Chapter 2

AN ISI RESEARCH FRAMEWORK: INFORMATION SHARING AND DATA MINING

Chapter Overview

To address the data and technical challenges facing ISI, we present a research framework with a primary focus on KDD (Knowledge Discovery from Databases) technologies. The framework is discussed in the context of crime types and security implications. Selected data mining techniques, including information sharing and collaboration, association mining, classification and clustering, text mining, spatial and temporal mining, and criminal network analysis, are believed to be critical to criminal and intelligence analyses and investigations. In addition to the technical discussions, the chapter also discusses caveats for data mining and important civil liberties considerations.

2.1 Introduction

Crime is an act or the commission of an act that is forbidden, or the omission of a duty that is commanded by a public law and that makes the offender liable to punishment by that law. The more threat a crime type poses on public safety, the more likely it is to be of national security concern. Some crimes such as traffic violations, theft, and homicide are mainly in the jurisdiction of local law enforcement agencies. Some other crimes need to be dealt with by both local law enforcement and national security authorities. Identity theft and fraud, for instance, are relevant at both the local and national level -- criminals may escape arrest by using false identities; drug smugglers may enter the United States by holding counterfeited passports or visas. Organized crimes, such as terrorism and narcotics trafficking, are often diffuse geographically, resulting in common security concerns across cities, states, and countries. Cybercrimes can pose threats to public safety across multiple jurisdictional areas due to the widespread nature of computer networks.

Table 2-1 summarizes the different types of crimes sorted by the degree of their respective public influence (Chen et al., 2004a). International and domestic terrorism, in particular, often involves multiple crime types (e.g., identity theft, money laundering, arson and bombing, organized and violent activities, and cyber-terrorism) and causes great damage.

2.2 An ISI Research Framework

We believe that KDD techniques can play a central role in improving counter-terrorism and crime-fighting capabilities of intelligence, security, and law enforcement agencies by reducing the cognitive and information overload. Knowledge discovery refers to non-trivial extraction of implicit, previously unknown, and potentially useful knowledge from data. Knowledge discovery techniques promise easy, convenient, and practical exploration of very large collections of data for organizations and users, and have been applied in marketing, finance, manufacturing, biology, and many other domains (e.g., predicting consumer behaviors, detecting credit card frauds, or clustering genes that have similar biological functions) (Fayyad and Uthurusamy, 2002). Traditional knowledge discovery techniques include association rules mining, classification and prediction, cluster analysis, and outlier analysis (Han and Kamber, 2001). As natural language processing (NLP) research advances, text mining approaches that automatically extract, summarize, categorize, and translate text documents have also been widely used (Chen, 2001; Trybula, 1999).

Many of these KDD technologies could be applied in ISI studies (Chen et al., 2003a; Chen et al., 2004b). Keeping in mind the special characteristics of crimes, criminals, and crime-related data, we categorize existing ISI technologies into six classes: *information sharing and collaboration*, *crime association mining*, *crime classification and clustering*, *intelligence text mining*, *spatial and temporal crime mining*, and *criminal network mining*.

Table 2-1. Crime types and security concerns.

	Crime Type	**Local Law Enforcement Level**	**National Security Level**
Increasing Public Influence ↓	Traffic Violations	Driving under influence (DUI), fatal/personal injury/property damage, traffic accident, road rage	-
	Sex Crime	Sexual offenses, sexual assaults, child molesting	Organized prostitution, people smuggling
	Theft	Robbery, burglary, larceny, motor vehicle theft, stolen property	Theft of national secrets or weapon information
	Fraud	Forgery and counterfeiting, fraud, embezzlement, identity deception	Transnational money laundering, identity fraud, transnational financial fraud
	Arson	Arson on buildings, apartments	-
	Organized Crime	Narcotic drug offenses (sales or possession), gang-related offenses,	Transnational drug trafficking, terrorism (bioterrorism, bombing, hijacking, etc.)
	Violent Crime	Criminal homicide, armed robbery, aggravated assault, other assaults	Terrorism
	Cyber Crime	Internet fraud (e.g., credit card fraud, advance fee fraud, fraudulent web sites), illegal trading, network intrusion/hacking, virus spreading, hate crimes, cyber-piracy, cyber-pornography, cyber-terrorism, theft of confidential information	

These six classes are grounded on traditional knowledge discovery technologies with a few new approaches added, including spatial and temporal crime pattern mining and criminal network analysis, which are more relevant to counter-terrorism and crime investigation. Although information sharing and collaboration are not data mining *per se*, they help prepare, normalize, warehouse, and integrate data for knowledge discovery and thus are included in the framework.

In Figure 2-1 we present our proposed research framework, with the horizontal axis being the crimes types and the vertical axis being the six classes of techniques (Chen et al., 2004a). The shaded regions on the chart show promising research areas, i.e., that a certain class of techniques is relevant to solving a certain type of crime. Note that more serious crimes

may require a more complete set of knowledge discovery techniques. For example, the investigation of organized crimes such as terrorism may depend on criminal network analysis technology, which requires the use of other knowledge discovery techniques such as association mining and clustering. An important observation about this framework is that the high-frequency occurrences and strong association patterns of severe and organized crimes such as terrorism and narcotics present a unique opportunity and potentially high rewards for adopting such a knowledge discovery framework.

Several unique classes of data mining techniques are of great relevance to ISI research. *Text mining* is critical for extracting key entities (people, places, narcotics, weapons, time, etc.) and their relationships presented in voluminous police incident reports, intelligence reports, open source news clips, etc. Some of these techniques need to be multilingual in nature, including the abilities for machine translation and cross-lingual information retrieval (CLIR). *Spatial and temporal mining and visualization* are often needed for geographic information systems (GIS) and temporal analysis of criminal and terrorist events. Most crime analysts are well trained in GIS-based crime mapping tools; however, automated spatial and temporal pattern mining techniques (e.g., hotspot analysis) have not been adopted widely in intelligence and security applications.

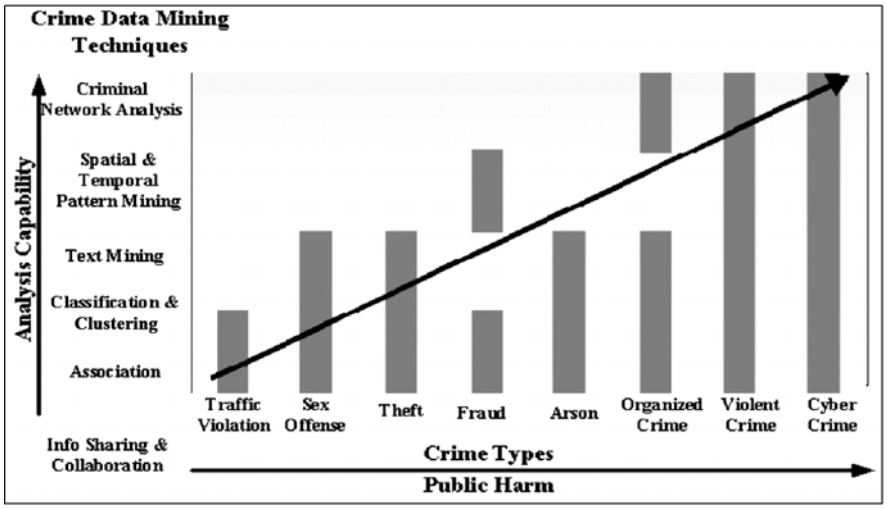

Figure 2-1. A knowledge discovery research framework for ISI.

Organized criminals (e.g., gangs and narcotics) and terrorists often form inter-connected covert networks for their illegal activities. Often referred to

as "dark networks," these organizations exhibit unique structures, communication channels, and resilience to attack and disruption. New computational techniques, including social network analysis, network learning, and network topological analysis (e.g., random network, small-world network, and scale-free network), are needed for the systematic study of those complex and covert networks. We broadly consider these techniques under *criminal network analysis* in Figure 2-1.

2.3 Caveats for Data Mining

Before we review in detail relevant ISI-related data mining techniques, applications, and literature in the next chapter, we wish to briefly discuss the legal and ethical caveats regarding crime and intelligence research.

The potential negative effects of intelligence gathering and analysis on the privacy and civil liberties of the public have been well publicized (Cook and Cook, 2003). There exist many laws, regulations, and agreements governing data collection, confidentiality, and reporting, which could directly impact the development and application of ISI technologies. We strongly recommend that intelligence and security agencies and ISI researchers be aware of these laws and regulations in research. Moreover, we also suggest that a hypothesis-guided, evidence-based approach be used in crime and intelligence analysis research. That is, there should be probable and reasonable causes and evidence for targeting particular individuals or datasets for analysis. Proper investigative and legal procedures need to be strictly followed. It is neither ethical nor legal to "fish" for potential criminals from diverse and mixed crime-, intelligence-, and citizen-related data sources. The well-publicized Defense Advanced Research Program Agency (DARPA) Total Information Awareness (TIA) program and the Multi-State Anti-Terrorism Information Exchange (MATRIX) system, for example, have recently been shut down due to their potential misuse of citizen data and impairment of civil liberties (American Civil Liberties Union, 2004; O'Harrow, 2005).

2.4 Domestic Security, Civil Liberties, and Knowledge Discovery

In an important recent review article by Strickland, Baldwin, and Justsen (Strickland et al., 2005), the authors provide an excellent historical account of government surveillance in the United States. The article presents new surveillance initiatives in the age of terrorism (including the passage of the USA PATRIOT Act), discusses in great depth the impact of technology on

surveillance and citizen's rights, and proposes balancing between needed secrecy and oversight. We believe this is one of the most comprehensive articles addressing civil liberties issues in the context of national security research. We summarize some of the key points made in the article in the context of our proposed ISI research. Readers are strongly encouraged to refer to (Strickland et al., 2005) for more details.

Framed in the context of domestic security surveillance, the paper considers surveillance as an important intelligence tool that has the potential to contribute significantly to national security but also to infringe civil liberties. As faculty of the University of Maryland Information Science department, the authors believe that information science and technology has drastically expanded the mechanisms by which data can be collected, and knowledge extracted and disseminated through some automated means.

An immediate result of the tragic events of September 11, 2001, was the extraordinarily rapid passage of the USA PATRIOT Act in late 2001. The legislation was passed by the Senate on October 11, 2001, by the House on October 24, 2001, and signed by the President on October 26, 2001. The continuing legacy of the then-existing consensus and the lack of detailed debate and considerations created a bitter ongoing national argument as to the proper balance between national security and civil liberties. The PATRIOT Act contains ten titles in 131 pages. It amends numerous laws, including, for example, expansion of electronic surveillance of communications in law enforcement cases; authorizing sharing of law enforcement data with intelligence; expansion of the acquisition of electronic communications as well as commercial records for intelligence use; and creation of new terrorism-related crimes.

However, as new data mining and/or knowledge discovery techniques become mature and potentially useful for national security applications, there are great concerns of violating civil liberties. Both the DARPA's TIA Program and the Transportation Security Administration's (TSA) Computer Assisted Passenger Prescreening Systems (CAPPS II) were cited as failed systems that faced significant media scrutiny and public opposition. Both systems were based on extensive data mining of commercial and government databases collected for one purpose and to be shared and used for another purpose; and both systems were sidetracked by a widely perceived threat to personal privacy. Based on much of the debate generated by these programs, the authors suggest that data mining using public or private sector databases for national security purposes must proceed in two stages – first, the search for general information must ensure anonymity; second, the acquisition of specific identity, if required, must be by court order under appropriate standards (e.g., in terms of "special needs" or "probable causes").

In their concluding remarks, the authors cautioned that secrecy in any organization could pose a real risk of abuse and must be constrained through effective checks and balances. Moreover, information science and technology professionals are ideally situated to provide the tools and techniques by which the necessary intelligence is collected, analyzed, and disseminated; while civil liberties are protected through established laws and policies.

In addition to the review article by Strickland et al., readers are also referred to an excellent book entitled: *No Place to Hide*, written by Washington Post reporter Robert O'Harrow (O'Harrow, 2005). He reveals how the government is creating a national intelligence infrastructure with the help of private information, security, and technology companies. The book examines in detail the potential impact of this new national security system on our traditional notions of civil liberties, autonomy, and privacy.

2.5 Future Directions

National security research poses unique challenges and opportunities. Much of the established data mining and knowledge discovery literature, findings, and techniques need to be re-examined in light of the unique data and problem characteristics in the law enforcement and intelligence community. New text mining, spatial and temporal pattern mining, and criminal network analysis of relevance to national security are among some of the most pressing research areas. However, researchers cannot conduct research in a vacuum. Partnerships with local, state, and federal agencies need to be formed to obtain relevant test data and necessary domain expertise for ISI research. Only after rigorous testing with scrubbed or anonymous data can selected techniques be field examined and verified by the domain experts (i.e., law enforcement personnel, intelligence analysts, and policy makers). These techniques should be used in actual investigations only after experts have confirmed their potential value. At this stage, the researcher-designed algorithms or systems are often much improved and refined, and are often operated and controlled by the domain experts with their own heuristics, know-how, and judgment.

2.6 Questions for Discussion

1. What are the newest and most promising data mining techniques and approaches? How much of the data mining research can be conducted using existing tools and software and how much should be built from scratch?

2. How can new, promising, and unique data mining techniques for national security be identified? What are some ways to develop an integrated, multi-disciplinary research team for algorithms development, system development, user interface design, user assessment, and organizational impact study?
3. How can agency partners who are willing to collaborate, i.e., providing data and domain expertise, be identified? How should a win-win scenario for the research partnership be created?
4. What are some ways and places to find help on privacy and civil liberties issues? How can rigorous data mining research be conducted in consideration of civil liberties?
5. What are some ways to work with industry partners on national security research?