DELHI TECHNOLOGICAL UNIVERSITY

DATA WAREHOUSE AND DATA MINING

IT-405

CASE STUDY PAPER

CRIMINAL DATA RECORD ANALYSIS

Submitted To:

Ms Priyanka Meel

Department of IT, DTU

Submitted By: (Group No:8)

Aaryaman Bajaj (2K17/IT/02)

Himanshi Nimesh (2K17/IT/52)

ABSTRACT

In the current era, the number of crimes occurring in the society and this criminal rate increases day by day. Crime has negatively influenced cultures. Crime control is essential for the welfare, stability and development of society. Law enforcement agencies are seeking the system to target crime structure efficiently. There is a tremendous growth of criminal data.

However, information overload hinders the practical analysis of criminal and terrorist activities. The intelligent crime data analysis provides the best understanding of the dynamics of unlawful activities, discovering patterns of criminal behaviour that will be useful to understand where, when and why crimes can occur. There is a need for the advancements in the data storage, collection, analysis and algorithm that can handle data and yield high accuracy. Data mining applied in the context of intelligence analysis holds the promise of alleviating such problems. We attempt to summarise the challenges arising during the analysis process, which should be removed to get the desired result.

INTRODUCTION

Criminal analysis and investigation is the process to explore and detect crime and unlawful relationships. There are lots of data related to the crime in police station records, information related to the particular crime or the essential information which is directly or indirectly related to crime should be extracted. Hence there is a need for such technology, which separates all these data from colossal content. Based on previously known (historical) crime and criminals relationship record, the criminal investigation team can extract useful information so that they can identify the facts related to the committed crime and minimise the future crime possibilities. Criminal investigation acts on criminal cases like murder cases, child abuse, threats, hacking, financial crime detection like money-laundering, terrorism funding, fraud, etc. So the criminal investigation team should use techniques so that they can predict the future crime trends based on available historical criminal data and in this way, the future crime rate will decrease. The need for criminal investigation is to identify and apprehend the criminal if a crime has been committed and provide the evidence to support a conviction in court.

The criminal investigation is the process to seek the methods, motives and identities of criminals and prove the guilt of a criminal. Crime investigation refers to the process to discover important information relevant to the crime. Investigation can be done by Evidence preservation, interviewing, record collection, electronic discovery, forensic anthropology, investigation and search warrants, email trace, criminal forensics, intelligence gathering, etc.

One challenge to intelligence agencies is the difficulty of analysing large volumes of data involved in criminal activities. Data mining holds can make it easy, convenient, and practical to explore vast databases for organisations and users.

Various technologies such as association, classification, clustering are used in criminal investigations in data mining. Crime investigation is done by using artificial intelligence methods. Lots of work has been done in this field like mining criminal databases to find investigation clues in the case of financial crime detection stolen automobiles.

DESIGNING A DATA WAREHOUSE

Kimball's design process consists of the following four-steps:

Step 1: Identify the business process- cyber crime investigation activity

Step 2: Determine the grain of a fact table, representing the level of the detail of the data warehouse data record to be analyzed. We can think about two choices - incidence and attack. An incidence is an abnormal activity that may or may not result in an attack. If we just want to analyze cyber attacks that actually resulted in crimes or damages, we can use the Attack fact table. On the other hand, if we analyze cyber crimes at each incidence level, we can use the Incidence fact table. Since many incidences, whether they may or may not result in any attack, are still important to track down, the Incident fact table is more powerful. The Incident fact table, however, may result in a larger number of rows than the Attack fact table.

Step 3: Identify the dimensions used to analyze the fact table. In the Attack schema, the dimensions can be Date, Attacker, Attacker Demographics, Attack pattern, Attack status, Law enforcement, Target, Target Agency, and Incidence Summary. In the Incidence fact table, the dimensions can be Date, Attacker, Attacker Demographics, Attack pattern, Attack status, Law enforcement, Target, Target Agency, and Attack. With this design, all the related incidences for a single attack can be easily aggregated for the attack.

Step 4. Identify the 'measure data' of the fact table. We select the same measure data for both Attack and Incidence fact tables. We first include Cyber Crime ID, which is the primary key of the source database from which the cyber crime data came. This attribute will be useful in connecting the source database and the data warehouse. This attribute thus supports real-time analysis using the data warehouse. Other selected measures are Loss in Dollars, Cost for fix, Actual Downtime, Cost for Downtime, and Cost for Exposed Confidential Data. The following are many-to-many relationships; the tools used by attackers, political affiliations joined by attackers, institutions the attacker attended, multiple Websites attacked by attackers, skills owned by attackers, etc. These data could be useful in analyzing cyber crimes.

Crime Analyses Using:

1. OLAP

OLAP enables a user to effectively extract and view information from different points-of-view. OLAP can locate the intersection of dimensions and report them. From the dimensional model, we can perform a number of analyses. If our focus is the attacker, then we can run queries that would tell us who has performed what certain types of attacks in the past, who tends to work in groups, and who would be a leader in those groups. We can query for recidivism, levels of technical skills, and affiliations. This last would be of particular interest to those agencies involved in anti-terrorist and homeland defence effort. If the focus of our investigations is attacks, then the model supports queries that would show which agencies were targeted, what tools were used, what was expected to be gained, and what types of skills were required for a given type of attack. Target-related investigations would be able to query for agencies that were highly targeted, and if the attacks were successful or vulnerable. The model also supports analysis of vulnerabilities, specifically addressing what systems, architectures, and operating systems that were most vulnerable. While the media press is generally full of articles saying which OS has a security problem, the query results would provide more reliable proof.

2. Data Mining

Although OLAP is a key component of the analytical process, it alone is not a sufficient tool for better understanding of cyber crime data and designing preventive methods against the cyber attacks. For example, to answer the following question “If a password theft attack happens, what is the type of attack most likely to happen next?”, it is very difficult to find a satisfactory answer based solely on the OLAP from the cyber forensic data warehouse. But the answer to the above question is very important to help the institutes reduce the damage caused by the attack.

Data mining techniques are used to identify patterns in a set of data. It looks for patterns where one event is connected to another event (association), patterns where one event leads to another later event (sequence or path analysis), and new patterns (classification). It can also offer a visual combination of newly documented facts (clustering), and analysis of patterns in data that can lead to reasonable predictions about the future (forecasting). Data mining can be applied to various log analysis and intrusion detection systems. A lot of mining algorithms and methods such as association algorithms, decision trees, and others can be applied for mining the cyber forensic data warehouse to derive insightful knowledge rules to help understand the attacks and protect the network security.

DATA MINING TECHNIQUES

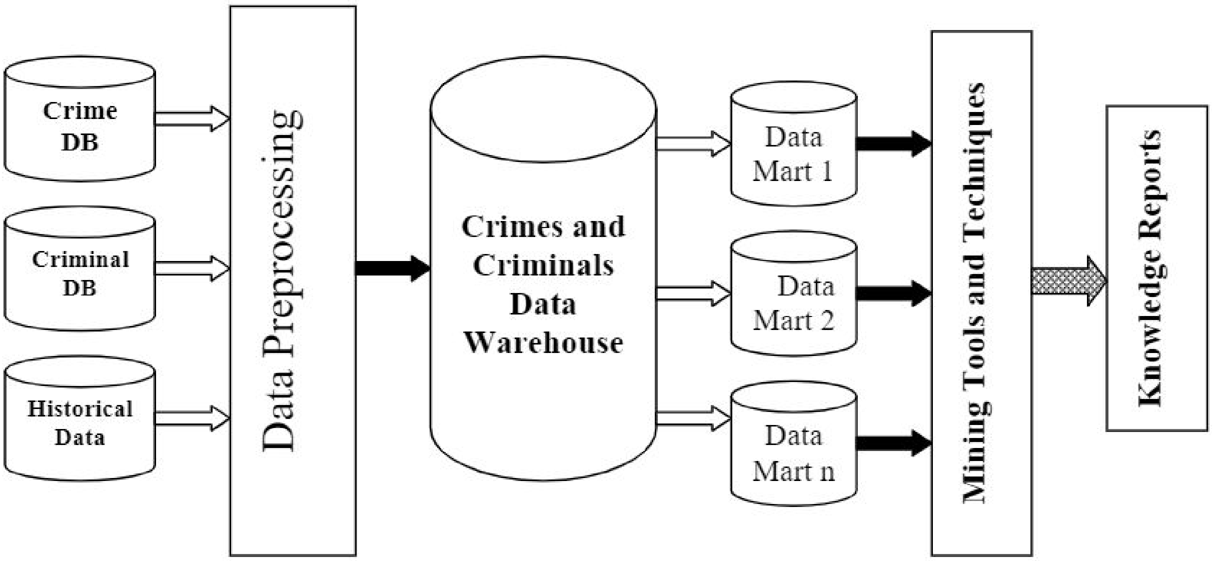
A combination of different data mining techniques is used to obtain more accuracy.

- Entity extraction: Entity extraction is the process of identifying the particular patterns so that they provide basic information for the crime analysis. Entity extraction is used to extract valuable information of person address, time, crime type, personal property, suspect description relevant to particular cases automatically. It is the process to identify the potential suspects.
- Link analysis: Link analysis is used to analyze the criminal incident & forms a network of the suspect. Social network analysis is used to analyze the associated elements of criminals in the criminal network for disrupting the network.
- Classification: Classification technique is a supervised machine learning method. It divides the dataset based on some predefined condition. It is the process to specify the class of the object to which it belongs. In mail spamming, classification is used. These are used for detection of specific activities of the criminals in large sized data sets, classify the crime activities into different categories and predict crime hotspots.
- The K-nearest neighbor algorithm (K-NN) is a classification algorithm used for classifying the objects. It is a simple machine learning algorithm. It is used to determine similarity between train and test record.
- Artificial neural network is the interconnection network of processing elements known as neurons. ANN mimics the cognitive, neurological functions of the human brain. Inputs are multiplied by weights and produce outputs as labels. Neural network techniques are used for entity extraction from the criminal data records. Its prediction accuracy is high. It is used to identify the crime hot spots of high level.
- The decision tree is a tree-like structure which demonstrates the flow of data where testing over the attributes is performed at each node and on the basis of condition correctly label the objects. The decision tree is used to detect suspicious mails and provide accuracy in classifying emails.
- Support vector machine (SVM) is a supervised ML algorithm that is used for classification problems by separating hyperplanes. SVM uses the Kernel function.
- The Naïve Bayes Belief network is a probabilistic model that is used for the classification. Bayesian theorem is also used in crime analysis. Its accuracy is good. It demonstrates the variable using directed acyclic graphs.
- Clustering: Clustering is an unsupervised machine learning algorithm. Clustering algorithm is used to group the records of similar type and dissimilar type of objects are grouped in different groups. Clustering provides efficiency in identifying crime zones and trends and in this way crimes can be controlled. Self organizing map, link analysis technique, hierarchical clustering, DB Scan, K-means clustering to detect hotspots, to automatically identify the association.
- K-mean clustering is used to partition the data into k- clusters based on their mean. Agglomerative algorithm and partitional algorithm are used for hierarchical clustering.
- An Intelligent agent is a computing agent which performs tasks autonomously. Agents monitor & identify the real time response and generate alert messages through emails & instant messages. When deformities are encountered, the agents deliver messages to alert the criminal investigators. It increases the efficiency & accuracy in criminal investigation.

- Text mining: Text mining is used to extract information from textual dataset. Natural language processing is used to identify the relevant entities. It compares phrases or sentences to extract associations within the criminal network.

SVM is used for identifying digital evidence related to computer crime. ANN provides higher accuracy than logistic regression when logistic applied to identify smuggling vessels. The SVM approach provides better accuracy as compared to a multilayer perceptron neural network. ANNs, decision trees and logistic regression are used for uncovering lies from statements of different types of crimes. These data mining techniques are used for auto insurance fraud. Nearest Neighbor, Decision tree, SVMs, Naïve Bayes are used. ANNs perform better as compared to a decision tree and SVM, and provide greater accuracy.

Framework:



CONCLUSION

There are a number of factors responsible for the rising of crimes at an alarming rate in India like illiteracy, poverty, unemployment, migration, frustration & corruption. Intelligence agencies search the database manually, which is a tedious task and consume more time. New advanced technologies & tools are used for combating crimes and to identify criminals.

Data analysis using data mining techniques needs a huge amount of historical data that may exceed the expectation from the model and the framework. New methodologies and analytical techniques should be explored to address the fundamental challenges of criminal data, and to leverage big data to facilitate criminal investigations. Big data analytics has the potential to transform the way that law enforcement and security intelligence agencies extract vital knowledge from multiple data sources in realtime to support their investigations.

We have presented dimensional models for a data warehouse for cyber forensics. We have also discussed ways of utilizing the data warehouse by considering the types of analysis as well as using OLAP and data mining technologies.

From the encouraging results, we believe that crime data mining has a promising future for increasing the effectiveness and efficiency of criminal and intelligence analysis. Many future directions can be explored in this still young field. More visual and intuitive criminal and intelligence investigation techniques can be developed for crime pattern and network visualization.

REFERENCES

1) https://www.researchgate.net/publication/236853080_Crime_Data_Analysis_Using_Data_Mining_Techniques_To_Improve_Crimes_Prevention_Procedures
2) https://pweb.fbe.hku.hk/~mchau/papers/CrimeDataMining.pdf
3) https://www.researchgate.net/publication/4355489_An_Analysis_of_Data_Mining_Applications_in_Crime_Domain
4) https://www.researchgate.net/publication/320060302_Crime_Data_Mining_an_Indian_Perspective
5) https://commons.erau.edu/cgi/viewcontent.cgi?article=1001&context=adfsl&httpsredir=1&referer=
6) https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwja0YaKlvrsAhWr6nMBHa0ACWAQFjABegQIARAC&url=https%3A%2F%2Fzenodo.org%2Frecord%2F1197513%2Ffiles%2FIJETMR18-CINSP-10.pdf&usg=AOvVaw0rD3nJNUPE62WIwtW9P-c4
7) https://www.researchgate.net/publication/2870463_Crime_Data_Mining_An_Overview_and_Case_Studies