On the Impact of Knowledge Discovery and Data Mining

Kirsten Wahlstrom

School of Computer and Information Science
University of South Australia
GPO Box 2471, Adelaide 5001, South Australia

kirsten.wahlstrom@unisa.edu.au

John F. Roddick

School of Informatics and Engineering
Flinders University of South Australia
PO Box 2100, Adelaide 5001, South Australia

roddick@cs.flinders.edu.au

Knowledge Discovery and Data Mining are powerful automated data analysis tools and they are predicted to become the most frequently used analytical tools in the near future. The rapid dissemination of these technologies calls for an urgent examination of their social impact. This paper identifies social issues arising from Knowledge Discovery (KD) and Data Mining (DM). An overview of these technologies is presented, followed by a detailed discussion of each issue. The paper's intention is to primarily illustrate the cultural context of each issue and, secondly, to describe the impact of KD and DM in each case. Existing solutions specific to each issue are identified and examined for feasibility and effectiveness, and a solution that provides a suitably contextually sensitive means for gathering and analysing sensitive data is proposed and briefly outlined. The paper concludes with a discussion of topics for further consideration.

1 Introduction

To paraphrase Terry Winograd (1992) we bring to our communities a tacit comprehension of right and wrong that makes social responsibility an intrinsic part of our culture. Our ethics are the moral principles we utilise to assert social responsibility and to perpetuate safe and just societies.

An important lesson of the Industrial Revolution was that the introduction of new technologies can have a profound effect on our ethical principles. In today's Information Revolution we strive to adapt our ethics to concepts as diverse as gender issues in cyberspace (Herring 1994) and the appropriate use of Management Information Systems (Wagner 1994). The recent emergence of very large databases, and their associated data analysis tools, presents us with yet another set of ethical challenges to consider.

The terms "Knowledge Discovery" and "Data Mining" are used to describe the 'non-trivial extraction of implicit, previously unknown and potentially useful information from data' (Cavoukian 1998). These tools use knowledge-based machine learning and statistics over very large databases to reveal interesting "nuggets" of information and have recently undergone a rapid increase in use.

The information nuggets are patterns or relationships in the data and those that satisfy a set of user-defined criteria for certainty and interest are used to corroborate 'proactive knowledge-driven business decisions' (Cavoukian 1998). These technologies are powerful tools for data analysis and trend prediction and are expected to be amongst the most significant tools of this type in the foreseeable future. Consequently, an examination of the relevant social issues is increasingly appropriate, if not urgent.

Although such issues have been identified as pertinent to KD and DM (Rainsford and Roddick 1999) with the exception of privacy (Brankovic and Estivill-Castro 1999, Fulda 1999), there has been little interest in examining them. This paper contributes to a foundation for future research in this area.

The following section of the paper presents the social issues; specifically privacy, database security, data accuracy and the research dilemma. The former three issues are directly related to Knowledge Discovery and the latter has a broader, academic relevance. This is followed by the presentation of a solution which offers contextual sensitivity and more informed organisational decision making. Finally, topics for further consideration are presented.

2 The issues

It should be clear that Knowledge Discovery itself is not socially problematic. Ethical challenges arise when it is executed over data of a personal nature. One concept this paper hopes to establish is the relevance of context when assessing cultural and personal impact. For example, the mining of manufacturing data is unlikely to lead to any consequences of a personally objectionable nature.

However, mining clickstreams of data obtained from oblivious web users instigates a variety of social dilemmas. Perhaps the most significant of these is the invasion of privacy.

2.1 Privacy

Complete privacy is not an inherent part of any society (Gavison 1984). This is because participation necessitates communication, which renders absolute privacy unattainable. Within the context of a society's communication practices, individual society members develop independent and unique perceptions of their own privacy (Rachels 1975). Hence, privacy exists within a society only because it exists as a perception of its members. This perception is crucial as it partly determines whether, and to what extent, a person's privacy has been violated.

An individual can maintain their sense of privacy by limiting their accessibility to others (Gavison 1984). This can be achieved by restricting the availability of their personal information. If a person considers the type and amount of information known about them to be inappropriate, then their perceived privacy is at risk. Thus, privacy can be violated when information concerning an individual is obtained, used, or disseminated, especially if this occurs without their knowledge or consent.

In 1980, the Organisation of Economic Cooperation and Development (OECD) produced a set of guidelines for the protection of personal data. These guidelines acknowledge and support the individual's prerogative to participate in the control of their personal information. Such participation enables individuals to pursue the protection of their perceived privacy.

One of the guidelines states that data cannot be used for any purpose other than that held out when the data were originally obtained from the individual. Another guideline states that the reason for collecting personal data should be made clear to the individual prior to collecting it. These guidelines enable an individual to refrain from providing personal data for any purpose they decide is inappropriate. KD has the potential to violate these principles.

Firstly, KD is commonly used with large amounts of historical data and thus uses data collected for one purpose for another purpose. Secondly, the specific purpose of a KD application is largely unknown until it has successfully revealed some previously unknown information. That is, its purpose is intrinsically related to the information it discovers. Finally, the information revealed during 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 their data and thus threaten to violate their sense of privacy.

It should be noted that the DM paradigm used in a KD application impacts on the potential for privacy violation. The three mining paradigms examined by Roddick and Lees (2001) differ in the manner in which the results of the mining process are handled as follows:

- the results of DM routines are fed directly back to the user;

- the results of DM are embedded within a process that interprets the results as being merely hints towards further properly structured investigation into the reasons behind the rules; and

- there is a KD process that accepts an hypothesis and attempts to refine it through the modification of the hypothesis as a result of DM.

The latter two techniques are ones in which the results of the complete KD process (including the subsequent research) have less potential for compromising an individual's perceived privacy than the former technique. This is due to the increased significance of the DM results in the former technique. If the DM had been applied over incorrect or sensitive data, the risk of an inappropriately modified hypothesis is increased.

As privacy is contextual and a product of individual perception, an infallible and universal solution to this dilemma is not possible. What is an acceptable solution for one person can be insufficient or unacceptable for another. However, there are measures that can be undertaken to enhance privacy protection.

One solution is the anonymisation of personal data (Clarke 1997). 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 render obsolete any legitimate KD applications dependent on identifiable data subjects, and prevent many mining activities altogether. Thus, an abolitionist policy is, we contend, inappropriate.

One proposed compromise with sufficient contextual sensitivity is the empowerment of individuals to dictate the amount and type of personal data they consider appropriate for an organisation to use for KD. Cavoukian (1998) suggests a two-option approach to this empowerment, where the organisation capturing the data provides the individual with the opportunity to permit or deny the use of their data for other purposes. These other purposes (for example, KD or data trade) could be identified and described, enabling informed decision-making on the part of the individual.

Current practices fall well short of this ideal. Commonly, an individual must adopt a proactive and assertive attitude in order to maintain their privacy, usually having to initiate communication with the holders of their data to apply any restrictions they consider appropriate. Furthermore, many individuals are unaware of the extent of the personal data stored by governments and private corporations. Such data can describe sensitive information such as credit rating, race, medical history, and so on. It is usually only when things go wrong that individuals exercise their rights to obtain this data and seek to eliminate or correct it.

2.2 Database security

In an ethical sense, database security is related to privacy. This is because database security inhibits the unauthorised dissemination of personal data thus further enhancing, albeit indirectly, an individual's capacity to regulate access to their data.

In terms of database security, two forms of KD operation need to be considered:

- those operating as authorised applications by an individual or organisation that holds and has full access to the data; and

- those operating as unauthorised applications by an individual or organisation that has access to the data only insomuch as has been permitted for other allowable purposes.

Note that the individual need not be external to the organisation that holds the data for the second operation to occur.

Conventional database security protects data via user authorisation techniques (O'Leary 1991) making no distinction between the degrees of sensitivity present in the database (Mills 1997). A more sophisticated model, Multi Level Security (MLS) extends conventional security measures by classifying data according to its confidentiality (Elmasri and Navathe 1994). The data in an MLS database is typically sorted into four security levels, with users permitted access only to their authorised level. This increases the protection of data from misuse by both authorised and unauthorised users.

Encryption is another popular database security technique. This approach has been offered commercially in updates to Oracle8i (Hammond 2000) which encrypt individual data items. Encryption is considered particularly applicable to databases accessible via the web, because of the increased data exposure and vulnerability inherent to this means of data access. It is not an infallible data security technique and a brief search of the literature will reveal numerous instances where data encryption has been insufficient.

Finally, auditing (Austin 1999) is used to record a database's transactions and who executed them. It is commonly used as a tool for enhancing database performance (Johnson 1999), as it identifies the busiest tables, privileges, etc but it can also be used to identify attempted database intrusions. However, auditing provides historical data for analysis rather than a means of detecting database intrusion as it occurs and its value as a preventative security measure is minimal.

Miller (1991) showed how users executing specific queries at their authorised security level in an MLS database could easily infer more sensitive information and later Thuraisingham (1997) discussed the possibility of this occurring during DM.

There exists a set of precautions that can enhance existing database security to improve the protection of personal data from unauthorised KD applications. Firstly, restricting mining applications to one security level in an MLS database can inhibit inference from less sensitive data to more sensitive data (Lin, Hinke, Marks and Thuraisingham 1996). Secondly, the introduction of noise to the data serves to corrupt the results of any symbolic learning techniques present in a DM tool (Miller 1991, O'Leary 1991). Finally, the introduction of instability to the data (O'Leary 1991) renders it unsuitable for mining by hindering the extraction of meaningful information. Note that these measures are reversible for authorised applications.

The combination of these precautions with conventional database security models only serves to discourage unauthorised KD by rendering it a complex and cost intensive exercise. Means of ensuring infallible protection of data from malicious applications are yet to emerge.

2.3 Data accuracy

The paper previously noted the OECD's guidelines (1980) for protecting personal data. One of the guidelines requires personal data to be precise, complete, and current in order to protect people from the harmful repercussions associated with poor data quality. This becomes all the more relevant when a KD application reveals information with detrimental repercussions for a data subject, especially as information is customarily taken as infallible regardless of whether it is in fact true or false (Gavison 1984).

Knowledge Discovery applications involve vast amounts of data, which are likely to have originated from many diverse, possibly external, sources. This means the initial quality of the data cannot be assured and it might be noisy, obsolete, inaccurate, or incomplete (Cavoukian 1998). Moreover, although data pre-processing (or cleaning) is undertaken before a mining application to improve data quality, people conduct transactions in a sporadic and largely unpredictable manner, which causes personal data to expire rapidly. In some cases, what is accurate data in one point in time is inaccurate shortly after that. When a KD application is executed over expired data inaccurate patterns are more likely to emerge, which can lead to negative consequences for an individual.

The data quality issue is difficult to resolve. The inaccurate data are undetected by the individual until he or she experiences some associated repercussion, such as denial of credit, or the withholding of a payment. It is also usually undetected by the organisation holding the data, as it lacks the contextual knowledge necessary for the exposure of inaccuracies.

The adoption of data quality management strategies by the organisation, coupled with the expedient correction of any inaccuracies reported by individuals and frequent data cleansing may go some way to solving the dilemma. Other solutions are apparent (for example, data matching) but they all have unsatisfactory implications for privacy protection.

2.4 The research dilemma

When conducting research, well documented and clearly defined ethical guidelines should be followed by those interested in the integrity of their work. One example of such guidelines is the Australian 'Joint NHMRC/AVCC Statement and Guidelines on Research Practice' (1997). This document presents ethical strategies for data storage and retention, authorship, publication, supervision of students and research trainees, disclosure of potential conflicts of interest, and research misconduct. It is intended to provide a national basis for Australian research institutions' ethics policies. Accordingly, Australian University research policies (see for example the University of South Australia's Council Policies - Research (1997)) cover human research procedures, genetic manipulation and recombinant DNA research, animal experimentation, radiation safety, biohazards, and responsible research practice (as outlined by the Office of National Health and Medical Research Council (1997)). However, while these documents thoroughly consider the practical issues immediately associated with research, they both fail to consider the sociological influence of research results generally, and of emerging information technologies in particular.

The term 'disruptive technologies' has been coined by business researchers (Christensen 1997) to refer to technologies that have an agitating effect on a market and the organisations competing within it. How is it this term has not been adapted by sociologists and computer scientists to describe technologies that disrupt our cultures? Examples of such technologies abound. A perusal of any computer ethics text (for example Baase (1997)) will reveal discourse regarding software accidents, risks, crime, hacking, and so on. However, these discussions typically cite documented incidents and fail to identify the possibility of evaluating a technology's potential for social disruption.

In 1965, Einstein lamented the impact of his research on humanity:

The release of atomic power has changed everything except our way of thinking...the solution to this problem lies in the heart of mankind. If only I had known, I should have become a watchmaker.

Later, Weizenbaum (1972) discussed the responsibility of the computer scientist to their society. He called for the recognition of social responsibility by computer scientists and considered the computer scientist ultimately responsible for the impact of his/her research. However, we argue that while the computer scientist's thorough understanding of their work provides them with the capacity to argue on the work's behalf, it does not provide the capacity to estimate the work's social impact. Only a combination of suitably qualified individuals, historical information, and a comprehensive understanding of the emerging technology can adequately provide such an estimate.

In the UK, the problem is being addressed by the Foundation for Information Policy Research (1999), an independent organisation examining the interaction between information technology and society:

Our goal is to identify technical developments with significant social impact, commission research into public policy alternatives, and promote public understanding and dialogue between technologists and policy-makers in the UK and Europe.

It combines information technology researchers with people interested in social impacts, and uses a strong media presence to disseminate its arguments and educate the public. The involvement of diversely qualified people, the consideration given to each emerging technology, and the active dissemination policy, clearly provides an effective method of assessing the cultural impact of technological developments.

3 User-defined sensitivity factors

Providing individuals with more control of their data is a solution with a strong contextual sensitivity that relieves both the privacy and data accuracy issues and offers a positive impact on the research dilemma. This is because the individual, with their unique perception of privacy and knowledge of what data values accurately describe their transactions and selves, is most qualified to determine the use of their data.

One means by which the use of data can be regulated, either by individuals or a set of advocates operating on behalf of the public, is by the introduction of user-defined sensitivity factors to data. This method proposes that when providing data, individuals would be provided with the means to dictate the sensitivity of each individual data item. For example, an individual (or the advocacy team) may consider that the attribute of sexual preference is of a highly sensitive nature and rates this data item accordingly. However, age may be rated lower and car ownership lower still.

The sensitivity of a proposition, P_n, over a set of attributes, A_{1-m}, can then be derived in two ways:

- specifically, from the sensitivity of the attributes; and,

- non-specifically, from the sensitivity of the propositions.

For example, the sensitivity of the proposition P_3 in the following rule

$$P_1(A_1, A_2) \wedge P_2(A_3) \rightarrow P_3(A_4)$$

can be measured specifically as a function of the sensitivities of the attributes A_1, A_2, A_3 and A_4, or non-specifically as a function of the sensitivities of propositions P_1 and P_2.

When propositions are executed over the sensitised data, results are presented with an indication their sensitivity. This enables the identification of insensitive propositions that can then be utilised freely in automated data analysis. Thus, organisations are better informed and are able to use their analyses with appropriate consideration for the individuals who provided the initial data.

Hence, data sensitivity factors provide a highly context sensitive means of collecting and analysing data, offering more pertinent privacy protection for individuals. They empower both individuals and organisations with the capacity to minimise the effects of a potentially detrimental technological application. This naturally extends to the minimisation of any negative cultural impact caused by emergent technologies that exploit personal data.

4   Further investigation

The primary consideration of future research should be at least maintenance, and preferably enhancement, of existing ethical standards. Solutions reconciling any social issues must not only be applicable to the ever-changing technological environment, but also flexible with regard to specific contexts and disputes. Also, many ethical issues overlap and have effects on each other; similarities and differences should be identified and exploited to derive solutions. Furthermore, we need to be able to identify ethical dilemmas as they arise and derive solutions in a timely, contextually sensitive and preferably concurrent manner.

In a practical, academic sense the introduction of user-defined sensitivity factors will be prototyped and an empirical evaluation of the prototype is planned. These efforts will serve to further reveal the suitability of this approach to privacy protection.

An interesting social dilemma not considered here is that contemporary Luddite philosophy (Baase 1997) in fact disempowers humanity, which is ironically its philosophical antithesis. Information technology is now a deeply rooted part of our culture, with new research constantly emerging to alter our ways of life. Neo-Luddites adopt an active abolitionist policy, which inadvertently encourages a fear of technology. This fear undermines the significance of people in our existing technologically rich culture. A philosophy that encourages people to master technology is far more empowering and offers support rather than dissuasion for people.

We exist in an environment of rapid change in which technology has an ever-increasing social relevance. The challenge now is to adapt our approaches to the application of new technologies, enabling us to use the tools technology provides wisely and with consideration for our society, its members, and its future.

5   References

AUSTIN, D. (1999): *Using Oracle8.* MacMillan Computer Publishing.

BAASE, S. (1997): *A gift of fire: social, legal, and ethical issues in computing.* Prentice-Hall.

BRANKOVIC, L. and ESTIVILL-CASTRO, V. (1999): Privacy Issues in Knowledge Discovery and Data Mining. Online, accessed 14 August 2000. URL: http://www.aice.swin.edu.au/events/AICEC99/webabstractsindex.html

CAVOUKIAN, A. (1998): Data mining: staking a claim on your privacy. Online, accessed 27 April 2000. URL: http://www.ipc.on.ca/Web_site.ups/Intro/Frames.htm

CHRISTENSEN, C.M. (1997): *The innovator's dilemma: when new technologies cause great firms to fail.* Harvard Business School Press.

CLARKE, R. (1997): Privacy and Dataveillance, and Organisational Strategy. *Proceedings of the Region 8 EDPAC'96 Information Systems Audit and Control Association Conference.* Perth, AUS, keynote speech, Promaco Conventions Pty Ltd.

EINSTEIN, A. (1965). Online, accessed 21 August 2000, URL: http://fys.ku.dk/~raben/einstein/index.html

ELMASRI, R. and NAVATHE, S.B. (1994): *Fundamentals of Database Systems (2nd ed).* Benjamin/Cummings.

FOUNDATION FOR INFORMATION POLICY RESEARCH (1999): Aims and Purposes. Online, accessed 20 July 2000. URL: http://www.fipr.org/aims.html

FULDA, J.S. (1999): Solution to a Philosophical Problem concerning Data Mining. *Computers and Society* **26(4):6-7.**

GAVISON, R. (1984): Privacy and the limits of the law. In *Computers, ethics & social values.* 332-351. JOHNSON, D. and NISSENBAUM, H. (eds). Prentice-Hall.

HAMMOND, M. (2000): Oracle, Sybase move to boost database encryption. Online, accessed 14 August 2000. URL: http://www.zdnet.co.uk/news/2000/2/ns-12714.html

HERRING, S. (1994): Gender differences in computer-mediated communication: Bringing familiar baggage to the new frontier. In *CyberReader.* 144-154. V. VITANZA (ed). Allen and Bacon.

JOHNSON J. (1999): Using Oracle Database Auditing to Tune Performance. Online, accessed 14 August 2000. URL: http://oracle.com/oramag/oracle/99-Nov/69dba.html

LIN, T.Y., HINKE, T.H., MARKS, D.G. and THURAISINGHAM, B.M. (1996): Security and Data Mining. *Database Security IX: Status and prospects. Proceedings of the Ninth International conference on Database Security,* **9:** 391:399. Chapman & Hall.

MILLER, M. (1991): A model of statistical database compromise incorporating supplementary knowledge. In *Databases in the 1990's.* 97-113. SRINIVASAN, B. and ZELEZNIKOW, J. (eds). World Scientific.

MILLS, T. (1997): Multi-Level Secure Database Management Schemes. Online, accessed 19 July 2000. URL: http://www.sei.cmu.edu/str/descriptions/mlsdms.html

OFFICE OF NATIONAL HEALTH AND MEDICAL RESEARCH COUNCIL (1997): Joint NHMRC/AVCC Statement and Guidelines on Research Practices.

Online, accessed 19 July 2000. URL:
http://www.health.gov.au/nhmrc/research/nhmrcavc.ht
m

O'LEARY, D. (1991): Knowledge Discovery as a Threat
to Database Security. In *Knowledge Discovery in
Databases*. 507-516. PIATETSKY-SHAPIRO, G. and
FRAWLEY, W. J. (eds). AAAI Press.

ORGANISATION OF ECONOMIC COOPERATION
AND DEVELOPMENT (1980). Guidelines governing
the protection of privacy and transborder flows of
personal data. Online, accessed 1 May 2000. URL:
http://www.oecd.org//dsti/sti/it/secur/prod/PRIV-
EN.HTM

RACHELS, J. (1975): Why privacy is important.
Philosophy and Public Affairs 4(4):323-333.

RAINSFORD, C. and RODDICK, J.F. (1999): Database
issues in knowledge discovery and data mining.
Australian Journal of Information Systems 6 (2):101-
108.

RODDICK, J.F. and LEES, B. (2001): Paradigms for
Spatial and Spatio-Temporal Data Mining. In
Geographic Data Mining and Knowledge Discovery.
MILLER, H. and HAN, J. (eds). Taylor & Francis.

THURAISINGHAM, B.T. (1997): Security issues for
data warehousing and data mining. In *Database
Security X: Status and Prospects, Proceedings of the
Tenth International Conference on Database Security*,
10:11-20. Chapman & Hall.

UNIVERSITY OF SOUTH AUSTRALIA (1997):
Council Policies - Research. Online, accessed 19 July
2 0 0 0 . U R L :
http://www.unisa.edu.au/adminfo/policies/indexpln.ht
m#NCPR

WAGNER, J.L. (1994): Ethical attitudes of MIS
personnel. In *Managing Social and Economic Change
with Information Technology. Proceedings of the 1994
Information Resources Management Association
International Conference*. 53-57. Idea Group
Publishing.

WEIZENBAUM, J. (1972): On the impact of the
computer on society. *Science* 176(12):609-614.

WINOGRAD, T. (1992): Computers, ethics and social
responsibility. In *Computing and Human Values:
Proceedings of the 1991 Conference*. BYNUM, T.,
MANER, W, and FODOR, J. L. (eds). Research Center
on Computing and Society.