An Intelligent Analysis of Crime through Newspaper Articles – Clustering and Classification

Ashwini B. Bais[1]
Computer Science and Engg.
TGPCET, Mohgao Nagpur, India

Prof. G. Rajesh Babu[2], Prof. Jayant Adhikari[3]
Faculty of t Computer Science and Engg.
TGPCET,Mohgao Nagpur, India

Abstract— Crime analysis is one of the most important activities of the majority of the intelligent and law enforcement organizations all over the world. Thus, it seems necessary to study reasons, factors and relations between occurrence of different crimes and finding the most appropriate ways to control and avoid more crimes. A major challenge faced by most of the law enforcement and intelligence organizations is efficiently and accurately analyzing the growing volumes of crime related data. The vast geographical diversity and the complexity of crime patterns have made the analyzing and recording of crime data more difficult. This paper presents an intelligent crime analysis system which is designed to overcome the above mentioned problems. Data mining is used extensively in terms of analysis, investigation and discovery of patterns for occurrence of different crimes. The proposed system is a web-based system which performs crime analysis through news articles. In this paper we use a clustering/ classification based model to automatically group the retrieved documents into a list of meaningful categories. The data mining techniques are used to analyze the web data.

Keywords: *clustering, data mining, law-enforcement, Classification*

*****

I. INTRODUCTION

Crime is one of the dangerous factors for any country. Crime analysis is the activity in which analysis is done on crime activities. Today criminals have maximum use of all modern technologies and hi-tech methods in committing crimes. It is impossible to find a country which has a crime-free society. As long as human beings have feelings they incline on attempting crimes. So the present society has also filled with various kinds of crimes. Hence, creation of data base for crimes and criminals is needed. Developing a good crime analysis tool to identify crime patterns quickly and efficiently for future crime pattern detection is challenging field for researchers.

Crime analysis has become one of the most vital activities of the modern world due to the high magnitude of crimes which is a result of technological advancements and the population growth. Law enforcement organizations and the intelligence gathering organizations all around the world usually collect large amounts of domestic and foreign crime data (intelligence) to prevent future attacks. As this involves a large amount of data, manual techniques of analyzing such data with a vast variation have resulted in lower productivity and ineffective utilization of manpower.

Hence, creation of data base for crimes and criminals is needed. Developing a good crime analysis tool to identify crime patterns quickly and efficiently for future crime pattern detection is challenging field for researchers. Data mining techniques have higher influence in the fields. There are various crime data mining techniques available [8] such as clustering techniques, association rule mining, sequential pattern mining, and classification and string comparison. Several web based crime mapping systems are available on the Internet such as narcotics network in Tucson police department, but majority of them have been custom made for legislative authorities in different countries and those systems are not accessible to parties outside that particular law enforcement or legislative authorities [9] [10]. This paper presents a web based crime analysis system. Sri Lankan English newspapers (Daily Mirror, The Island, and Ceylon Today) are used as the source for details of crime incidents. Newspaper articles are crawled using a focused crawler and they are classified using a SVM based classifier. Required entities are extracted from classified crime articles and duplicate detection is performed. By using preprocessed data, crime analysis operations are performed and results are displayed using web based GUI. Unlike most systems, this system is open to anyone who is interested in crime analysis. When newspapers are considered, they contain articles only for a subset of total crime population.

Most of the time the police and other interested parties are more concerned about major crime incidents rather than minor crime incidents when taking decisions. Therefore crime analysis results based on newspaper articles will be useful to interested parties (police, researchers, investors and tourists) as means of assistance for their respective tasks even though newspapers cannot reveal the exact number of crimes. The proposed system cannot be directly validated using records of the police department because police records include both major and minor crime incidents. The proposed system is based on newspaper articles so it includes only a subset of total crime incidents. So individual components of the proposed system are evaluated and results of that evaluation are used to measure the effectiveness of proposed system.

There are several significant reasons for crime analysis such as to identify general and specific crime trends, patterns, and series in an ongoing, timely manner, to maximize the usage of limited law enforcement resources, to access crime problems locally, regionally, nationally within and between law enforcement agencies, to be proactive in detecting and preventing crimes and to meet the law enforcement needs of the changing society. There are various crime data mining techniques available [8] such as clustering techniques, association rule mining, sequential pattern mining, and classification and string comparison.

II. EXSISTING SYSTEM

There are several existing systems which use crime data mining techniques for crime analysis such as, regional crime analysis program , data mining framework for crime pattern identification [12] and narcotics network in Tucson police department [8]. In [7] a collection of criminal analysis steps are given. Among them, steps such as hotspot detection, crime comparison, crime pattern visualization are significant. In crime pattern visualization, a time series can be drawn between the crime frequency and the time and using it interesting crime trends can be identified. In addition to these steps, [7] has given some other analysis steps such as crime clock, outbreaks detection and nearest police station detection. Using the above techniques, crime data can be analyzed more effectively and efficiently and law enforcement organization and other interested parties will be able to get more accurate decisions based on them. An intelligent crime identification system is described in [11] which can be used to predict possible suspects for given crime. They have used five types of agents namely, message space agent, gateway agent, prisoner agent, criminal agent and evidence agent.

III. LITERATURE SURVEY

Kaumalee Bogahawatte and Shalinda Adikari (2013) proposed the criminal identification system for identify the criminal (ICIS) This paper highlights the use of Clustering and classification for effective investigation of crimes. The system uses an explicit clustering mechanism on the available evidences. Naïve Bayesian classification has used to identify most possible suspect/ suspects for crime incidents which used the explicit clustering that can potentially identify a criminal based on the evidences collected from the crime spot. The solution has provided for three crime categories namely robbery, burglary and theft out of 21 categories of grave crimes [1].

Luís Filipe da Cruz Nassif and Eduardo Raul Hruschka (2013) In this paper present an approach that applies document clustering algorithm's to forensic analysis of computerized in police investigations. They illustrated the well known six algorithms for document clustering i.e.(K-means, K-medoids, Single Link, Complete Link, Average Link, and CSPA) [2] applied to five Real-world datasets obtained from computers seized in real-world investigations and they performed some experiment with different combination of parameter for relevant result. By applying so many algorithm the scalability may be an issue.

Qusay Bsoul ,Juhana Salim, Lailatul Qadri Zakaria (2013) The author take the data from all type of criminal news ,stories Divers dataset and other resource. The aim of this paper to automatically group together similar document in one cluster using different type of extraction and clustering algorithm. The author used k-means, k-medians and k-means++ and hierarchical clustering algorithm. They developed the new technique called Lemmatization algorithm. This algorithm used for catching the important word from the two lists of prepositions first list includes proceeding verb and other nouns [3]. But the author not developed the decision making tree and there is not a concept of outlier detection.

Jyoti Agarwal,Renuka Nagpal,Rajni Sehgal (2013), In this paper analysis is done by performing K-means clustering algorithm on crime data set using rapid miner tool they do crime analysis by considering crime homicide and plotting it with respect to year and got into conclusion that homicide is decreasing from 1990 to 2011[4]. From that clustered results it is easy to identify crime trend over years and can be used to design precaution methods for future. They provide the crime trend over year not the criminal and not specified the rule for identifying the criminal. Where the Open rapid miner tool used for reading the criminal excels sheet of crime.

In [5] an improved method of classification algorithms for crime prediction has proposed by A. Babakura, N. Sulaiman and M. Yusuf. They have compared Naïve Bayesian and Back Propagation (BP) classification algorithms for predicting crime category for distinctive state in USA. In the first step phase, the model is built on the training and in the second phase the model is applied. The performance measurements such as Accuracy, Precision and Recall are used for comparing of the classification algorithms. The precision and recall remain the same when BP is used as a classifier.

In [6] researches have introduced crime analysis and prediction using data mining. They have proposed an approach between computer science and criminal justice to develop a data mining procedure that can help solve crimes faster. Also they have focused on causes of crime occurrence like criminal background of offender, political, enmity and crime factors of each day. Their method steps are data collection, classification, pattern identification, prediction and visualization.

IV. OBJECTIVE

The proposed system having following objectives:

A. *To implement web crawling to crawl news article*

Crawling news articles from given newspaper is perform using crawler. So that the required content of the crawled article is stored in database for future processing.

B. *To implement new approach for data preprocessing using NLP*

The main aim of Natural Language Processing (NLP) to convert the human language into a formal representation that easy for computer to manipulate. This is used for preprocessing the data.

C. *To implement document classification using SVM based classifier*

Document classification is defined as assigning predefined categories to text documents, where documents can be news stories, technical reports, web pages, etc., There are a lot of algorithms available for text classification and among them SVM (support vector machine) is used for document classification.

V. PROPOSED WORK

The proposed system consists of seven major components. They are as follows. High level architecture of the system including major components is given in Fig. 1.

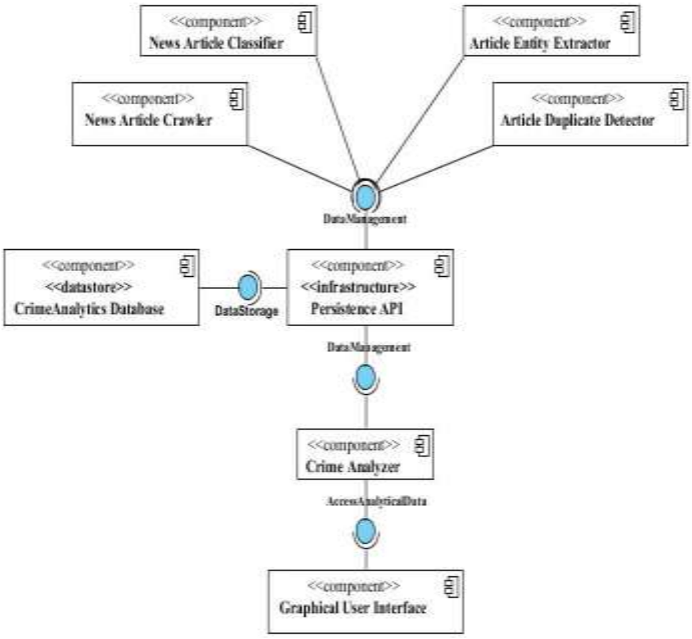

Fig. 1.□ High level architecture of the proposed system

A. Crawler

The main responsibility of the crawler is to crawl news articles of a given newspaper. Required content of the crawled articles is stored in the database for further processing. A focused crawler has been implemented by extending a generic crawler known as Crawler4j [14]. Jsoup [15] is a java library that can be used for extracting and manipulating HTML data.

B. Document Classifier

Responsibility of document classifier is to classify crawled newspaper articles as crime and non-crime articles. Classified articles will be stored in data base for entity extraction. Documents have to be processed before using them for the classification process. Weka library has been used for this purpose. Documents have been transformed to feature vectors while removing stop words from them using tf-idf transformation. Stemming/ lemmatization has been performed in order to reduce words into their base forms. LibSVM library has been used to implement the classifier.

C. Entity Extractor

This module is used to extract important entities from the classified newspaper articles. From each crime article, entities such as crime date, location, police, court, victim count etc. are extracted if possible. Several ancillary processes have been carried out to do the required preprocessing on the document corpus to prepare documents for named entity extraction GATE (General Architecture for Text Engineering) has been used for text processing as well as for entity extraction.

D. Duplicate Detector

The main purpose of this module is to identify exact/near duplicates of newspaper articles and remove them from the database. Newspaper articles have been represented using 64 bit simhash values. The entire contents of newspaper article have not been used to generate simhash values as noise may distort the simhash value. Therefore in order to generate the corresponding simhash value, extracted entities of each article have been used. Crime type, crime date and crime location is used to generate the representation of the document. To hash each feature, murmer hash implementation has been used.

E. Database Handler

All database transactions are handled using this module. This has been implemented using Hibernate framework .

F. Analyzer

Analyzer module will perform crime analysis operations on processed crime articles. It will perform Hot spot detection, Crime comparison, Crime pattern visualization .

G. Web based GUI

This module is used to visualize crime statistical details of the previous years. The client side of the GUI uses a JavaScript library called high charts for visualizing data with maps, graphs and pie charts. The server side of the web application has been written in JavaScript in order to interact with client side JavaScript library efficiently. The server side analyses data fetched from the database and converts them to a format for visualization. The web application is implemented as a single page web application (SPA).

VI. Natural Language Processing

NLP is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. We used form NLP

1) *Part of Speech Tagging :* A part of speech – particularly in more modern classifications, which often make more precise distinctions than the traditional scheme does – may also be called a word class, lexical class, or lexical category.

2) *Chunkinh :* Grouping the extracted information .

VII. Classification Of Crime Crime

Crime is defined as “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”. Crime is referred to as a comprehensive concept that is defined in both legal and nonlegal sense.

Classification of Crime

- *Traffic Violations:* Driving under the influence of alcohol, fatal / personal injury / property damage traffic accident, road rage.
- *Sex Crime:* Sexual offences .
- Fraud: Forgery and counterfeiting, frauds, embezzlement, identity deception.
- *Arson:* Arson on buildings Drug Offences: Narcotic drug offences (sales or possession).
- *Violent Crime:* Criminal Homicide, armed robbery, aggravated assault, other assaults.
- *Cyber Crime:* Internet frauds, illegal trading, network intrusion / hacking, virus spreading, hate crimes, cyber piracy, cyber pornography, cyber-terrorism, theft of confidential information.

VIII. Clustering

Clustering can be considered the most important unsupervised learning problem occurred in data mining for criminal document clustering; so, all other problem of clustering is deals with finding a structure in a collection of unlabeled data A loose definition of clustering could be “the process of organizing objects into groups whose members are similar to each other”. A cluster is defined as the collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

The proposed system uses adaptive Bisecting K-means Algorithm for finding K clusters and it is as follows:

Algorithm:

1. Initialize: randomly select k of the n data points as the medoids

2. Assignment step: Associate each data point to the closest medoid.

3. Update step: For each medoid m and each data point o associated to m swap m and o and compute the total cost of the configuration (that is, the average dissimilarity of o to all the data points associated to m). Select the medoid o with the lowest cost of the configuration.

4. Pick a cluster to split

5. Find 2 sub-clusters using the basic k-Means algorithm (Bisecting step)

6. Repeat step 2, the bisecting step, for iterative times and take the split that produces theclustering with the highest overall similarity.

7. Repeat steps 4, 5 and 6 until the desired number of clusters is reached.

The critical part is which cluster to choose for splitting. And there are different ways to proceed. Bisecting k-means clustering algorithm can be used to partition the item from the e-commerce site as per the client interests and his previous searches. A partition clustering algorithm obtains a single partition of the data instead of a clustering method, such as the dendrogram produced by a hierarchical technique.

IX. Conclusion

Crime data is a sensitive domain where efficient clustering techniques play vital role for crime analysts and law-enforcers to precede the case in the investigation and help solving unsolved crimes faster. Similarity measures are an important factor which helps to find unsolved crimes in crime pattern. Partition clustering algorithm is one of the best method for finding similarity measures. This paper deals detailed study about importance of clustering and similarity measures in crime domain.This paper proposed a web based crime analysis system. The proposed system performs crime analysis operations such as hotspot detection, crime comparison and crime pattern visualization. Graphical user interface of the system uses graphs and diagrams to display the results which make crime analysis a very simple task. Then law enforcement officers and other interested users will be able to use this system effectively and efficiently for crime analysis processes. Also this is a public accessible system so that anyone who is interested in this area will be able to use this system.

REFERENCES

[1] Kaumalee Bogahawatte and Shalinda Adikari "Intelligent Criminal Identification System " The 8th International Conference on Computer science and Education (ICCSE 2013) April 26-28.Colombo ,shri Lanka

[2] Luís Filipe da Cruz Nassif and Eduardo Raul Hruschka "Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection" IEEE Transactions On Information Forensics And Security, Vol. 8, No. 1, JANUARY 2013

[3] Qusay Bsoul ,Juhana Salim, Lailatul Qadri Zakaria "An Intelligent Document Clustering Approach to Detect crime Patterns" The 4th International Conference on Electrical Engineering and Informatics (ICCSE 2013)

[4] Jyoti Agarwal,Renuka Nagpal,Rajni Sehgal "Crime Analysis using K-Means Clustering" International Journal of Computer Applications (0975 – 8887) Volume 83 – No4, December 2013

[5] Anshu Sharma, Raman Kumar "The obligatory of an Algorithm for Matching and Predicting Crime - Using Data Mining Techniques" IJCST Vol. 4, Issue 2, April - June 2013

[6] Cluster Analysis available at http://en.wikipedia.org/wiki/Cluster_analysis.

[7] P. Chamikara, D. Yapa, R. Kodituwakku and J. Gunathilake, "SLSecureNet : intelligent policing using data mining techniques," International Journal of Soft Computing and Engineering, vol. 2, no. 1, pp. 175-180, 2012.

[8] Chen, W. Chung, J. Xu, G. Wang , Y. Qin and M. Chau, "Crime data mining: a general framework and some examples," IEEE ExploreComputer, vol. 37, no. 4, pp. 50-56, 2004.

[9] Crime Mapping and Reporting System [Online]. Available: https://www.crimereports.com/

[10] Intelligent Mapping System [Online]. Available: http://maps.met.police.uk/

[11] S. Adhikari and K. Bogahawatte, "Intelligent criminal identification system," in The 8th International Conference on Computer Science & Education, Colombo, Sri Lanka, 2013, pp. 633-638.

[12] V. Nath, "Crime pattern detection using data mining," in Web Intelligence and Intelligent Agent Technology Workshops, Hong Kong, 2006, pp. 41-44.

[13] Miss. Aparna N. Gupta , Prof. ArtiKarndikar, "A Review : Study of Various clustering Technique in web usage mining" , International

Journal of Advance Research In Computer And Communication Engineering, Vol. 3 , Issue 3 , March 2014.

[14] Crawler4j [Online]. Available: https://code.google.com/p/crawler4j/

[15] JSoup [Online]. Available: http://jsoup.org/