JOURNAL OF OBJECT TECHNOLOGY

Online at www.jot.fm. Published by ETH Zurich, Chair of Software Engineering ©**JOT, 2003**

Vol. 2, No. 3, May-June 2003

On US Homeland Security, Data Mining and Civil Liberties

Won Kim, Cyber Database Solutions, Austin, Texas

PROLOG

The US Defense Department's DARPA (Defense Advanced Research Projects Agency) has recently launched a Total Information Awareness (TIA) Initiative as part of national efforts to safeguard the US homeland from future terrorist threats. The vision behind the Initiative is to integrate public and private databases of all data relevant to monitoring the activities of potential terrorists and their supporters, and querying, analyzing, and mining the data. The databases are to include such information as records of deposits and withdrawals from banks, money transfers through banks, enrollment in schools, records of entry into and departure from the US, records of travels via airlines and car rentals, purchase records for goods and services using credit cards, etc. Further, US President Bush has authorized the creation of Terrorist Threat Integration Center, in part to allow integrated access to various databases of the US federal government, including the Department of Homeland Security, the CIA, the FBI, the INS (Naturalization and Immigration Services), etc. in order to identify terrorist suspects and supporters, and to track their activities in a more timely and coherent manner. What the US government is actually attempting to do is to create a central database out of hitherto disparate databases, and query, analyze, and mine the central database to determine terrorist activities in a much more timely, coherent, and accurate manner than has been possible up to now.

However, for consumption by the general public, the news media have been reporting that the objective of the TIA initiative and the Terrorist Threat Integration Center is to "develop massive data mining systems to monitor all data about all Americans". This overly simplified portrayal of the US initiatives in the news media is leading the public to equate the US initiatives to data mining technology, and, in turn, data mining technology to a major threat to civil liberties and privacy. Some US Senators and some news media are calling for a moratorium on the use of data mining technology in the TIA initiative in particular and on US homeland security in general until the issues of civil liberties and privacy are fully debated and addressed. However, portraying the current US initiatives as "developing massive data mining systems" is misleading and

Cite this column as follows: Won Kim: "On US Homeland Security, Data Mining and Civil Liberties", in *Journal of Object Technology*, vol. 2, no. 3, May-June 2003, pp. 33-37.
http://www.jot.fm/issues/issue_2003_05/column4

injurious to the large scientific community working on the research and development of data mining technology. The creation and use of the kind of information systems envisioned by the US government require many technologies, including database management, information retrieval, online analytical processing, query reporting, data visualization, statistical analysis, data mining, speech recognition, image/pattern recognition, image/pattern matching, natural language understanding, natural language translation, knowledge ontology, data warehousing, data integration, world-wide web, etc. In other words, data mining is merely one element in the array of required information technologies.

1 DATA MINING AND CIVIL LIBERTIES?

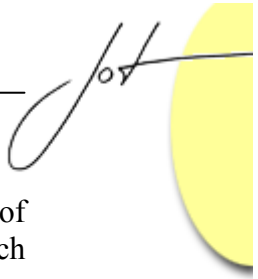
I believe that a debate on the impact of US national initiatives on civil liberties and privacy is necessary and healthy, and legitimate issues must be satisfactorily addressed. However, ill-founded or unfounded concerns and remedies proposed on the basis of such concerns should be quickly set aside.

Data mining is the process and a set of techniques for discovering valuable information, often called "intelligence", that is not explicitly stored in the database. It is based on longstanding theories of statistics. Data mining has proven to be very useful in detecting fraudulent uses of credit cards, fraudulent insurance claims, profiling customer purchasing behaviors, segmenting customers, anticipating mechanical failures of automobile parts, discovering new chemical groups and even new stars, etc. Data mining technology can indeed help in this process of discovering valuable intelligence in the form of terrorist profiles, patterns of activities such as movement of money and people, etc.

Data mining is not as magical or even automatic as the marketing literature of the vendors of data mining software and services tend to portray it. The patterns of data that data mining automatically discovers are often useless and irrelevant, and must be carefully analyzed and interpreted before useful intelligence can be obtained. Also before data mining is applied, the stored data has to be carefully organized and cleansed of dirty data – both very time-consuming and tedious processes. Further, data mining is far from infallible; it can generate lots of false positives. For example, fraud-detecting data mining systems used by credit-card companies often stop, and inconvenience, legitimate credit card users from making legitimate charges to their credit cards when their purchasing patterns deviate from their norm, such as the amounts, store locations, and frequency of purchases.

To be sure, when misused, data mining can potentially be used to violate civil liberties and privacy of people. But in this regard data mining is not alone. Many other information technologies can potentially violate civil liberties and privacy. I will discuss a few of them below.

One is database management technology. Whereas data mining attempts to automatically discover intelligence from a database, that is, deduce information that is not



explicitly stored in a database, database management allows people to retrieve some of the stored data. People can issue queries and retrieve data that satisfies certain search conditions from a database, for example, "find people who traveled more than six times during August 2001", "find people under the age of 45 who have no jobs but who have more than $100,000 in their bank accounts in Florida", etc. Database management systems from IBM, Microsoft, Oracle, etc. have been used all over the world during the past 20 years for managing all sorts of data in governments, businesses, educational institutions, etc.

Another is the radio frequency identification (RFID) technology. By attaching tiny chips that contain product or person identifiers on products and objects that people carry around or wear, locations of the products and people can be identified and such information can be transmitted wirelessly over the Internet and stored in a database. This technology can reveal the location.of a person (e.g., in a cancer treatment center in Houston), when the person may not wish for it to be known. Such information may have an adverse effect on the person's employment status with his/her employer.

The Internet has been used often by people to post such information as stolen credit card numbers, bank account numbers, medical records, host keys for activating commercial software products, even social security numbers and personal identities, etc. of unsuspecting people and businesses.

Should we place a moratorium also on the use of RDBs, RFID tags, and even the Internet because they can be used to violate civil liberties and privacy of people? I do not believe anybody will agree to it any more than the notion that we should place a moratorium on the use of cell phones, airplanes, emails, etc. because terrorist suspects are known to use them.

2 DATABASES AND CIVIL LIBERTIES?

Credit card companies, airline companies, rental car companies, hotels, banks, telephone companies, schools, stores, hospitals and doctors offices, etc. maintain in their computers records of customers and customer transactions. Further, federal, state, and municipal government departments and agencies maintain records of people, properties, finances, taxes, etc. When such data is centralized, disparate data can be correlated, and analyzed from additional perspectives. Then danger to privacy in theory increases.

To prevent violation of privacy, should we prohibit businesses, schools, and the government from storing data in their computers? Should we prevent conglomerates from centralizing data from multiple subsidiaries? I do not believe so. The reality today is that businesses and the government cannot operate, certainly cannot operate cost-effectively and competitively, without storing and making use of all sorts of data. In other words, whether people like it or not, lots of data about what they do are already in many databases, and there really is not much anyone can do to change the current situation in a reasonable way.

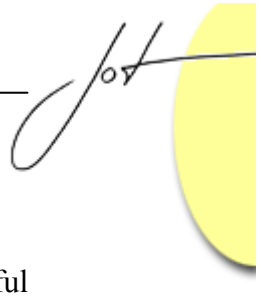
3 MISUSE OF DATA AND CIVIL LIBERTIES

I believe that the emerging debate about civil liberties and privacy should cover the entire array of information technologies, and be focused solely on the misuse of stored data. It certainly should not single out data mining, database management, the Internet, and not even the creation of a national database as envisioned by the TIA Initiative.

The apparently all-encompassing scope of the databases that the TIA initiative envisions creating and accessing has led to widespread concerns about the impact of a TIA initiative on civil liberties of Americans. There are at least two legitimate concerns. One is the possibilities of use of the databases for other purposes than anti-terrorism. The temptations and opportunities for such inappropriate use of the database exist, especially since the databases will contain "all data about everything that everyone does who resides in the US". Another is the inevitable false positive results of queries, analyses, and mining, that is, the wrong identification of otherwise innocent people as terrorist suspects or supporters. These are both consequences of misuse of data. I believe that the real danger to civil liberties lies mostly in the misuse of data, including the use of wrong data, the use of data in unauthorized ways, the wrong and unauthorized dissemination of data, and wrong conclusions from data. Misuse of data is where legal, technological, and procedural means can and should be devised and applied. Laws should be created and enforced to deter and punish illegal or unauthorized use and dissemination of data. Both the government and businesses should create and follow procedures to safeguard data from unauthorized use and dissemination by insiders and hackers from outside, and to take compensatory measures in the event that misuse of data has materially caused harm to customers and citizens. Technologies should be developed and applied to secure the data from hacking, to identify those who misuse data or disseminate data inappropriately, to minimize false positive results, to maintain and enhance integrity of stored databases, etc.

The following are some of the ways in which misuse of data can occur.

1. Wrong data may be entered into a database, such as a wrong entry date into the US, a misspelled name, incomplete address, etc.
2. Data may become outdated, for example, a person's address, travel record, etc.
3. Some poorly designed software do not allow wrong data to be corrected, for example, a person erroneously recorded as dead.
4. People may issue wrong or incomplete queries against a database, for example, "find a Saudi Arabian named Omar Mohamed" instead of "find a 25-year old Saudi Arabian named Abdul Saheed".
5. Analysts may erronesouly select an invalid result of data mining or OLAP analysis.
6. People may inappropriately or illegally disseminate both raw data and information obtained from the data (whether wrong or right) for money or otherwise, online or offline.



EPILOG

In my view, protecting civil liberties does not include protecting criminal and unlawful activities, such as committing electronic commerce frauds, money laundering, tax evasion, running child pornography sites on the Internet, scamming, hacking, spreading computer viruses, etc. I think it is entirely fair to deploy information technologies and query and mine databases to deter and detect people who commit such activities. In other words, such activities need not be protected under the broad umbrella of civil liberties and privacy. I think freedom from spam emails and (hardcopy) junk mails that people receive is arguably a privacy issue when senders of such mails obtained the email or postal addresses from entities that did not receive permission from the people involved to disseminate such addresses. The use of information technology in general, and data mining and database management technologies in particular, should not result in making life difficult for people who may be erroneously selected as suspects of unlawful or criminal activities. It also should not result in unwanted exposure of personal information, that is information unrelated to any unlawful or criminal activities, such as subscriptions to certain types of publication, memberships in certain types of organization, financial information, etc. To prevent such violation of civil liberties and privacy, and to minimize damages to people involved, laws, procedures, technologies must all be brought to bear.

About the author

Won Kim is President and CEO of Cyber Database Solutions (www.cyberdb.com) and MaxScan (www.maxscan.com) in Austin, Texas, USA. He is also Dean of Ewha Institute of Science and Technology, Ewha Women's University, Seoul. Korea. He is Editor-in-Chief of ACM Transactions on Internet Technology (www.acm.org/toit), and Chair of ACM Special Interest Group on Knowledge Discovery and Data Mining (www.acm.org/sigkdd). He is the recipient of the ACM 2001 Distinguished Service Award.