of their vulnerability. However, this would rule out any legitimate Knowledge Discovery (KD) applications dependent on identifiable data subjects, and prevent many data mining activities altogether\(^9\). Such an abolitionist policy would be inappropriate for medical research purposes, for example. Furthermore, it is hard to anonymise without distorting data (except in trivial cases).

In 1980, the Organisation of Economic Cooperation and Development (OECD) produced a set of guidelines\(^10\) for the protection of personal data. These guidelines acknowledge and support the individual’s prerogative to participate in the control of his or her personal information. Such participation enables the individual to pursue the protection of his/her privacy as he/she perceives it. The guidelines are designed to enable an individual to refrain from providing personal data for any purpose he/she considers inappropriate.

One of the guidelines states that data cannot be used for any purpose other than that held out when the data was originally obtained from the individual\(^11\). Due to the nature of mining for data techniques, however, this objective might now be difficult to achieve. Another guideline states that the reason for collecting personal data should be made clear to the individual prior to its collection. This also seems to be incompatible with the nature of some data mining tools.

\(^7\) (This is so even if the individual subject is anonymised) – on anonymisation of personal data see Anonymising Personal Data – Ian Walden – International Journal of Law and IT June 2002

\(^8\) Privacy and Dataveillance, and Organisational Strategy – Clarke R. 1997 – Information systems audit and control association conference, Promaco Conventions Pty Ltd., Western Australia

\(^9\) On the impact of knowledge discovery and data mining – Kirsten Wahlstrom and John F. Riddick, 2nd Australian Institute of Computer Ethics Conference 2000

\(^10\) www.oecd.org/dstl/sti/it/secur/prod/priv-en.htm

\(^11\) This is also reflected in the UK’s Data Protection Act
Similarly, Article 6 of the EU General Protection Directive (95/46/EC) lays it down that personal data must be collected for specified, explicit and legitimate purposes and not be further processed in a way incompatible with those purposes. Since data mining is based on the extraction of unknown patterns from a database, the data miner does not know, and cannot know, at the outset, what personal data will be of value, and what relationships will emerge. The requirement of the Directive may effectively rule out a data mining exercise.

KD uses data collected for one purpose for another purpose. The specific purpose of a KD application is largely unknown until it has successfully revealed some previously unknown information. The information revealed by KD may be considered inappropriate (in terms of type and quantity) by the individual data subjects whom it describes. These contraventions diminish the individual’s capacity to participate in the control of the use made of his/her personal data, and thus threaten to violate his/her sense of privacy.

The European Parliament and Council’s General Directive of October 24th 1995 governing the processing of personal information, in chapter 1, article 1 provides that “Member states shall protect the fundamental rights and freedoms of natural persons, and in particular their right to privacy, with respect to the processing of personal data”. “Processing of personal data” includes “collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction” [Chapter 1, article 2(b)]. Among other things, the Directive limits the purposes for which personal information on an individual may be collected to “specified, explicit and legitimate purposes” and guarantees a right of access of the individual to such data.

Several features of the Directive pose serious problems for data mining. These features are:

1. Without the consent of the subject, data processing is legitimate only if needed for the performance of a contract to which the data subject is a party, for compliance with a legal obligation, or for the performance of a task that is carried out in the public interest.
2. The subject has a right to access the relevant data.
3. The subject has the right to demand that inaccurate data be corrected or removed.

A problem with data mining is that of data accuracy. The OECD’s guidelines require personal data to be precise, complete and current in order to protect people from the harmful repercussions associated with poor quality of data. This becomes all the more relevant when a data mining application/tool reveals information with detrimental repercussions for a data subject, especially as information is customarily taken as accurate regardless of whether or not it is in fact so.\(^{12}\)

Suppose an individual suffers harm simply because incorrect data has been used. Should the injured individual have a right of redress or a right to compensation? If so, who should adjudicate and assess the damage done? Perhaps a government agency or an independent body appointed specifically to deal with such issues is needed.

Most data protection laws (including the EU’s Data Protection Directive) have understandably been drafted with little, if any, specific consideration being given to data mining. For example, data protection laws prohibit collection of data for future use. However, data mining tools can help collect material that could be of future use without this being immediately apparent. Hence, the regulation of data mining operations tends to be incidental or accidental.

\(^{12}\) Privacy and the limits of the law – Gavison R. – Computers, ethics and social values 1984 p332-351
In fact, considerable uncertainty surrounds the exact way in which the laws apply to data mining. This is part of a larger problem, namely that data protection laws still tend to lack sufficiently detailed rules that take account of the quality and architecture of information systems.

In Innovation(Mail Order) Limited v Data Protection Registrar, the UK Data Protection Tribunal stated that “fair obtaining” means that at the time that information is collected the data user needs to inform the data subject of certain matters that will enable the subject to decide whether to provide the information or not. In particular, this includes information about the intended use(s) for the data, unless such use(s) can be considered obvious. It is often the case that results obtained by data mining tools raise the question as to how the results can be used.

It is worth noting, that the US Privacy Act of 1974 provides little restraint on data mining of commercial databases. The Act does not apply to private sector databases. The Act was passed in response to concerns about the creation and dissemination of large government databanks of personal information. This means that its protections apply only where the government is creating a ‘system of records’. But data mining currently being developed does not involve the creation of government databases. Data mining can be conducted in such a way that the data never leaves private hands and so the protection of the act is avoided. US government agencies can secure various scans of data held by private corporations without bringing that data into a centralised government database.

**iii) Concluding observations, including suggested remedies**

Subject to some important but understandable exceptions, the UK’s Data Protection Act -to take one example- gives the data subject the power to control the use by others of his personal data. It is based on the premise that the data subject can refuse to disclose personal information for use for a purpose he considers inappropriate (e.g. he/she can prevent processing likely to cause him/her damage or distress –article 10 of DPA).

As we have seen, however, personal information he/she has allowed to be made available for a specific purpose may subsequently, via data mining, be used (together with personal information from a large number of other subjects) for a purpose or purposes for which he/she was not aware and of which the data miner also may not have been aware. Moreover, it may not be practical for the data miner to obtain consent because of the large number of subjects involved.

A different public policy approach may therefore have to be adopted. Basically, that approach might be to establish arrangements which can control or regulate data mining, for example (i) by requiring data mining projects to be licensed, and (ii) by monitoring the use of new information generated by licensed data mining projects.

---

14 29 September 1993 (case DA/92 31/49/1)
16 I believe that precedents for this approach may be found in areas such as experiments on animals.
If it is considered necessary to have a regulatory body, it would have to be established by
government, and its members appointed by it. Membership would no doubt be drawn from
various constituencies such as law, computer science, sociologists, ethicists, etc. After a
period of years it might be possible for the body to introduce block exemptions for specific
categories of projects. It may also be desirable that some types of data mining projects or
programmes be allowed only after a privacy assessment has been approved by an appropriate
body.

As Baroness Mary Warnock\textsuperscript{17} has said, the central question is whether the potential benefits
of a practice –in our case data mining- are outweighed by the possible harms. This calls for a
case by case approach. In my view, a regulatory body is likely to be more effective than a
court of law (unless a specialised court is set up for the purpose, as was the case in
competition law with the establishment of the Restrictive Practices Court\textsuperscript{18} in 1956
legislation)

The importance of privacy issues connected with the use of data mining techniques is growing
rapidly as public concerns are intensifying as regards terrorist threats to national security and
also the prevention of crime.

Unless there are effective privacy safeguards, data mining has the potential to be used for
Spying on members of the public without being subject to the judicial or procedural
constraints which limit the extent to which traditional surveillance techniques are permitted
to infringe privacy. Existing arrangements in relation to national security do not pay enough
attention to the protection of personal privacy.

As Mr Ian Brown(Chairman of Privacy International-human rights group) has recently\textsuperscript{19}
stated there is a “growing surveillance culture”. A further illustration of this culture is the
recent example of a pilot scheme in England (UK) which uses an automatic number plate
recognition (ANPR) system. The present Data Protection laws may not be able to face the
challenge of the new technologies.

Another form of mining for data is Radio Frequency Identification (RFID). RFID tags are
tiny chips that can be embedded in anything from a Coca Cola can to a shirt label. The chips
act as transponders. When an RFID ‘reader’ emits a radio signal, the (associated) tags in the
vicinity respond by transmitting the detailed sorted data they have collected. This could relate
to the shopping habits, behaviour and decisions of individuals. Thus the use of RFID raises
data protection issues.

One possible solution concerning the use of RFID tags at the consumer level would be to treat
RFID tag chips as ‘Computers’, in which case the UK Computer Misuse Act (part 1
unauthorised access) and Data Protection Act would apply. The rules of data protection would
have to be complied with, including those involving the informed consent of the customers

Some experts have proposed that the RFID tag problem could be solved by ‘killing’ the tags
at the point of sale, rendering them inoperable. However, the proposed EU Intellectual

\textsuperscript{17} Baroness Mary Warnock, \textit{Philosophy and Ethics} in Cookson, C, Nowak, G and Thierbach – Genetic
Engineering 1993 Munich- European Patent Office

\textsuperscript{18} The court was terminated a few years ago.

\textsuperscript{19} The Guardian newspaper, \textit{They have your number}, 29\textsuperscript{th} July 2004
http://www.guardian.co.uk/humanrights/story/0,7369,1271120,00.html
Property (IP) Enforcement Directive includes a measure that would make it illegal for European citizens to deactivate the chips in RFID tags, on the ground that the owner of the tag has an intellectual property right in the chip. De-activating the tag could arguably be treated as an infringement of that right.

RFID does not raise only data protection issues. In certain circumstances its use could also break European human rights law. For example, section 8 of the Human Rights Act states that every individual has a right to privacy in his or her private life, home and correspondence. As Justice Louis Brandeis stated, in Olmstead v US (1928) “The right to be left alone – the most comprehensive of rights, and the right most valued by a free people”.

References

- Clarke R. -Privacy and Dataveillance, and Organisational Strategy –1997 – Information systems audit and control association conference, Promaco Conventions Pty Ltd.,Western Australia
- Gavison R - Privacy and the limits of the law. – Computers, ethics and social values - 1984
- Ian Walden – Anonymising Personal Data –International Journal of Law and IT June 2002
- Kirsten Wahstrom and John F. Riddick- On the impact of knowledge discovery and data mining, 2nd Australian Institute of Computer Ethics Conference 2000
- Privacy’s Gap: The Largely Non-Existent Legal Framewrok for Government Mining of Commercial Data – Center for democracy and technology – 28 May 2003
- Mark Oliver- They have your number –Guardian Unlimited -29 July 2004 [http://www.guardian.co.uk/humanrights/story/0,7369,1271120,00.html]
- OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data-[http://www.oecd.org/document/18/0,2340,en_2649_34255_1815186_1_1_1_1,00.html]