JOURNAL OF OBJECT TECHNOLOGY

Online at www.jot.fm. Published by ETH Zurich, Chair of Software Engineering ©**JOT, 2002**

Vol. 1, no. 5, November-December 2002

On Database Technology for US Homeland Security

Won Kim, Cyber Database Solutions, USA

In the wake of the September 11 terrorist attacks, the US government has initiated a wide array of measures to forestall replay of such attacks. One such measure is to encourage development and adoption of information technology to detect and apprehend suspects and security risks, and to manage emergencies should such an attack elude prevention. One element of information technology that is indispensable, particularly in preventing terrorism, is database technology.

For the purpose of this article, database technology is defined broadly as software and methodologies for modeling and storing large volumes of data of arbitrary types and structures and responding to queries and update requests against the data. Then database technology goes well beyond relational database technology to encompass technologies for storing and searching multimedia data (e.g., images, audio, speech, graphics, animation, video), geospatial data, temporal (time series) data, free-form text files, semi-structured data such as HTML and XML data, automatically discovering data in data (i.e., data mining), etc.

Understandably, technologies such as data mining, multimedia data search, speech understanding, document feature extraction, etc. are attracting keen interest as elements of technological solutions to enhancing US homeland security. For example, data mining may be used to automatically detect unusual patterns in the movement of terrorists through different cities and countries (five countries on the list of terrorists-harboring states in two weeks) or the movement of funds through chains of banks across countries. Multimedia technology, such as face recognition or image matching technology, may be used to identify people and automobiles photographed in airports, monuments, power plants, etc. Text mining technology may be used to automatically summarize free-form textual documents for indexing, classification, and subsequent search to help piece together paper trail left by terrorists and their supporters. Document feature extraction technology can be used to identify emails and other documents that contain certain keywords or combinations of keywords to alert authorities to potential terror and other criminal activities. Large-scale government systems for email and other communications snooping, such as the US National Security Agency's Echelon and US Federal Bureau of Investigation's Carnivore, are in operation. All these high-tech solutions are clearly of

Cite this column as follows: Won Kim: *On Database Technology for US Homeland Security,* in Journal of Object Technology, vol. 1, no. 5, November-December 2002, pages 43-49.
http://www.jot.fm/issues/issue_2002_11/column4

value to preventing terrorism and thereby enhancing homeland security. They work to varying degrees of sophistication, including various data mining algorithms (neural networks, decision trees), text summarization tools, data visualization tools, text search engines, some information classification tools, as well as systems for face recognition, image matching, voice matching, finger print matching, etc. However, many of the high-tech solutions require significant further research and development, especially in the areas of natural language understanding and content search and similarity matching of non-textual multimedia data (images, video, audio, and broadcast).

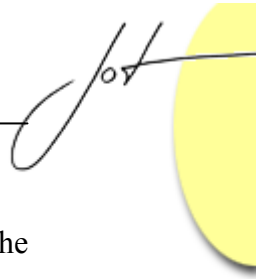
Further, before this "high-tech" side of database technology can be brought to bear, some serious work remains to be done on what researchers may view as the "low-tech" side of the technology. This includes work that actually requires significant innovation and work that is only massive and tedious. The purpose of this article is to bring out practical problems that can potentially render the "high-tech" side of database technology ineffective or even useless, and point to ways to solving these problems. The "low-tech" database issues discussed in the remainder of this article include legacy databases, federated databases, data quality, fuzzy queries, file classification, and data mining.

Before proceeding, I would like to remark that I do not know if any of the "low-tech" database issues I discuss in this article have actually been addressed, and to what extent, in confidential systems that the government may have created or adopted. The basis of my discussions is my knowledge of the capabilities and limitations of today's commercial database systems and the status of academic and industrial research into database technology. Further, I would like to remark that pursuit of homeland security using database and communications technologies inevitably raises issues of privacy of people. This article will ignore such issues.

Legacy Databases

The federal, state, and municipal governments in the US all maintain large numbers of databases to provide services to and monitor citizens and non-citizens alike. The US federal government has a large number of agencies with overlapping functions. A state government is organized into many departments, each department into bureaus, each bureau into programs, etc. The federal, state, county, and city governments have their own police forces and judiciary branches. One of the first requirements for the purpose of ensuring homeland security is to ensure that each database at every level of the government that has high relevance to homeland security be up to date and easily accessible. These include state motor vehicle registration, driver's license registration, police records, court records, immigration and naturalization registration, etc.

One serious problem today is that many of the government databases are legacy desktop database systems, such as Dbase, FoxPro, etc. In other words, these databases are not networked with other government databases, and as a result it is difficult to cross reference information across different agencies or departments. To allow easier and faster access and cross referencing, the legacy desktop databases of high relevance to homeland security at each level of government should be migrated and integrated into data marts or data warehouses managed by enterprise relational database systems. Various tools for



migrating data from legacy databases to a data mart or data warehouse have been on the market.

Federated Databases

The government does not have all the data relevant to homeland security. Business enterprises of all types, including, in particular, financial institutions, airline companies, shipping companies, telephone companies, rental car companies, hotels, educational institutions, religious organizations, charity organizations, etc. also maintain their databases, and some of the data they have can be of high relevance to homeland security. To detect security risks, potential terrorists, movements of funds, etc. requires, preferably easy and fast, cross referencing of specific data in multiple databases across organizations, possibly across countries even. For example, funds transfer activities stored in a chain of financial institutions and charity and religious organizations need to be cross referenced and correlated against databases of terrorist suspects in the databases of the CIA, FBI, NSA, INS, etc.

However, it is not reasonable to expect the disparate databases maintained by thousands of government agencies and hundreds of thousands of non-governmental organizations in the US to be integrated into a single humongous data warehouse. It is even more unlikely that databases of more than one country will be integrated into a single data warehouse. These databases will always be maintained independently. However, it is desirable to bring at least some of the independent databases into a federation of cooperating databases. For example, some of the databases of the CIA, FBI, NSA, INS, etc. may well be brought into federated databases. In the research literature, a virtually integrated global schema is assumed over the schemas of all the independent databases in the federation. Such an assumption is in general not practical, given the reality of database administration practices in different organization, and the difficulty associated with authorization and levels of security to assign to allow access to different databases. The capabilities of a federated database for the purpose of homeland security may be rather limited. Each participating site has the schemas of other databases so that it may initiate a cross-referencing or correlating query against some combination of other sites, and receive answers, preferably in real-time. A non-invasive "federation" layer should augment the database management systems that manage independent databases of all organizations that own data that are relevant to homeland security.

Data Quality

Databases almost always contain dirty data. Dirty data include missing data (e.g., gender data in questionnaire), wrong data (e.g., passport number with a missing digit, misspelled name), non-standard data (e.g., weight and money in non-US format), etc. Dirty data come about for a variety of reasons, including data entry error, non-update of changes (e.g., address change, financial status), intentional wrong entry, etc. Error checking and timely update are often skipped because of the overhead associated with them. Automatic error checking before data is entered into a database slows down database insertion.

Manual error checking (e.g., of misspelled names, wrong age, wrong phone number) obviously entails time and human resource overheads.

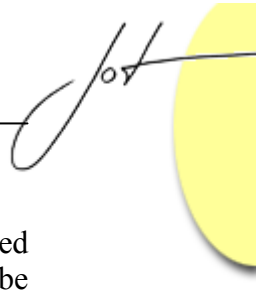
In the context of homeland security, one type of particularly troublesome dirty data is the names of people from such regions as Asia, Middle East, Africa, and Europe. Mohammad, Muhammad, Muhamed, and Mohamed; (former Prime Minister of China) Chou Enlai, Chu Eun-Lai, and Chou En Lai; Tsichristzis, Tsikrichis, and Tsichritsis – each set of these names may refer to one and the same person. Also, often only partial names (e.g., only the surname) may be stored in a database. Today's database systems are designed largely to do exact matches; that is, to find the exact string or number specified in a query. As a result, if the query asks for "Chou Enlai", but the person's information is stored in the database under the spelling "Chu Eun-Lai", no match will be found. Another type of troublesome dirty data is out-of-date addresses and affiliations. As people move about, updates in the database tend to fall behind and become obsolete. Efforts to search for a suspect will be delayed if the person's address and affiliation are, say, two years out of date.

Data quality procedures should really be defined and followed to maintain the highest plausible quality of data at all times. Such procedures would include preventing dirty data from entering the database, and detecting and correcting dirty data that do make it into the database. Value range testing may be done automatically (on such data as social security number, passport number) to prevent out-of-range data from entering the database. A spellchecker may be run on text data (such as names and addresses). Redundant data may be created for cross validation. Database triggers may be written and stored in the database to automatically update changeable data as early as possible after a certain event occurs.

Fuzzy Queries

A fuzzy query is a query against incomplete, imprecise, or even dirty data in a database, or a query whose search conditions are imprecise. Today's database systems are designed largely for 'precise' queries against a database of 'precise and complete' data. Range queries (e.g., 'Age BETWEEN 20 AND 30'), disjunctive queries (e.g., 'Name = "Mohammad" OR Name = "Muhammad"), and regular expression queries (e.g., 'Address CONTAINS "land"') do allow for some 'imprecision' in queries. However, these fall far short as fuzzy queries. The need for fuzzy queries arises for two reasons: as observed earlier, the stored data is often dirty or imprecise, and the query conditions can often be imprecise. An example is "find a person whose last name sounds like 'Napalu', is perhaps 'middle aged', and who drives an 'old' 'white-ish' car whose license plate contains the letters 'TR'".

To support fuzzy queries in the context of homeland security, many possible spellings of the names of people originating from various regions of the world, and descriptive terms for automobiles, people, incidents, etc. need to be maintained in a sort of "names thesaurus" along with rules for matching similar names and descriptions. Further, certain types of data in databases need to be re-organized in accordance with proper levels of abstraction or categorization, or support for searches in implicit levels of



abstraction or categorization must be provided. Suppose that 'pistol' and 'rifle' are stored in the Weapons-Training field in a table. A system to support fuzzy queries should be made to recognize that 'pistol' and 'rifle' are generalized into 'guns', and in response to a query 'find persons who received Weapons-Training in guns', the system should return all those who are trained in the use of 'pistols' and 'rifles', even if 'guns' may not appear anywhere in the Weapons-Training field.

Fuzzy queries also involve geospatial and temporal search conditions, such as 'near', 'within', etc. Fortunately, much research has been done on spatial and temporal data management, including spatial and temporal search conditions and spatial indexing mechanisms.

Supporting fuzzy queries will also require serious research into performance techniques. Today's relational database systems, as they are largely designed to support precise queries against a precise database, and use such 'precise' access support mechanisms as indexing, hashing, and sorting. Such mechanisms are used for fast selective searches of records within a table and for joining two tables based on 'precise' matching of values in join fields in the tables. The imprecise nature of the search conditions and/or the stored data makes such access mechanisms largely useless. There are some obvious techniques that can be used to address the resulting performance and scalability problems, such as compressed storage of data and partitioned storage of data across multiple computers, etc. However, additional research is necessary to go beyond these.

Information Classification

Indexing is an essential mechanism for reducing the search space to find desired data in a large database, be it a corporate database, a government agency's database, or the entire World Wide Web. Database systems create and maintain indexes on user-specified fields of a table to expedite searches that involve the indexed fields. Similarly, information retrieval (document retrieval) systems create and maintain indexes, in the form of lists of words that appear in free-form textual documents, to expedite searches of documents that contain certain words or combinations of words. Internet search engines create keywords representing HTML documents, and use them as indexes into such documents. Research into semantic Web is receiving significant attention these days. Its goal is to make searches possible based on semantic understanding of user's requests and of the stored Web documents.

In the interim, a good information classification technique should serve as a powerful 'high-level' indexing mechanism to support fast and accurate identification of free-form textual documents. For example, a document that discusses a meeting between Muslim militants and al Qaeda operatives in Malaysia in 2000 may be classified under several subject categories rather than just one: one about Muslim militants, one about al Qaeda, and one about Islam in Malaysia, etc. Often, classifying a document under only one subject will tend to make the document irretrievable. For example, it is often difficult to retrieve even old emails if they have been saved in certain folders under names deemed

appropriately representative of the contents of the emails. At a later point one often does not recall the names of the folders, and may attempt to recall the names in terms of some of the main messages in the mails that one recalls. Today's text mining technology allows for feature extraction (proper names, etc.), keyword frequency count, and even summarization of free-form texts. Summarizations of documents provide a reasonable basis for classifying the documents.

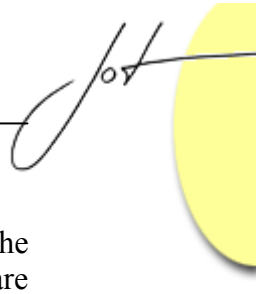
Today, technology for matching a sample image, sound, or video clip against a stored database of images, sound, and video is not mature. Many technical challenges remain in dealing with shapes, texture, color, size, background isolation, scene shifting, etc. Again, until the technology matures, the tedious but reasonably effective means must be used to deal with multimedia data. These include manually tagging all photographs, images, voice recordings, video clips, etc. that are relevant to homeland security, and store the tags in the database for easy and fast searches. Such tags may also be properly classified into appropriate subject categories for use as indexes. Thumbnails of such data may also be produced and stored for quick browsing to eliminate the need to load full data that are not needed.

Data Mining

Data mining is automatic discovery of intelligence from raw data stored in computer systems. It has been used to detect fraudulent uses of credit cards and phone cards, fraudulent claims of insurance benefits, for customer churn prediction, for customer segmentation, etc. Given the reality of dirty data in databases, and the objective of homeland security, data mining techniques need to be retooled or redirected. In particular, tolerance of errors in data has to be an important criterion in the selection of data mining algorithms. Further, whereas data mining algorithms are typically used to discover unforeseen broad patterns and trends, the search for security risks, terrorist suspects and movement of funds will tend to require discovery of exceptional and unusual patterns (outliers). As such, data mining algorithms better-suited for discovery and analysis of outlier data would be more appropriate. Moreover, allowance must be made for dirty data when determining the amount of data for training data mining models. Without these considerations, the results of data mining may not be reliable.

Concluding Remarks

This article discussed several database technology issues that need to be addressed to help the US government combat terrorism on US soil. This is merely the tip of the iceberg. The problem of protecting the US homeland from terrorism is truly monumental. The US is a country with vast borders, is a country into and out of which an incredible number of people and materials flow daily, and is a country founded on freedom and liberty for individuals. These make it impossible to guarantee security. Although the enemy numbers only about a few thousands, they are fanatic, determined, cunning, faceless and stateless. To make the problem even more difficult, there is also the danger of US-bred terrorism from the far-right militants such as those who bombed the federal building in Oklahoma City.



The damages they can inflict on the lives, way of life, economy, and the infrastructure of the US, and indeed the entire industrialized modern world, are potentially enormous.

Database technology is only one of several technologies that can be brought to bear in combating terrorism. Other technologies include communications (tracking down mobile phone users, prepaid phone card users, tapping phone and fax communications, etc.), security (securing computer and communications networks, tracking down hackers and cyber attackers, tracking down identities of people posting messages on the Internet, encrypted communication, cracking encrypted messages, etc.), robotics (for search and rescue), biometrics, multimedia technology (voice analysis and matching, face recognition, etc.), weapons and hazardous materials detection, etc.

Although clearly indispensable, technology is only one element in the arsenal for combating terrorism and securing the US homeland. Certain laws that prevent government agencies from cooperating (e.g., between the CIA and the FBI) and that prevent authorization for wiretapping, etc. have to be relaxed. Key personal identifications, such as passports, green cards (permanent resident visas), driver's licenses, and social security cards, should be made more difficult to forge. The nation's economic and social infrastructure, such as the energy supply network, water supply network, transportation systems, etc., should all be made more secure.

Beyond these the US needs to forge and maintain close alliance with other governments to receive intelligence and thwart terrorism. As various political pundits have claimed, the fundamental solution may very well lie in addressing the root causes behind the rise of Islamic militancy, namely the disenfranchised youth of the Muslim countries, the excesses of the Western industrialized world, and the Israel-Palestine issue.

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.