University of Sindh Journal of Information and Communication Technology (USJICT)
Volume 2, Issue 1, January 2018
ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

A Survey of Data Mining Techniques for Crime Detection

Shamaila Qayyum, Hafsa Shareef Dar
Department of Software Engineering, International Islamic University, Islamabad
shamaillla@gmail.com, hafsa_darr@yahoo.com

Abstract: In large datasets, data mining is one of the most powerful ways of knowledge extraction or we can say it is one of the best approaches to detect underlying relationships among data with the help of machine learning and artificial intelligence techniques. Crime Detection is one of the hot topics in data mining where different patterns of criminology are identified. It includes variety of steps, starting from identification of crime characterization till detection of crime pattern. For this purpose, various crime detection techniques have been discussed in literature. In this paper, we have selected widely adapted data mining techniques that are specifically used for crime detection. The analytical study is presented with an extraction in form of strengths and weakness of each technique. Each technique is specific to its use. This survey would serve as a helping guide to researchers to get state of the art crime detection techniques in data mining along with pros and cons.

Keywords: Data mining, Crime detection, Classification, Clustering, Association, Prediction, Constraint, Association rules

I. INTRODUCTION

With the advancement of technology, criminals are also adapting smarter ways to commit crimes. With time, the crime rate has been tremendously increased instead of decreasing. The technology where, has helped the people in making their lives easier, has also helped the criminals in making new plans for their crimes. It is becoming trivial to get smarter and technical ways to investigate and prevent these crimes. One of the most common ways of crime that can be seen nowadays, is not just within the streets, but in the world of connectivity; the internet. It is deemed important to discover and adapt the ways that help in efficient discovery and prevention of such crimes. The criminal acts range from the street crimes to massive terror attacks and undoubtedly to offend large databases and systems. This all lie under the umbrella of crime[1]. Crimes can be defined for example as network attacks [2], Hacking [3], network intrusion [2], cyber fraud [2, 4],deceptive information [5].

Data mining refers to the extraction, discovery and analysis of meaningful patterns and rules from a very large amount of data. It is emerging as very useful tools for crime detection [6]. Data mining is a very powerful tool to undermine the activities of criminals by analyzing the criminal's record and information and preventing the crimes in future[7]. Data mining for crime detection is considered one of the most important research area as described in [8] despite data mining being a new and evolving field itself [9]. Data mining is thought of being very helpful and accurate in understanding the crime trends as compared to Humans. Humans are often error prone specially when overworked; computers prove to be more accurate as compared to human. Mining information through computers is more convenient and less costly than hiring and training people for the purpose of collecting and analyzing existing crime information. Data Mining can help the investigators, no matter if they are experienced or not, to explore large databases efficiently[10]. This not only helps them in tracking and solving crimes but also predicting crimes in advance[11]. Data mining techniques offers some predictive models that manipulate the hidden information and can predict the trends [12].

Hosseinkhani et al.,[13] has suggested some data mining techniques that can be used for crime detection. These techniques are clustering, association rule mining, deviation detection, classification and string comparator. The crime detection data mining techniques as presented by Hsinchun et al., are entity extraction, clustering, association rule mining, sequential pattern mining, deviation detection, classification, string comparator and social network analysis [14]. Hossein Hassani et al., have later presented a review of some existing crime data mining techniques which included; entity extraction, cluster analysis, association rule, classification and social network analysis [15]. The aim of this study is to get an insight of the techniques that are followed for crime data mining. Previously, survey on data mining techniques was conducted with only five techniques that covered small scope of data mining. In this survey, we intend to cover each possible technique used for crime detection.

In this study the existing techniques of data mining for crime detection and investigation are thoroughly observed. Different data mining techniques are discussed and analyzed for their usage. They are then further compared for their strengths and weaknesses under different

circumstances of usage. The differences and commonalities of these techniques have also been discussed.

II. EXISTING CRIME DATA MINING TECHNIQUES:

Crime can range from the simple street crimes to internationally planed crimes[8]. Crime data mining, as compared to usual data mining, is more concerned with privacy [16]. In structured data, the patterns are identified through different traditional data mining techniques such as association, classification, prediction, clustering and outlier analysis[17]. Advanced data mining handles both structured and unstructured data for pattern recognition [14, 18]. In this section we analyze the existing data mining techniques that are used for crime detection and investigation.

A. Entity extraction

It is used to identify persons, vehicles, texts basically by identifying patterns [19]. It is the process of extracting data from text documents [15]. In computer forensics it can help in identifying programs write by hackers. This is done by grouping similar programs. However it requires a huge amount of clean data to produce good results[3]. The main approaches for entity extraction are machine learning, statistic based, rule based and lexical lookup [19]. Machine learning techniques use algorithms to extract knowledge and derive patterns e.g, decision trees, neural networks[20], entropy maximization [21] and hidden markov models [22]. Rule based systems are the structural, contextual or lexical hand crafted rules to identify entities[23]. Statistical based systems used training dataset to obtain statistics for identifying the occurrence of particular patterns[24]. Lexical loops systems maintain popular entities of interest and look up in text the phrases that are specified in their lexicons[21].

B. Clustering techniques

They help in maximizing or minimizing the interclass similarities by grouping the data items into the classes based on their characteristics. In criminal investigation, it can help in identifying criminals that follow a set pattern for committing a crime [25]. Clustering functions by obtaining the distance measurement among objects, such as Euclidean distance, Minkowksi distance and Manhattan distance. Different algorithms of cluster group the data into hierarchical manner or partition them as per requirements[26]. Apart from these two approaches, other main approaches for clustering are: density based, grid based, model based and constraint based.

C. Association rule mining

It presents the patterns as the rules by finding frequent occurrences of items within a dataset. This is helpful in identifying network attacks [2]. It was initially built to observe interesting co-occurrences in market data[23]. This is very helpful in investigating the simultaneous occurrences of events[27]. The strength of association can be measured in terms of support; the applicability of rule to given dataset and confidence; the frequency of appearance of one data in transactions that contain another data[28].

D. Sequential pattern mining

Sequential pattern mining discovers frequently occurred sequence of items at different intervals of times. It is helpful in network intrusion detection. For meaningful results, a large amount of structured data is required[2]. Ayres et al., have provided an algorithm that finds all possible sequences in the transactions data, very quickly. They have used depth first traversal combined with the bitmap representation to achieve this [29].

E. Deviation detection

It is also known as outlier detection. It studies data that has clear distinction from rest of the data set. This technique is very helpful in fraud detection and other crime analysis[2, 4]. Aggarwal, in his research has proposed an evolutionary outlier detection algorithm which works by selection then crossover and then mutation methods [30]. Whereas, Arnind et al., have proposed a linear method of deviation detection in larger database, which provides solution by simulating a mechanism that is familiar to human beings [31].

F. Classification

Classification has some predefined classes. This data mining technique finds out some common characteristics of different crime entities. These are then organized into the predefined classes. A predefined classification scheme is required for this approach[4]. Classification is very helpful in predicting crimes and identifying crime entities with a very less amount of time. However, it needs a predefined scheme for classification and complete training with testing data, as only this could bring accuracy to the predicted results [4]. Classification aims to discover a set of rules from the dataset. Classification can be carried out either by Decision Trees[32-34], Support Vector Machines [35], [35]Naïve Bayes Rules[36] or Neural Networks [37-42].

G. String comparator

It computes the similarity among the database records by comparing their textual fields. It helps identifying deceptive information. However it requires a large amount of computations [5]. It returns a numerical value by comparing two strings [23].

H. Social network analysis

Social network analysis analyses the role and interaction of nodes within a conceptual framework. It can be used to identify criminal's roles by creating a network.

It can also help in analyzing the flow of information among these entities, though it won't help in identifying networks' true leaders[5].it reveals the structure within some text, by presenting some interlinked entities[43]. This shows that people have participated or communicated somewhere[44]. The most widely used techniques for SNA are: Degree; number of nodes connected to any node [45], Density; number of edges in a specific area as compared to the overall number of edges[46] and Centrality; the importance of a node within a structure[47].

Hossein Hassani et al., have reviewed the data mining techniques for crime. This review cover the techniques: entity extraction, cluster analysis, association rule, classification and social network analysis [15]. In [48] a tool was discussed which is based on Natural Language Processing technique for detection of white collar crimes. However, a comparative analysis of all above mentioned crime data mining techniques is still missing in the literature.

III. ANALYSIS

In this section, comparative analysis of each technique is presented on the basis of its strength and weakness. The strength and weakness of each technique has been extracted from literature review in which authors and researchers have identified positive and negative impacts of particular technique. Table 1 is a complete analysis of each technique:

TABLE 1 STRENGTH AND WEAKNESS OF EACH CRIME DATA MINING TECHNIQUE

TECHNIQUE	STRENGTH	WEAKNESS
Entity extraction	Machine learning makes it easier	large amount of clean data required
Clustering	Detect outliers without any required label data	Computational cost is high. Its effectiveness also depends upon the method used
Association Rule mining	Support	It is used for the most accurate classification rules
Sequential pattern mining	Wide range of applicability	large amount of structured data is required
Deviation detection	Widely applicable in fraud detection	Sometimes its data dependency becomes a hurdle
Classification	Very less time consumption	Predefined scheme of classification and complete training dataset required
String Comparator	Accuracy in terms of numerical value	Large amount of computation required
Social network analysis	focus on relationships between actors rather than attributes of actors	Won't identify network's true leaders

The mentioned techniques have been deeply studied to know the pros and cons of any technique while adapting it in crime detection. Entity Extraction is enhanced by machine learning techniques but polluted data can be a hurdle to it so its weakness is requirement of clean data. The strength of clustering is detection of outliers without labeled data but as this process is costly and its effectiveness depends on selected method as well. Association Rule Mining is yet another technique which basically supports classification and its weakness is its specific nature to classification rules only. Sequential pattern mining has wide range of applicability in all areas and hence, large amount of structured data is required for its execution. Deviation detection is widely used in fraud detection but data dependency in some areas is still a question unsolved. Classification technique is conventional technique with very less time consumption and weakness is predefined scheme of classification that requires complete training data set. String comparator accuracy is great when we consider numerical values but it requires high computations. Social network analysis focuses on relationships between actors rather than their attribute which makes it more direct but unfortunately it doesn't identify network's true leader in the system.

IV. CLASSIFICATION OF EXISTING TECHNIQUES

In order to understand the techniques used for crime data mining, we first present classification of these techniques. This classification contains the data mining techniques that are specifically used for crime data mining and are stated in the literature. The classification of the techniques along with the methods they use is shown in Fig 1.

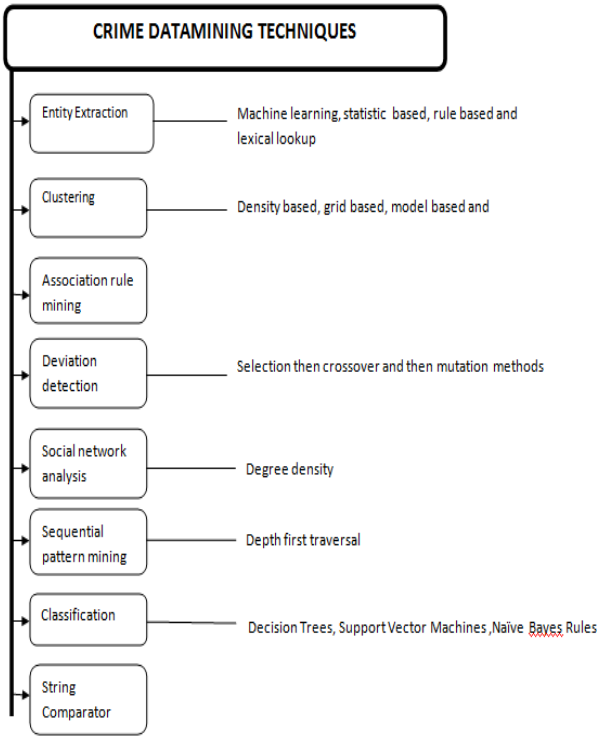

Figure 1. Classification of crime data mining techniques

In order to clearly understand the usage of these techniques, it must be clear to the researchers and investigators that what kind of usage each technique possess. In order to take the full advantage of these techniques and to make their selection easier for the researchers and investigators, we have thoroughly studied the techniques that are helpful in crime data mining; we have then prepared a comparison of these techniques based on their usage. We have identified all the areas in which crime data mining techniques can be used and have further identified which data mining technique is used in each scenario. We have thus, further classified these techniques based on their usage. The usage of each crime data mining technique is presented in Table 2.

TABLE 2 USAGE BASED CLASSIFICATION OF CRIME DATA MINING TECHNIQUES

USAGE	DATA MINING TECHNIQUE	REFERENCE
Identification of programs written by hackers	Entity extraction	[11, 15, 16, 17, 18, 19, 20, 21]
Identifying criminals following a set pattern	Clustering	[22, 23]
Identifying network attacks	Association rule mining	[20, 24, 25, 26]
Intrusion detection	Sequential pattern mining	[24, 27]
Fraud detection	Deviation detection	[24, 28, 29, 30]
Predicting crimes	Classification	[28, 31-33, 36-41]
Identifying deceptive information	String Comparator	[20, 42]
Identifying Criminals role	Social Network analysis	[42, 43, 44, 45, 46, 47]

This classification not only helps researchers but also investigators, to choose the right technique for their scenario, without wasting time. To understand these techniques in more details and to have a comparative analysis of these details, we have identified some attributes of these techniques, after a thorough study of the techniques. These attributes are defined as follow:

i. Timeliness is one of the most important attribute. Most investigators are concerned with the timely response of a data mining technique.
ii. Data dependency is another important attributes. At times the techniques require a lot of data to produce accurate results, but the required amount of data may not be available.
iii. Accuracy is very important in identifying the true criminals. So this attribute is also very important for analyzing crime data mining techniques.

A. Comparison of Existing Crime Data Mining Techniques

This section aims to compare and analyze the existing techniques for crime data mining, on the basis of proposed thematic taxonomy. We have learnt that there are various techniques used for detection of crime, and every technique is used in different scenario. So the real comparison is according to the scenario in which they are used, and they may not need to be compared for the common attributes. But however, it is important to understand the comparison of these techniques for effective selection and flawless investigation. The comparison is based on three values – timeliness, data dependency and accuracy of the technique. Table 3 shows the comparison of these techniques based on the proposed taxonomy.

TABLE 3 COMPARISON OF CRIME DATA MINING TECHNIQUES

COMPARISON OF CRIME DATA MINING TECHNIQUES			
ATTRIBUTE/TECHNIQUE	TIMELINESS	DATA DEPENDENCY	ACCURACY
Entity extraction	Less time consumption	Huge amount of data required	Accurate
Clustering	Moderate Time consumption	Less data dependency	Accurate
Association Rule mining	Moderate	Moderate	Accurate
Sequential pattern mining	Moderate	Huge amount of data required	Accurate
Deviation detection	Less time consumption	Moderate	Accurate

Classification	Less time consumption	Complete training set data required	Accurate on data availability
String Comparator	More time consumption	Moderate	Accurate
Social network analysis	Moderate	Moderate	No true leader identified

[2]

The comparison presented in table 2 is quite elaborative in terms of attribute's timeliness, data dependency and accuracy. Entity extraction is first attribute which has high data dependency with minimum time consumption and optimal accuracy. Clustering is one of the renowned techniques in data mining, but for crime detection its time consumption is moderate having less data dependency and is accurate. Association rule mining is yet another technique used for crime detection having moderate timeliness and data dependency with accurate results. Sequential pattern mining is used as a technique. According to the comparison it has moderate time consumption with huge amount of data dependency but gives accuracy in results. The technique named deviation detection has less time consumption moderate data dependency and accuracy is optimal. Similarly, classification has less timeliness, and data is dependent on complete training dataset and accuracy depends on availability of data. String comparator has more time consumption, moderate data dependency with accurate results and social network analysis technique has moderate time consumption and data dependency with unidentified accuracy level.

V. CONCLUSION

We have discussed the use of crime data mining using different techniques like clustering, association rules, sequential patterns and others. We identified eight techniques along with their strengths and weakness. Our contribution here was to provide comparative analysis of these techniques. This analysis is helpful as a researcher guide for studying and identifying crime data mining techniques. There are some limitations of each technique including computational efforts, cost, structured data and rules. On the basis of three attributed criteria: timeliness, data dependency and accuracy, we have analyzed each technique. Based on the strength and weakness identified, it is concluded that each technique is specific to crime detection data mining scenario and has significant attributes. One of the facts occurred during this research is that crime data they require data mining experts and data analysts equipped with sufficient knowledge of data mining and they need to collaborate with detectives in early phases of crime detection.

VI. FUTURE WORK

Crime data is an important area in which efficient crime detection data mining techniques play vital role for analyst and law enforcers to proceed the case in investigations and help resolving criminal cases. The scope of this study can be further enhanced by working on criminal investigation data set like FBI and crime detection of counter terrorism measures. Another enhancement of this technique is to implement in any integrated ERP software

REFERENCES

[1] P. Kanellis, Digital crime and forensic science in cyberspace: IGI Global, 2006.

[2] W. Lee, S. J. Stolfo, and K. W. Mok, "A data mining framework for building intrusion detection models," in Security and Privacy, 1999. Proceedings of the 1999 IEEE Symposium on, 1999, pp. 120-132.

[3] A. Gray, S. MacDonell, and P. Sallis, "Software forensics: Extending authorship analysis techniques to computer programs," 1997.

[4] O. De Vel, A. Anderson, M. Corney, and G. Mohay, "Mining e-mail content for author identification forensics," ACM Sigmod Record, vol. 30, pp. 55-64, 2001.

[5] G. Wang, H. Chen, and H. Atabakhsh, "Automatically detecting deceptive criminal identities," Communications of the ACM, vol. 47, pp. 70-76, 2004.

[6] J. Hosseinkhani, S. Ibrahim, S. Chuprat, and J. H. Naniz, "Web Crime Mining by Means of Data Mining Techniques," Research Journal of Applied Sciences, Engineering and Technology, vol. 7, pp. 2027-2032, 2014.

[7] P. Thongtae and S. Srisuk, "An analysis of data mining applications in crime domain," in Computer and Information Technology Workshops, 2008. CIT Workshops 2008. IEEE 8th International Conference on, 2008, pp. 122-126.

[8] H. Chen, W. Chung, Y. Qin, M. Chau, J. J. Xu, G. Wang, et al., "Crime data mining: an overview and case studies," in Proceedings of the 2003 annual national conference on Digital government research, 2003, pp. 1-5.

[9] V. Pinheiro, V. Furtado, T. Pequeno, and D. Nogueira, "Natural language processing based on semantic inferentialism for extracting crime information from text," in Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on, 2010, pp. 19-24.

[10] U. Fayyad and R. Uthurusamy, "Evolving data into mining solutions for insights," Communications of the ACM, vol. 45, pp. 28-31, 2002.

[11] S. V. Nath, "Crime pattern detection using data mining," in Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on, 2006, pp. 41-44.

[12] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI magazine, vol. 17, p. 37, 1996.

[13] J. Hosseinkhani, M. Koochakzaei, S. Keikhaee, and J. H. Naniz, "Detecting suspicion information on the Web using crime data mining techniques," International Journal of Advanced Computer Science and Information Technology, vol. 3, pp. 32-41, 2014.

[14] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: a general framework and some examples," Computer, vol. 37, pp. 50-56, 2004.

[15] H. Hassani, X. Huang, E. S. Silva, and M. Ghodsi, "A review of data mining applications in crime," Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 9, pp. 139-154, 2016.

[16] H. Kargupta, K. Liu, and J. Ryan, "Privacy sensitive distributed data mining from multi-party data," in International Conference on Intelligence and Security Informatics, 2003, pp. 336-342.

[17] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques: Elsevier, 2011.

[18] G. Gupta, Introduction to data mining with case studies: PHI Learning Pvt. Ltd., 2014.

[19] M. Chau, J. J. Xu, and H. Chen, "Extracting meaningful entities from police narrative reports," in Proceedings of the 2002 annual national conference on Digital government research, 2002, pp. 1-5.

[20] S. Baluja, V. O. Mittal, and R. Sukthankar, "Applying Machine Learning for High-Performance Named-Entity Extraction," Computational Intelligence, vol. 16, pp. 586-595, 2000.

[21] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman, "Exploiting diverse knowledge sources via maximum entropy in named entity recognition," in Proc. of the Sixth Workshop on Very Large Corpora, 1998.

[22] S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, et al., "Algorithms that learn to extract information: Bbn: Tipster phase iii," in Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998, 1998, pp. 75-89.

[24] I. H. Witten, Z. Bray, M. Mahoui, and W. J. Teahan, "Using language models for generic entity extraction," in Proceedings of the ICML Workshop on Text Mining, 1999.

[25] R. V. Hauck, H. Atabakhsb, P. Ongvasith, H. Gupta, and H. Chen, "Using Coplink to analyze criminal-justice data," Computer, vol. 35, pp. 30-37, 2002.

[26] R. T. Ng and J. Han, "E cient and E ective Clustering Methods for Spatial Data Mining," in Proc. of, 1994, pp. 144-155.

[27] H. Yun, D. Ha, B. Hwang, and K. H. Ryu, "Mining association rules on significant rare data using relative support," Journal of Systems and Software, vol. 67, pp. 181-191, 2003.

[28] P.-N. Tan, Introduction to data mining: Pearson Education India, 2006.

[29] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, "Sequential pattern mining using a bitmap representation," in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002, pp. 429-435.

[30] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in ACM Sigmod Record, 2001, pp. 37-46.

[31] A. Arning, R. Agrawal, and P. Raghavan, "A Linear Method for Deviation Detection in Large Databases," in KDD, 1996, pp. 164-169.

[32] C. J. Stone, "Classification and regression trees," Wadsworth International Group, vol. 8, pp. 452-456, 1984.

[33] J. R. Quinlan, C4. 5: programs for machine learning: Elsevier, 2014.

[34] J. R. Quinlan, "Induction of decision trees," Machine learning, vol. 1, pp. 81-106, 1986.

[35] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, pp. 273-297, 1995.

[36] P. Langley, W. Iba, and K. Thompson, "An analysis of Bayesian classifiers," in Aaai, 1992, pp. 223-228.

[37] M. D. Richard and R. P. Lippmann, "Neural network classifiers estimate Bayesian a posteriori probabilities," Neural computation, vol. 3, pp. 461-483, 1991.

[38] G. P. Zhang, "Neural networks for classification: a survey," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 30, pp. 451-462, 2000.

[39] H. Gish, "A probabilistic approach to the understanding and training of neural network classifiers," in Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on, 1990, pp. 1361-1364.

[40] P. A. Shoemaker, "A note on least-squares learning procedures and classification by neural network models," IEEE Transactions on Neural Networks, vol. 2, pp. 158-160, 1991.

[41] E. A. Wan, "Neural network classification: A Bayesian interpretation," IEEE Transactions on Neural Networks, vol. 1, pp. 303-305, 1990.

[42] B. Widrow, D. E. Rumelhart, and M. A. Lehr, "Neural networks: applications in industry, business and science," Communications of the ACM, vol. 37, pp. 93-106, 1994.

[43] J. Mena, Investigative data mining for security and criminal detection: Butterworth-Heinemann, 2003.

[44] A. M. Fard and M. Ester, "Collaborative mining in multiple social networks data for criminal group discovery," in Computational Science and Engineering, 2009. CSE'09. International Conference on, 2009, pp. 582-587.

[45] M. K. Sparrow, "The application of network analysis to criminal intelligence: An assessment of the prospects," Social networks, vol. 13, pp. 251-274, 1991.

[46] A. Iriberri and G. Leroy, "Natural language processing and e-government: extracting reusable crime report information," in Information Reuse and Integration, 2007. IRI 2007. IEEE International Conference on, 2007, pp. 221-226.

[47] K. Chan and J. Liebowitz, "The synergy of social network analysis and knowledge mapping: a case study," International journal of management and decision making, vol. 7, pp. 19-35, 2005.

[48] Maartin B., et. al., Performance Evaluation of a Natural Language Processing approach applied in White Collar Crime, Springer adfa, p 1., Berlin, 2011