International Journal of Advanced Research Methodology in Engineering & Technology,
Volume 1, Issue 2, March 2017, ISBN 978-1-63535-889-6

Review on analysis of crime by extracting crime information from newspaper articles

Ashwini B. Bais[1]
Computer Science and Engg.
[1]Tulsiramji Gaikwad-Patil College of Engineering &Technology,
Mohgao Nagpur, India
ash.abhi1998@gmail.com

Jayant Adhikari[2]
Computer Science and Engg.
[2]Tulsiramji Gaikwad-Patil College of Engineering &Technology,
Mohgao Nagpur, India
adhikari.jayant@gmail.com

***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 and 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—component; formatting; style; styling; insert (key words)

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. Data mining techniques have higher influence in the fields. There are various crime data mining techniques available [2] 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 [3] [4]. 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.

II. PROBLEM DEFINITION

To address this problem, we propose the system that will improve the efficiency and reduce the delay to identifying the criminal by improving the data mining techniques. By providing the combine approach of rule engine and outlier detection.

Crime domain is very sophisticated so, proper input data preprocessing and document clustering is very important. So many data mining techniques are available but in or proposed work included the NLP for extracting the action word.

III. EXSISTING SYSTEM

There are several existing systems which use crime data mining techniques for crime analysis such as, regional crime analysis program [9], data mining framework for crime pattern identification [7] and narcotics network in Tucson police department [2]. In [1] 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, [1] 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 [6] 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.

IV. 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.

V. RELATED WORK

Crime data mining is the application of data mining techniques for crime analysis [11]. Various researches have been carried out in this domain and few of them are given in [7], [8], [12] and [13]. Crimes can be divided into subcategories based on different criteria. In [13] eight crime categories are given. They are traffic violations, sex crimes, theft, fraud, arson, drug offenses, cybercrimes and violent crimes. They have given definitions for each category in local law enforcement level and national law enforcement level. IPTC [14] (international press telecommunication council) too has given a different categorization where crimes are divided into war crime, corporate crime, organized crime etc. There

are various crime data mining techniques available [8]. The most commonly used methods are, entity extraction, clustering techniques, association rule mining, sequential pattern mining and classification. There were many efforts to analyze different types of crimes using automated techniques but there is no unified framework describing how to apply those techniques to different crime types. In [8], they have proposed a framework which includes a relationship between the crime data mining technique and crime type characteristics.

VI. Preliminaries

A. Web Crawling

Web crawler is an Internet bot. It systematically browses the World Wide Web typically for the purpose of web indexing [15]. Crawlers can be selective about the pages they fetch (Ex- crawl only the pages of selected newspaper sites) and are then referred to as preferential or heuristic-based crawlers [16]. Preferential crawlers built to retrieve pages within a certain topic are called topical or focused crawlers.

B. Document Classification

Document classification is defined as assigning predefined categories to text documents, where documents can be news stories, technical reports, web pages, etc., and categories are most often subjects or topics [17]. To perform an effective document classification, syntactic (arrangement of words) and semantic (meaning of words) aspects of the natural language have to be addressed [18]. There are a lot of algorithms available for text classification and among them SVM (support vector machine) is the best [19].

C. Entity Extraction

This is the main step of transforming the unstructured data into the structured format. A named entity is typically a name of a person, a place, an organization or a date. Extraction of named entities involves identification of small chunks of appropriate texts and classification of them into one of such predefined categories of interest [20].

D. Duplicate Detection

Duplicates (i.e. documents that are exact duplicates of each other due to mirroring and plagiarism) are easy to identify by standard check summing techniques. A more difficult problem is the identification of near-duplicate documents. Two such documents are identical in terms of content but differ in a small portion of the document [21].

VII. 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.

VIII. 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.

IX. 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] R. Krishnamurthy and S. Kumar, "Survey of data mining techniques on crime data analysis," International Journal of Data Mining Techniques and Applications, vol. 1, no. 2, pp. 117-120, December 2012.

[12] 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.

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

[14] International Press Telecommunications Council [Online]. Available: http://www.iptc.org/site/Home/

[15] Web Crawler [Online]. Available: http://www.sciencedaily.com /articles/w/web_crawler.htm

[16] G. Pant, I Srinivasan and F. Menczer, "Crawling the Web," in Web Dynamics.: Springer Science & Business Media, 2004, pp. 153-178.

[17] W. Zhang, T. Yoshida and X. Tang, "Text classification based on multiword with support vector machine," Knowledge-Based Systems, vol. 21, no. 8, pp. 879-886, 2008.

[18] K. Cheng, S. Pan and F. Kurfess, "Ontology-based semantic classification of unstructured documents," Adaptive Multimedia Retrieval in Computer Science, vol. 3094, pp. 120-131, 2004.

[19] K. Mertsalov and M. McCreary, "Document classification with support vector machines," 2009.

[20] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," Linguisticae Investigationes, 2007.

[21] S. Charikar, "Similarity estimation techniques from rounding algorithms," in ACM symposium on Theory of computing, New York, 2002, pp. 380-388.