International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 2 459 - 463

A Survey on Document Clustering For Identifying Criminal

[1]Ms. H. N. Gangavane
Department of Computer Science
& Engineering
BDCOE
[1]*harsha24hng@gmail.com*

[2]Prof. Ms. M. C. Nikose
Department of Computer Science
& Engineering
BDCOE
[2]*monalinikose@gmail.com*

Abstract— Crimes are a social nuisance and cost our society dearly in several ways. Crime investigation has very significant role of police system in any country. Developing a good crime analysis tool to identify crime patterns quickly and efficiently for future crime pattern detection is required. This paper presents combine approach of clustering, outlier detection and providing the rule engine to identify the criminals.

Data mining is the computer-assisted process to break up through and analysing large amount of data. Then extracting the meaning of the data. It is also the process of analysing data from different perspectives and summarizing it into useful information. Data mining plays an important role in terms of prediction and analysis. Clustering is the task of grouping a set of objects in such a way that objects in the same groups are more similar to each other than to those in other groups. The law enforcers have to effectively meet out challenges of crime control and maintenance of public order. Hence, creation of data base for crimes and criminals is needed.

Keywords- *Data mining; Clustering; Outlier detection; Rule engine.*

*****

I. INTRODUCTION

In recent years, volume of crime is becoming serious problems in many countries and today's world criminals have maximum use of all modern technologies and hi-tech methods in committing crimes. The law enforcers have to effectively meet out challenges of crime control and maintenance of public order. 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 required.

The Police Department is responsible for enforcing criminal and traffic law, enhancing public safety, maintaining order and keeping the silence. Alternatively Crime investigation has very significant role of police system. With the increased complexity of crime investigation process police officers have to tolerate a lot of pressure than early days. The most incredible risk for the police department is investigating crimes with the current technologies, because they still use conventional instruction manual processes to handle crimes that are doing with the use of advanced technologies [11].

To enhance the pattern discovery on the data set the clustering is very important task. It is a main task of exploratory data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions.

Therefore Clustering is formulated as a multi objectives optimization problem. The appropriate clustering algorithm and parameter settings depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi objectives optimization that involves trial and failure. It will often be necessary to modify data pre-processing and model parameters until the result achieves the desired properties. Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy, botryology and typological analysis. While in data mining, the resulting groups are the matter of interest, in automatic classification primarily their discriminative power is of interest. This often leads to misunderstandings between researchers coming from the fields of data mining and machine learning, since they use the same terms and often the same algorithms, but have different goal. Document clustering technique such as extractions and Clustering approaches to overcome the difficulty in designing a general purpose document clustering for crime investigation [13].

Various data mining techniques are used for the pre-processing the data set. Data pre-processing techniques are mainly used for producing high-quality mining results. Raw data are being pre-processed before mining because data are in different format, collected from various sources and stored in the data bases and data warehouses. Major steps involved in data mining are data cleaning, data integration, data transformation and data reduction. Our work focus on improve pre-processing technique using NLP. After pre-processing, finally standard data underwent the process of mining and hence better results have obtained. The outlier is an object which is far away from the cluster many researchers are used the clustering techniques there is no separate tool which used for detecting outliers for renovation of it and not provide the rule engine to better result for identifying the criminal.

In proposed project we want to combine all approaches like preprocessing using NLP, clustering, outlier detection and rule engine on criminal data set which gives us

an efficient result for identifying the criminal. Our proposed work is divided into six step like collecting the data set of criminal, then apply the preprocessing using NLP, after that apply the clustering approach for better result to the document, then detect the outlier if any present, then apply rule engine for identifying criminal and the last result evaluation and optimization if required.

II. LITERATURE SURVEY

In the recent decade, a great deal of scientific researches and studies have been performed on crime data mining. The results are usually emerged in the aspect of new software applications for detecting and analyzing crime data

Kaumalee Bogahawatte and Shalinda Adikari 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**,** 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, 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 , 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.

Sotarat Thammaboosadeea, Bunthit Watanapa Nipon, Charoenkitkarna (2012) This paper proposes a framework to identify criminal. They used two – stage classifier. First layer of classifier used "modular Artificial Neural Network algorithm". Second layer of classifier used legal level attribute and the data is collected from the relevant law articles consisting of sentences and range of punishments, given facts discovered in the criminal case of interest [5]. After collecting the relevant data they apply the two stage-classifier on the data for classifying the crime in civil law. They also focus on the decision making, data mining technique and criminal law but on the article law sentences.

Uttam Mande, Y.Srinivas J.V.R.Murthy , In this paper the data set is used images which captured by CCT camera, binary clustering and classification used for analyzing the data and the criminal data is used from Andhra Pradesh police department. The identification of criminal is based on the facial evidence obtained through the CCT cameras and witness/clue at the crime spot using a Generalized Gaussian Mixture model. If the witness is available, at the crime incident, then they provide a novel methodology of constructing a image with the features. Then comparing with the faces in the database and if a match is obtained, it tries to present the details regarding the criminal and his identification marks [6].But the issue is they not used any type of document they used only images.

Malathi. A and Dr. S. Santhosh Baboo,The author of paper are used MV algorithm and Apriori algorithm to detecting the crime pattern. They use semi supervised learning technique in this paper for knowledge discovery from accuracy the crime records and to help increase the predictive. By scientific study of crime, criminal behavior and law enforcement are used to identifying the crime characteristics. They develop the crime pattern analysis tool having four steps data cleaning, clustering, classification and outlier detection to detect the crime pattern and future prediction of crime. They predict the of the crime pattern with related data set only. The main focus of author was to develop a crime analysis tool that assists the police in

- o Detecting crime patterns and perform crime analysis
- o Provide information to formulate strategies for crime prevention and reduction
- o Identify and analyze common crime patterns to reduce further occurrences of similar incidence

Sukanya.M, T.Kalaikumaran and Dr. S. Karthik, In this paper the data set is classified according to crime place, crime type and crime time. They deal with only spatial clustering algorithm and structured crime classification for classify the criminal activities. They are concentrated on find out hotspot of crime, like of burglary crime. Above classification are help to identify the criminal on the bases of witness and clue at the crime spot. This will help to the police department to identify crime place then they can provide prevention and more security to that particular area and GIS is used to visualize the hotspot. But completely crime not in controlled.

Shaym Varan Nath used the clustering technique for identifying the crime pattern. They used method for identification on the basis of expert based semi-supervised learning method and developed the scheme for weighting the significant attributes [9]. They developed weighting scheme for attribute which deals with limitation of various out of box clustering tools and techniques. It helps to plot the geospatial plot to identify the geographical area of the crime. In this the author used K-means algorithm for clustering. But the limitation of this paper is, it will used to identify the crime pattern which is help only the detective for working in initial phases.

S. Yamuna, N. Sudha Bhuvaneswari (2012) they proved the methodology to predict and analyze the crime. Data mining applied in the context of law enforcement and intelligence analysis holds the promise of alleviating crime related problem [10]. The clustering techniques are used DB-Scan algorithm and K-mean algorithm for predicting the crime then it was create the link between the analytical parts of information and finally give the report. The results of this data mining could potentially be used to lessen and even prevent crime for the forth coming years [10].

Subhash Tatale, Sachin Sahare used the Business intelligence for identifying the crime pattern and crime recognition tool which link with combination if data mining techniques for identifying the criminal. They involved major step form data mining i.e. data cleaning; data integration, data transformation and data reduction [11]. This system used communication power of multistage agent for agent systems to increase the efficiency in identifying possible suspects [11].

Uttam Mande Y.Srinivas J.V.R.Murthy methodology was used identify the criminal by mapping the crime using Auto correlation. The author was collected data from andhra pradesh they used the techniques of data mining and clustering. For crime analysis the link with crime, criminal behavior, evidence, crime place etc, then this link mapped with the previous knowledge for criminal identification. These technologies have hindered the effective analysis about the criminals. They also used k-mean algorithm for document clustering.

Anshu Sharma, Raman Kumar[14] they used K-Mean lustering algorithm and neural network and classification respectively.They proposed neural networks for classification. They conclude that the detecting effect and accuracy is more efficient than the classification model.

Andrew Skabar and Khaled Abdalgader[15] used a novel fuzzy clustering algorithm for sentence-text clustering that operates on the relational input data. They proposed the algorithm for clustering named as “Fuzzy Relational Eigenvector Centrality Based Clustering Algorithm [15]. This algorithm is capable for identifying the overlapping clusters of semantically related sentences. But the issue of this algorithm is time complexity.

Mohammad J. Sawar, Umair Abdullah, Aftab Ahmed(2010) Digital forensic is the process of uncovering and interpreting electronic data for use in a court of law[16]. The main objective of this paper preserving any evidence in its most original form while performing a structured investigation by collecting information identifying it and validating . This information was the purpose of reconstructing past events. Digital forensics deals with the analysis of artifacts on all types of digital devices [16]. The Clustering method used to automatically group the retrieved documents into a list of meaningful categories. Document clustering involves descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document cluster is generally considered to be a centralized process. They proposed extracting document and got a brief knowledge in this paper.

III. PROPOSAED WORK

Crime is a very sophisticated domain and identification of criminal is very difficult. In previous system are created with pattern and agent base criminal identification. Increase in the size of crime information that has to be stored and analysed. Different methods and structures used for recording crime data. The data available is inconsistent and are incomplete thus making the task of formal analysis is more difficult. Improvement in clustering can improve the classifier performance. So changes in the K-Means clustering algorithm which is used by various authors can improve the result. K-Means clustering algorithm can be improved by assigning weights to the seeds. This can be done to improve the overall performance of the model.

Because of today's world is not free from crime and the existing System for Criminal identification is very slow and somewhat un-optimized. There is also problems in to identifying techniques that can accurately and efficiently analyze this increasing amount of crime data. To identifying the perfect cluster other approach will be required. Document clustering with new approach for identifying the criminal is required. Some author used the decision tree but not performed the outlier detection and some provide this facility then they don’t use the rule engine. These problems in existing system are motivated to us to create an efficient system for criminal identification.

The main motivation of this work has been to investigate possibilities for the improvement of the effectiveness of document clustering by finding out the main reasons of ineffectiveness of the already built algorithms and get their solutions. Since crime domain is very sophisticated, proper input data preprocessing and document clustering is very important. So many data mining techniques are available but in our proposed work included the Natural Language Processing for extracting the action word. By providing the combined approach of rule engine and outlier detection we will improve the efficiency and reduce the delay to identify crime.

1. To develop the clustering algorithm.

   The document clustering plays an important role for any system because it helps in organizing documents in groups according to their similarity of content. It is apply on various documents for different goal. Here we used the Criminal Data Set or documents for clustering which very sophisticated. So applying new approach for documents clustering is required.

2. To implement data preprocessing by 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.

3. To Dealing with outliers.
   Many authors are used the data mining techniques to clustering but they somewhat ignore the outlier detection. In this current approach we used the technique to detect the outlier if any.

4. To Develop rule engine for criminal identification
   To improve the efficiency of the system and utilization power of rule based systems' rule based engine are used.

We believe that crime data mining has a promising future for increase in the effectiveness and efficiency of criminal and intelligence analysis. Visual and intuitive criminal and intelligence investigation techniques can be developed for crime pattern. As we have applied clustering technique of data mining for crime analysis we can also perform other techniques of data mining such as classification using NLP.

IV. WORKFLOW OF THE PROPOSED SCHEME

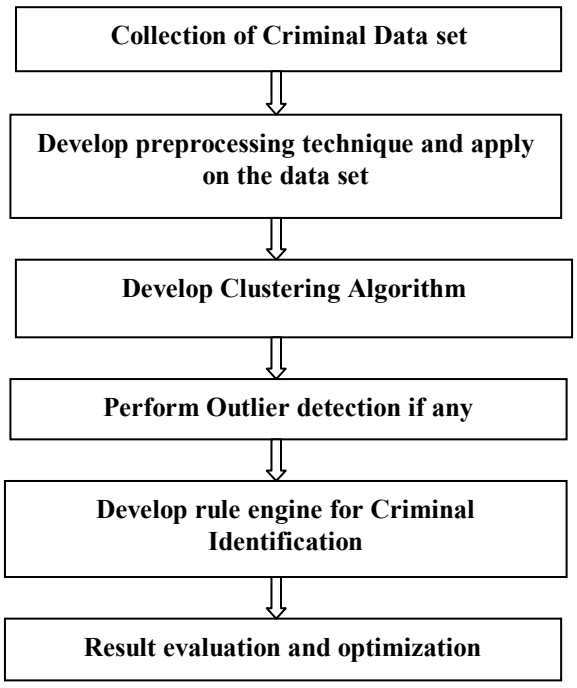

Fig.1 Flow chart of proposed system.

I. Collection of criminal data set -

In that step we collect the criminal Data set from various sources like National Archive of Criminal justice data (NACJD) which is provide they facilitate research in criminal justice and criminology, through the preservation, enhancement, and sharing of computerized data resources; through the production of original research based on archived data, Denver Open data catalog and other on line research. Also we can collect the data from Communities and Crime Data Set. The per capita violent crimes variable was calculated using population and the sum of crime variables considered violent crimes in the United States: murder, rape, robbery, and assault. There is apparently some controversy in some states concerning the counting of rapes. These resulted in missing values for rape, which resulted in incorrect values for per capita violent crime. These cities are not included in the dataset. Many of these omitted communities.

II. Develop preprocessing technique.

After collection of criminal document we have to remove unwanted words. In this step the preprocessing of given data is done by using Natural Language Processing (NLP) that will first perform part of the speech(POS) tagging algorithm then applies to chunking technique in order to filter out only action words.

III. Develop clustering Algorithm

The goal of a document clustering scheme is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances (using an appropriate distance measure between documents). A distance measure (or, dually, similarity measure) thus lies at the heart of document clustering.

The large variety of documents makes it almost impossible to create a general algorithm which can work best in case of all kinds of datasets. Document clustering is being studied from many decades but still it is far from a trivial and solved problem. The challenges are:

1. Selecting appropriate features of the documents that should be used for clustering.
2. Selecting an appropriate similarity measure between documents.
3. Selecting an appropriate clustering method utilizing the above similarity measure.
4. Implementing the clustering algorithm in an efficient way that makes it feasible in terms of required memory and CPU resources.
5. Finding ways of assessing the quality of the performed clustering.

Document clustering is very important so, in third step we develop a clustering algorithm for creating the cluster from input data.

IV. Perform outlier detection if any

Outlier detection in streaming data is very challenging because streaming data cannot be scanned multiple times and also new concepts may keep evolving. Irrelevant attributes

can be termed as noisy attributes and such attributes further magnify the challenge of working with data streams. This is the post processing techniques in which the clustered data is processes by using Silhouetted techniques. This technique like as template matching.

V. Develop a rule engine

The rule technological landscape is becoming ever more complex, with an extended number of specifications and products. It is therefore becoming increasingly difficult to integrate rule-driven components and manage interoperability in multi-rule engine environments. The described work presents the possibility to provide a common interface for rule-driven components in a distributed system. The authors' approach leverages on a set of discovery protocol, rule interchange and user interface to alleviate the environment's complexity. In this step we create a rule engine which apply if – else type rule to the clustered data so, that the criminal should be identified.

VI. Result evaluation and optimization

In this step of module the result should be evaluated like accuracy and delay. This will be optimized if required.

V. CONCLUSION

The literature reviewed has been concluded that the preprocessing techniques are used but not NLP is used for structured the crime data and this data can analysis by two techniques K-Means clustering algorithm is used for clustering and pattern detection for criminal identification etc. Improvement in clustering algorithm can improve the classifier performance. Outlier detection and Rule Engine tool s are used for criminal identification with evidence of crime. These overall performances improve the identification of criminals.

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 "*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] Sotarat Thammaboosadeea, Bunthit Watanapa Nipon, Charoenkitkarna "*A Framework of Multi-Stage Classifier for Identifying Criminal Law Sentences*" Proceedings of the International Neural Network Society Winter Conference (INNS-WC 2012)

[6] Uttam Mande, Y.Srinivas, J.V.R.Murthy "*Criminal Identification System Based On Facial Recognition Using Generalized Gaussian Mixture Model.*" Asian Journal Of Computer Science And Information Technology 2: 6 (2012) 176– 179.

[7] Malathi. A and Dr. S. Santhosh Baboo "An Enhanced Algorithm to Predict a Future Crime using Data Mining" International Journal of Computer Applications (0975 – 8887) Volume 21– No.1, May 2011

[8] Sukanya.M, T.Kalaikumaran and Dr.S.Karthik "*Criminals and crime hotspot detection using data mining algorithms: clustering and classification*" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1, Issue 10, December 2012

[9] Shaym Varan Nath " *Crime pattern detection using Data mining*" Oracal Co opration

[10] S.Yamuna, N.Sudha Bhuvaneswari "*Datamining Techniques to analyze and predict crime*". The International Journal of Engineering And Science (IJES) Volume 1 Issue 2 Pages 243-247 2012 ISSN: 2319 – 1813 ISBN: 2319 – 1805

[11] Subhash Tatale, Sachin Sakhare " *Intellectual Crime Recognition System* ."IOSR Journal of Computer Science (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727 PP 40-45

[12] Priyanka Gera, Rajan Vohra, "*City Crime Profiling Using Cluster Analysis*" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5145-5148

[13] Uttam Mande Y.Srinivas J.V.R.Murthy "*An Intelligent Analysis Of Crime Data Using Data Mining & Auto Correlation Models*" International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 Vol. 2, Issue 4, July-August 2012, pp.149-153

[14] 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

[15] Andrew Skabar, and Khaled Abdalleder "*Clustering Sentence level Text using a novel fuzzy relation clustering algorithm*" IEEE transaction on Knowledge and data engineering ,vol ,25, No.1.,January - 2013

[16] Mohammad J. Sawar, Umair Abdullah, Aftab Ahmed"*Enhanced Design of a Rule Based Engine Implemented using Structured Query Language*" Proceedings of the World Congress on Engineering 2010 Vol I WCE 2010, June 30 - July 2, 2010, London, U.K.