WEB DATA MINING: VIEWS OF CRIMINAL ACTIVITIES

B.M. Monjurul Alom

Assessment Research Centre, Melbourne Graduate School of Education
Level 8, 100 Leicester Street, The University of Melbourne, Victoria 3010 Australia

ABSTRACT: *Web data mining discovers valuable information or knowledge from the web hyperlink structure, page content and usage data. Along with the swift popularity of the Internet, crime information on the web is becoming increasingly flourishing, and the majority of them are in the form of text. A major challenge facing all law-enforcement and intelligence-gathering organizations is accurately and efficiently analyzing the growing volumes of crime data. Detecting, exploring crimes and investigating their relationship with criminals are a big challenge to the present world. The evaluation of the different dimensions of widespread criminal web data causes one of the research challenges to the researchers. Criminal web data always offer convenient and applicable information for law administration and intelligence department. The goal of crime data mining is to understand patterns in criminal behavior in order to predict crime anticipate criminal activity and stop it. This paper describes web data mining which includes structure mining, web content mining, web usage mining and crime data mining. The occurrences of criminal activities based on web data mining process is also presented in this paper. The presented information on different criminal activities can be used to reduce further occurrences of similar incidence and to stop the crime.*

KEYWORDS: Crime data, Web Mining, Clustering, Pattern Analysis, Classification, Crime Control.

INTRODUCTION

The beginning of the World Wide Web has prompted an outstanding increase in the usage of the Internet. The resulting growth in on-line information combined with the almost unstructured web data necessitates the development of efficient web data mining process (Kambayashi, Mohania et al. 2003). The overall goal of the web data mining process is to extract knowledge from different types of web data set in a coherent structure.

Web data mining focus on three issues; web structure mining, web content mining, and web usage mining. Web structure mining involves mining the web document's structures and links. Web content mining is the process of information discovery from sources across the World Wide Web. Web usage mining is the process of finding out what users are looking for on the Internet; some users might be looking at only textual data, whereas some others might be interested in multimedia data or other types of data. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better help the needs of Web-based applications (Cooley, Mobasher et al. 1997). Usage data captures the identity or source of Web users along with their browsing behaviour at a Web site.

Web usage mining is gaining importance due to its vast applicability. It is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. In the public sector, web data mining applications were initially used as a means to detect fraud and waste, but have grown to also be used for purposes such as measuring and improving program performance. Web usage mining is relatively independent, but not isolated, category, which mainly describes the techniques that discover the user's usage pattern and try to predict the user's behaviors.

Web usage mining in the study and analysis of criminology becomes very challenging as criminals are becoming technologically stylish in committing crime using web communication. Therefore major challenge faced by intelligence and law enforcement agencies is the complications in analysing large volume of data involved in crime and extremist activities. Crimes are a social annoyance and charge our society extremely in several ways obstructing the peace within the nation.

Data mining in the study and analysis of criminology can be categorized into main areas, crime control and crime suppression. As data mining is the appropriate field to apply on high volume crime dataset and knowledge gained from data mining approaches will be useful and support police force. The goal of crime data mining is to understand patterns in criminal behavior in order to predict crime, anticipate criminal activity and prevent it. Among a large set of crimes that happen every year in a major city, it is challenging, time-consuming, and labor-intensive for crime analysts to determine which ones may have been committed by the same individual.

This paper describes the issues on identifying criminal activities through web data mining. The main objective is the extraction of crime patterns by analysis of web usage data, prediction of crime and anticipation of crime rate using different data mining techniques. We have used criminal datasets provided by Australian bureau of statistics, in order to analyse crime patterns and to provide more information to reduce the crime. The remainder of this paper is organized as follows: related work is described in section 2, web data mining structure is presented in section 3. Crime data mining techniques is presented in section 4. Section 5 presents experimental results. The paper concludes with a discussion and final remarks in section 6.

RELATED WORK

Data mining, which is referred to as knowledge discovery in databases. Researchers in many different fields, including database systems, knowledge-base system, artificial intelligence, knowledge acquisition, statistics, spatial databases, and data visualization, have shown great interest in data mining. Furthermore, server emerging applications in information providing services, such as on-line services and World Wide Web, also call for various data mining techniques to understand user behavior and to increase the business opportunities. In response to such demand, data mining with respect to database perspective is described by (Chen, Han et al. 1996) represents the survey of several data mining techniques developed in several research communities and those implemented in applicative data mining systems.

Web mining an application of data mining provides an overall view of web mining especially web usage mining consists of three phases: Pre-processing, Pattern discovery and Pattern analysis, is presented by (Dalal, Kumar et al. 2014). They have described the user navigation patterns discovery and analysis. They have also explained a mining system called 'Web SIFT' to make it easier to understand the methodology of how to apply data mining techniques to large Web data repositories in order to extract usage patterns.

Nazlena Mohamad Ali et al. describes the method referred as Visual Interactive Malaysia Crime News Retrieval System; their key purposes are to build crime-based event; explore the use of crime based event in improving the classification and clustering; develop an interactive crime news retrieval system; visualize crime news in an effective and interactive way; integrate them into a usable and robust system and evaluate the usability and system performance to address crime data and the eventual goal of combating the crimes (Ali, Mohd et al. 2011). Web structure mining and web content mining is described in (Ko, Yueh-Ming et al. 1998).

A system for Web usage mining called WEBMINER is presented by (Cooley, Mobasher et al. 1997). They have described application of data mining techniques to the World Wide Web, referred to as Web mining, has been used in two distinct ways. The first, called Web content mining, is the process of information discovery from sources across the World Wide Web. The second, called Web usage mining, is the process of mining for user browsing and access patterns. They have also define web mining and present an overview of the various research issues, techniques, and development efforts.

An intelligent analysis of web crime data using data mining as described in (Sharma and Sharma 2012) represents a clustering or classification based model to anticipate crime trends. This model elaborates the consequence to extract the attributes and relations in the web pages and reconstruct the situation for crime mining. Crime analysis tool for Indian scenario using different data mining techniques that help law enforcement department to efficiently handle crime investigation is presented by (Malathi and Baboo 2011). The proposed method describes and analyzes crime data to identify actionable patterns and trends very accurately. A framework for the crime and criminal data analysis and detection using decision tree algorithms is described by (Agarwal, Nagpal et al. 2013). This framework tends to help specialists in discovering patterns and trends, making forecasts, finding relationships and possible explanations, mapping criminal networks and identifying possible suspects.

A clustering or classify based model to anticipate crime trends is described in (Malathi and Baboo 2011). The presented data mining techniques are used to analyze the city crime data from Police Department. The results of this data mining technique were potentially used to diminish and even stop crime. The possibility of using clustering technology for auditing automated fraud filtering has been addressed by (Thiprungsri and Vasarhelyi 2011). The purpose of that filtering is to inspect the use of clustering technology to automate fraud purifying during an audit. They have used cluster analysis to help auditors focus their efforts when evaluating group life insurance claims. A large set of crimes that occur every year in a major city therefore it is challenging, time-consuming, and labor intensive for crime analysts to determine which ones may have been committed by the same individual is described in (Wang, Rudin et al. 2013). They provide a pattern detection algorithm that discovers crimes within a database. In criminal justice systems, incompatible content and information formats often create barriers to data access and utilization that make knowledge management a complex and daunting process. To address these problems, (Hauck, Atabakhsh et al. 2002) have described in the Coplink project, which bridges the gap between conducting technology research and helping police officers fight crime. Several user studies have shown that this system also increases searching and browsing in the engineering and biomedicine domains.

Sukanaya M et al. have presented the approach to find the hotspot of the criminal activities and the criminals by using clustering and classification algorithms in (M, T et al. 2012). This paper also elaborates the identification of hotspot of criminals activities which help the police department to provide more security to the particular area to prevent crimes in future. The procedure of extracting knowledge and information from large set of data that applying artificial intelligence method to find unseen relationships of data is described by (Hosseinkhani, Koochakzaei et al. 2014). (Senator, Goldberg et al. 1995) elaborates the Financial Crimes Enforcement Network (FIN-CEN) AI system (FAIS) links and evaluates reports of large cash transactions to identify potential money laundering. The objective of FAIS is to discover previously unknown, potentially high-value leads for possible investigation. FAIS integrates intelligent human and software agents in a cooperative discovery task on a very large data space. It is a complex system incorporating several aspects of AI technology, including rule-based reasoning and a blackboard.

The techniques that can automatically detect identity deception are described by (Wang, Chen et al. 2004). Most of the existing techniques are experimental and cannot be easily applied to real applications because of problems such as missing values and large data size. The authors propose an adaptive detection algorithm that adapts well to incomplete identities with missing values and to large datasets containing millions of records. The approach can identify deception having incomplete identities with high precision. In addition, it demonstrates excellent efficiency and scalability for large databases. A data mining framework for adaptively building Intrusion Detection (ID) models is described by (Lee, Stolfo et al. 1999). The central idea is to utilize auditing programs to extract an extensive set of features that describe each network connection or host session, and apply data mining programs to learn rules that accurately capture the behavior of intrusions and normal activities. These rules can then be used for misuse detection and anomaly detection. (Vel, Anderson et al. 2001) describe an investigation into e-mail content mining for author identification, or authorship attribution, for the purpose of forensic investigation. Their focus of discussion is on the ability to discriminate between authors for the case of both aggregated e-mail topics as well as across different e-mail topics. An extended set of e-mail document features including structural characteristics and linguistic patterns were derived and, together with a Support Vector Machine learning algorithm, were used for mining the e-mail content. Experiments using a number of e-mail documents generated by different authors on a set of topics gave promising results for both aggregated and multi-topic author categorization.

Valuable criminal-justice data in free texts such as police narrative reports are currently difficult to be accessed and used by intelligence investigators in crime analyses. It would be desirable to automatically identify from text reports meaningful entities, such as person names, addresses, narcotic drugs, or vehicle names to facilitate crime investigation. (Chau, Xu et al. 2002) have described a neural network-based entity extractor, which applies named-entity extraction techniques to identify useful entities from police narrative reports. Preliminary evaluation results demonstrated that their approach is feasible and has some potential values for real-life applications. Their system achieved encouraging precision and recall rates for person names and narcotic drugs, but did not perform well for addresses and personal properties.

WEB DATA MINING STRUCTURE

Web data mining consists of web content mining, web structure mining, and web usage mining. Each section is described in the following:

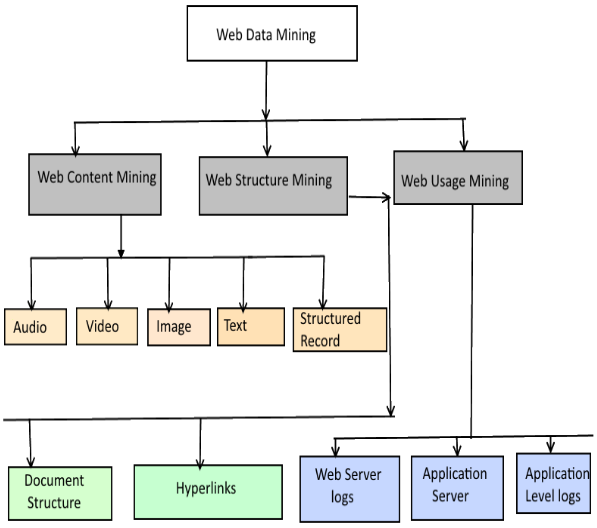

Figure 1. Web data mining system structure.

Web Content Mining

Web content mining is the mining, extraction and integration of useful data, information and knowledge from web page content. It may consist of text, images, audio, video, or structured records such as lists and tables. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages.

Web content mining is distinguished from two different points of view: information retrieval view and database view (Wikipedia 2009). The research works done from information retrieval view is for unstructured data and semi-structured data. It shows that most of the researches use bag of words, which is based on the statistics about single words in isolation, to represent unstructured text and take single word found in the training corpus as features. For the semi-structured data, all the works utilize the HTML structures inside the documents and some utilized the hyperlink structure between the documents for document representation. As for the database view, in order to have the better information management and querying on the web, the mining always tries to infer the structure of the web site to transform a web site to become a database.

Web Structure Mining

Web structure mining is the process of determining structure information from the web. In another words, web structure mining is the process of using graph theory to analyze the node and connection structure of a web site (Wikipedia 2009). Web structure mining terminologies:

Web graph: directed graph representing web.

Node: web page in graph.

Edge: hyperlinks.

In degree: number of links pointing to particular node.

Out degree: Number of links generated from particular node.

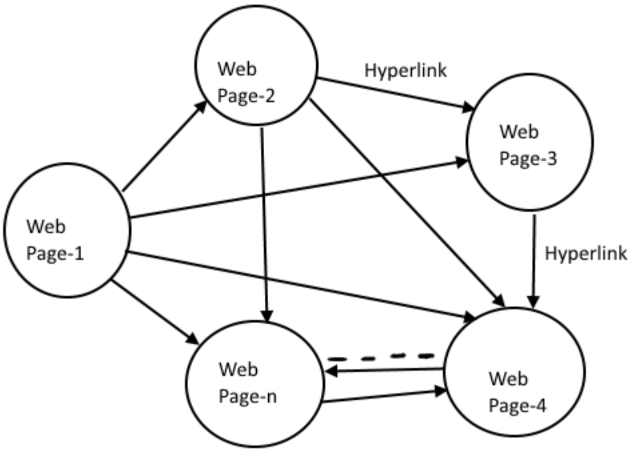

Fig. 2: Web structure mining as a graph.

Web structure mining can be further divided into two kinds based on the kind of structure information used.

Hyperlinks

A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an intra-document hyperlink, and a hyperlink that connects two different pages is called an inter-document hyperlink.

Document Structure

In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts have focused on automatically extracting document object model (DOM) structures out of documents.

Web Usage Mining

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered:

Web Server Data

The user logs are collected by the Web server. Typical data includes IP address, page reference and access time.

Application Server Data

Commercial application servers have significant features to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

Application Level Data

New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the categories above.

CRIME DATA MINING STRUCTURE

Data Analysis

Crime data mining is defined as analytical processes which provide relevant information relative to crime patterns and trend associations to assist personnel in planning the deployment of resources for the prevention and suppression of criminal activities.

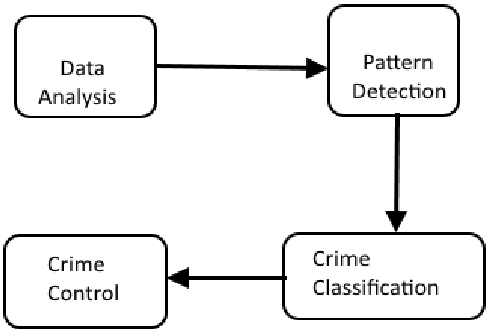

Figure 3: Crime data mining structure.

Pattern Detection

Analyze crime to inform law enforcers about general and specific crime trends in timely manner. Link analysis is a good start in mapping terrorist activity and criminal intelligence by visualizing associations between entities and events. Link analyses often involve seeing via a chart or a map the associations between suspects and locations whether by physical contacts, communications in a network, thru phone calls, financial transactions, or via the Internet and e-mail. Criminal investigators often use link analysis to begin to answer such questions as “who know whom and when and where have they been in contact?”

Intelligence analysts and criminal investigators must often correlate enormous amounts of data about individuals in fraudulent, political, terrorist, narcotics and other criminal organizations. A critical first step in the mining of this data is viewing it in terms of relationships between people and organizations under investigation. One of the first tasks in data mining and criminal detection involves the visualization of these associations, which commonly involves the use of link analysis charts. Extraction of crime patterns by analysis of available crime and criminal data.

Crime Classification

Crime place, crime types and crime time.

Crime Control and Suppression

Traditional data mining techniques such as association analysis, classification and prediction, cluster analysis, and outlier analysis identify patterns in structured data (Han 2005).

Experimental Results

Many open source data mining suites are available such as R, Tanagra, Weka, KNIME, Orange, Rapid miner. We have used data mining software called 'Orange' for data analysis. Orange is used to compare the predicted data patterns for the accuracy of the results from the given data sets. Orange is a component-based visual programming software for data mining, machine learning and data analysis. The datasets have been used in the experiments is as CSV format.

Crime dataset used for crime analysis is an offence presented by Australian Bureau of Statistics from 2008 to 2014 (Statistics 2013-14) in Australia. The datasets we have used in our experimental analysis, sample of these datasets are presented in Table 1. In Table 1, first column presents the year when the crime occurred. Crime pattern and the area (where the crime occurred) are presented in second and third columns respectively. Fourth and fifth column describes the total occurrence and the percentage of the crime, based on the population on that area. The datasets have been used in the experiments as CSV format.

From figure 4, it is clearly shown that the growth of crime is most on the year 2012-2013 and least on the year 2008-2009. The growth of the distribution of the crime based on the type of offence is presented on the figure 5. Figure 5, shows the most dominating crime is the Sexual Assault. Figure 6, presents the variation of the crime committed based on the region. The dissemination of the crime towards the different area is presented on the figure 7. Figure 8 presents the distribution of the crime to different years.

Table I. Sample of Crime Data Sets

Year	Offence Type	Region	Total	Rate (%)
2008-09	Physical Assault	NSW	153800	2.8
2008-09	Physical Assault	Victoria	142500	3.3
2008-09	Physical Assault	Queensland	100700	3
2008-09	Physical Assault	South Australia	35500	2.8
2008-09	Physical Assault	Western Australia	64500	3.8
2008-09	Physical Assault	Tasmania	16000	4.1

2008-09	Physical Assault	Northern Territory	7100	5.7
2008-09	Physical Assault	ACT	7500	2.8
2008-09	Face-To-Face-Threatened Assault	NSW	203400	3.7
2008-09	Non-Face-To-Face-Threatened Assault	NSW	51800	2.2
2013-14	Non-Face-To-Face-Threatened Assault	ACT	3900	1.3

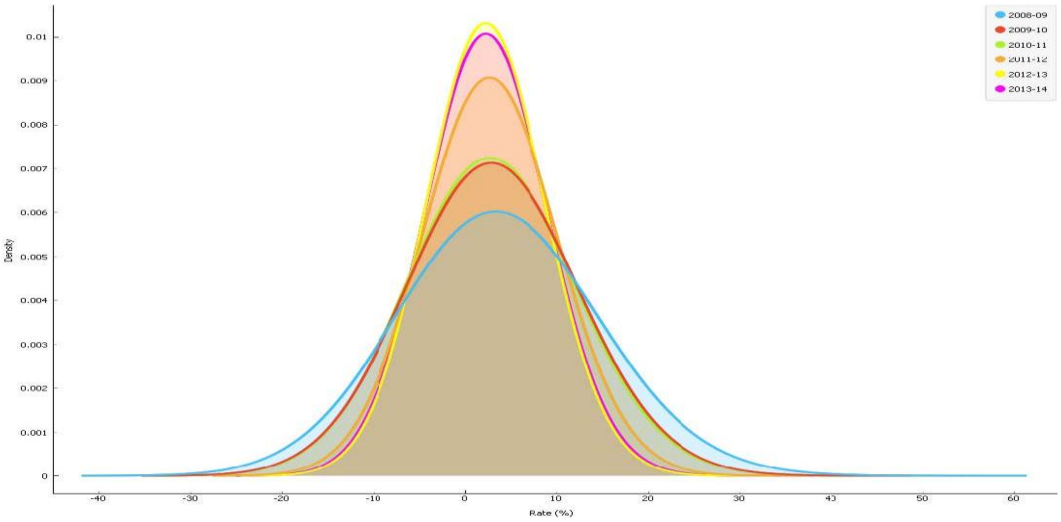

Figure 4. The growth of crime based on year.

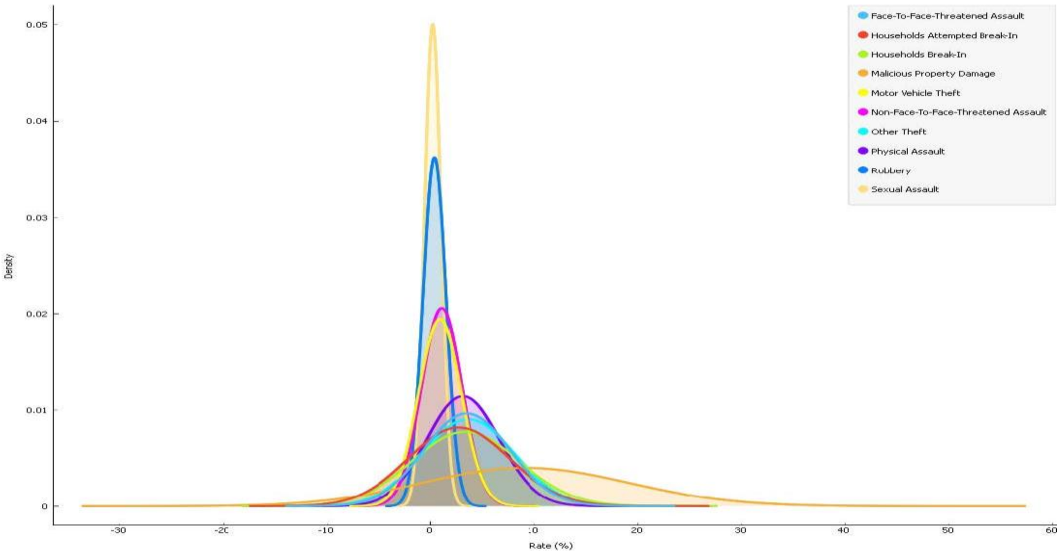

Figure 5. The frequency of crime.

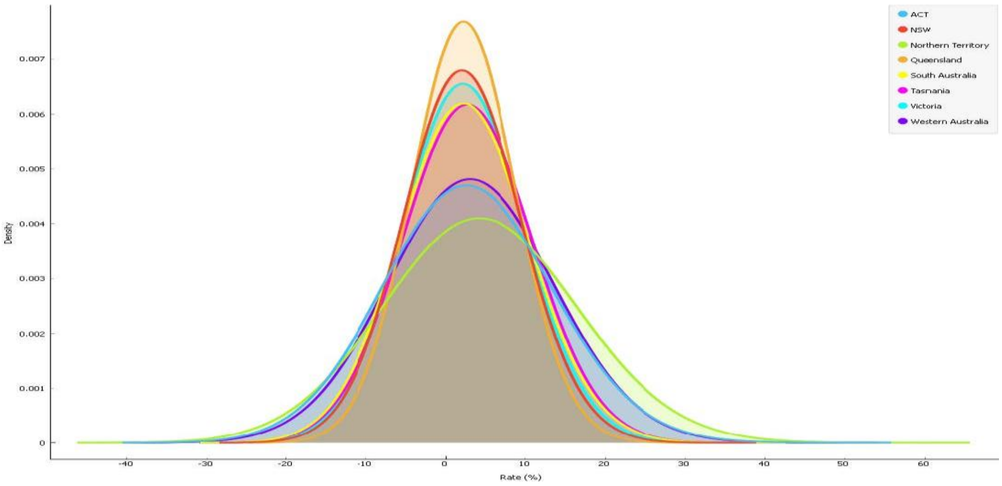

Figure 6. The growth of crime based on region.

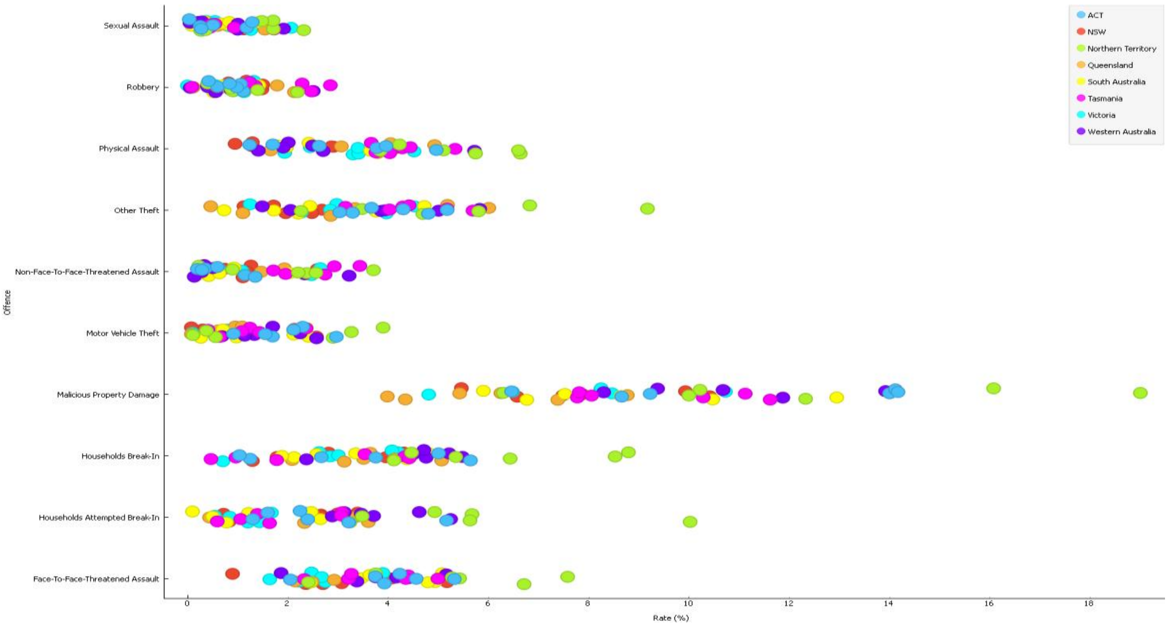

Figure 7. The distribution of crime to the corresponding region.

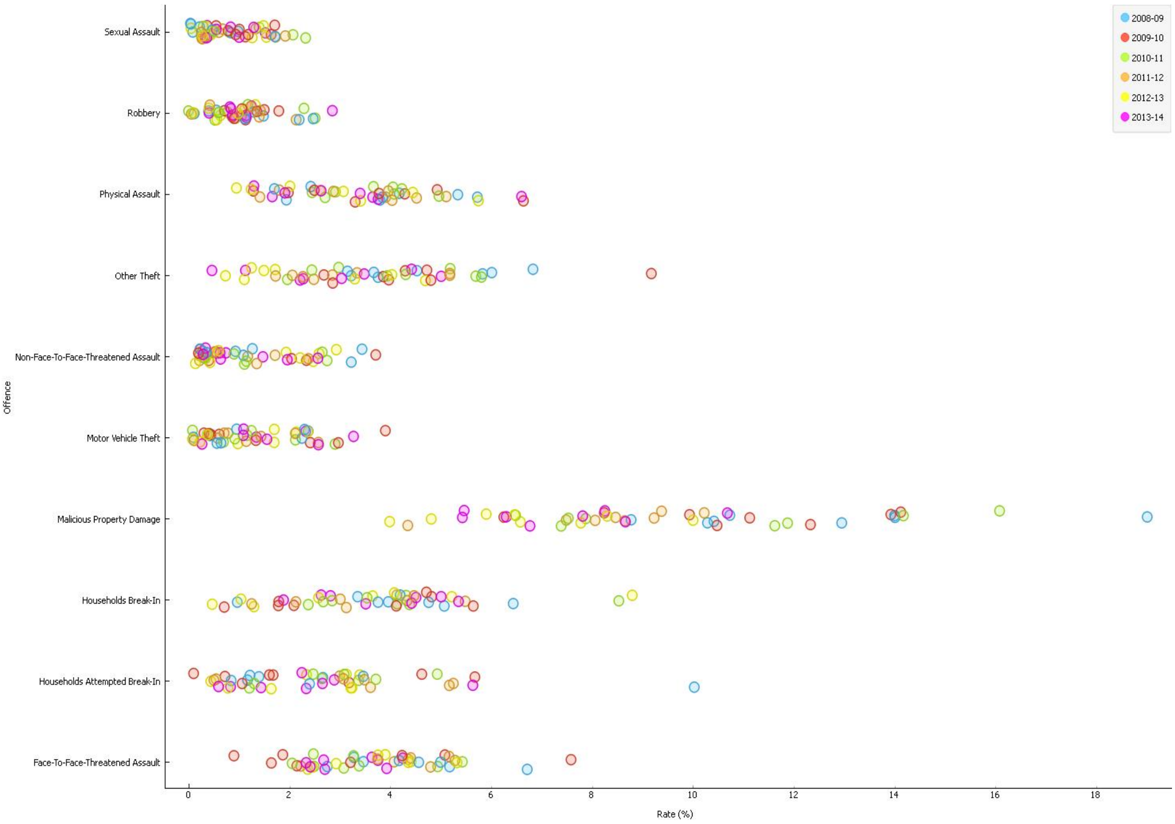

Figure 8. The distribution of crime to the corresponding year.

CONCLUSIONS

As the web and its usage continues to grow, therefore it is a great challenge to analyze web data and extract all manner of useful knowledge from it. A key challenge facing all law-enforcement and intelligence gathering organizations is accurately and efficiently analyzing the growing volumes of crime data. As information science and technology progress, sophisticated data mining and artificial intelligence tools are increasingly accessible to the law enforcement community. The past ten years have seen the emergence of web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it.

Web data mining including content structure and crime data mining is describe in this paper. We have also presented the web based criminal information through web data mining process to reduce further occurrences of similar incidence and to stop the crime. Web data mining process has been used to find the regions where the most criminal activities are occurred and the types of the criminal activities. This information is useful to make laws or regulations to diminish criminal activities in the society. The information can be used to stop same occurrence on the same spot in future.

Many future directions can be explored being as a very attractive young research domain. For example, more visual and intuitive criminal and intelligence investigation techniques can be developed for crime pattern and network visualization.

REFERENCES

Agarwal, J., R. Nagpal and R. Sehgal (2013). "Crime Analysis using K-Means Clustering." International Journal of Computer Applications **83**(4): 1-4.

Ali, N. M., M. Mohd, H. Lee, A. F. Smeaton, F. Crestani, S. Azman and M. Noah (2011). i-JEN: Visual Interactive Malaysia Crime News Retrieval System. Visual Informatics: Sustaining Research and Innovations, Springer Berlin Heidelberg**:** 284-294.

Chau, M., J. J. Xu and H. Chen (2002). Extracting Meaningful Entities from Police Narrative Reports. The National Conference for Digital Government Research.

Chen, M. S., J. Han and P. S. Yu (1996). "Data mining: An overview from database perspective." IEEE transaction on Knowledge and Data Engineering **8**(6): 866 - 883.

Cooley, R., B. Mobasher and J. Srivastava (1997). Web Mining: Information and Pattern Discovery on the World Wide Web

IEEE International Conference on Tools with Artificial Intelligence, Newport Beach, CA, IEEE.

Dalal, S., S. Kumar and V. Dixit (2014). "Web mining an application of data mining." International Journal of Computer Science and Information Technology Research **2**(3): 455-460.

Hauck, R. V., H. Atabakhsh, P. Ongvasith, H. Gupta and H. Chen (2002). "Using Coplink to analyze criminal-justice data." IEEE Computer **35**(3): 30-37.

Hosseinkhani, J., M. Koochakzaei, S. Keikhaee and J. H. Naniz (2014). "Detecting Suspicion Information on the Web Using Crime Data Mining Techniques." International Journal of Advanced Computer Science and Information Technology **3**(1): 32-41.

Kambayashi, Y., M. Mohania and W. Wob (2003). Research Issuses in Web Data Mining. Data Warehousing and Knowledge Discovery. Prague, Czech Republic, Springer. **5:** 303-312.

Ko, M.-T., Yueh-Ming and Huang (1998). Extracting Classification Knowledge of Internet Documents with Mining

Term Associations: A Semantic Approach SIGIR'98,. Melbourne, Australia, ACM**:** 241-249.

Lee, W., S. J. Stolfo and K. W. Mok (1999). A Data Mining Framework for Building Intrusion Detection Models. IEEE Symposium on Security and Privacy. Oakland, CA**:** 120-132.

M, S., K. T and K. S (2012). "Criminals and crime hotspot detection using data mining algorithms: clustering and classification." International Journal of Advanced Research in Computer Engineering & Technology **1**(10): 225-227.

Malathi, A. and S. Baboo (2011). "Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters " Global Journal of Computer Science and Technology **11**(11): 1-7.

Malathi, A. and S. Baboo (2011). "An Enhanced Algorithm to Predict a Future Crime using Data Mining." International Journal of Computer Applications **21**(1): 1-6.

Senator, T. E., H. G. Goldberg, J. Wooton, M. A. Cottini, A. F. U. Khan, C. D. Klinger, W. M. Llamas, M. P. Marrone and R. W. H. Wong (1995). "FinCEN Artificial Intelligence System: Identifying Potential Money Laundering From Reports of Large Cash Transactions " Applied Technology for Improved Operational Effectiveness International Technology **16**(4): 21-39.

Sharma, A. and S. Sharma (2012). "An Intelligent Analysis of Web Crime Data Using Data Mining." International Journal of Engineering and Innovative Technology **2**(3): 203-206.

Statistics, A. B. o. (2013-14). http://www.abs.gov.au/ausstats/abs@.nsf/mf/4530.0. Crime Victimisation, Australia.

Thiprungsri, S. and M. A. Vasarhelyi (2011). "Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach." The International Journal of Digital Accounting Research **11**: 69-84.

Vel, O. d., A. Anderson, M. Corney and G. Mohay (2001). "Mining E-Mail Content for Author Identification Forensics." ACM SIGMOD **30**: 55-64.

Wang, G., H. Chen and H. Atabakhsh (2004). "Automatically detecting deceptive criminal identities." Communications of the ACM **47**(3): 70-76.

Wang, T., C. Rudin, D. Wagner and R. Sevieri (2013). Learning to Detect Patterns of Crime. Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg**:** 515-530.