International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 2, No 2, 2014. ISSN: 2348-6600
Reference ID: IJCS-059 PAGE NO: 343-347.

# A New Approach on Crime Detection with Data Mining and Cloud Computing

J. Mohana Sundaram[#1], C. Thirumoorthy[*2]

[#1]*Assistant Professor, Department of Computer Science, PSG College of Arts & Science, Coimbatore*

[#1] mohanpsgphd@gmail.com

[*2]*Assistant Professor, Department of Computer Science Hindustan College of Arts & Science – Coimbatore*

[*2] cthirumoorthymca@gmail.com

***Abstract*— The increase in the population causes unemployment that makes people to earn money in whatever way against the law. The invention of new technologies and this unemployment problem make the country to move towards the path of disasters. It is an emerging concern about national security, needed to have still more attention to identify the crimes. Intelligence agents, various public and private departments are involved in the process of crime detection and in the enforcement of laws. It is needed to collect data on crimes, classify data according to the crime type, analyzed to identify important areas of crime to concentrated and finally to find the prediction s over the protection of future crimes. The process acquired through the ocean of cloud computing, analyzed in cluster analysis and applied with the various data mining techniques to extract the needed data in the context of law enforcement and intelligence analysis. A model for this process has been designed to organize, analysis, and to make predictions over the data for crime detection.**

***Index Terms* - Crime detection, cloud computing, data mining techniques, enforcement of laws, Intelligence agents.**

## I. INTRODUCTION

Internet is often said to be a cloud and the term "Cloud Computing" arises from that analog. Cloud computing is defined as the dynamic provisioning of capabilities (hardware, software or services) from third parties over a network[3]. It is said that clouds are the services based on hardware offering to compute, network and store where hardware management facilities are highly available and Infrastructure capacity is highly elastic. The cloud is different from other traditional models where the users not needed to put hands to manage over their own resources. Cloud computing is involved in identifying all the possible crimes.

Data Mining allows analyzing data in many different directions or angles categorize the analyzed data and summarize the identified relevant needed data for the task. Data mining is a process used to find the future trends and behaviors and helpful in making proactive, knowledge-driven decisions and it involves massive data collection process as a prior step and identifying he patterns needed through different algorithms of data mining.

Cluster analysis is the process of finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters[1]. A good clustering method will produce high quality clusters with high intra-class similarity. The clustering result quality depends on both the similarity measure used by the method and its implementation. It is also measured by its ability to discover some or all of the hidden patterns. It is an essential need to have collaborated work of the above said concepts in order to identify the crime and to make in with timely decisions.

## II CLOUD COMPUTING – AN OVERVIEW

The improvements in the science field results in the management of huge data that are computerized and stored in databases. The stored data contains hidden knowledge that is very important and useful for many decision making approaches. The conventional statistical methods which are not possible to unravel this knowledge and are also time consuming. Moreover, it may produce improper conclusions ultimately affect decision making. So it is an essential need to collect the appropriate data and made used over efficient techniques to arrive with the better solutions.

### *A. Cloud Computing Vs Internet Technologies*

Cloud computing is said to be a mode of delivering hardware, software or services on a virtualized and scalable platform using Internet Technologies. It enables client



Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 2, No 2, 2014. ISSN: 2348-6600
Reference ID: IJCS-059 PAGE NO: 343-347.

organizations to have access to highly available services with charges that vary based on utilization. It eliminates the organizations to invest on highly expensive technological equipments and minimizes the risk. The services of a cloud computing can be provided to the users in three major categories such as software, platform and the infrastructure. Now in the field of technological the cloud computing is said to be a great beneficiary for obtaining a large scale of data. It is benefit because of its qualities such as paying according the usage, instant scalability, security, reliability etc.

## III MASSIVE DATA COLLECTION

It is an essential need to have an eye over the national security at all the times. When compared to the previous decades the crime rate is increased as like that of a speed of a rocket. It is increased not only in its rate, also improved in with the latest technologies. So it is the social responsibility and duty of the Government and various intelligent agents to improve their quality by implementing new technologies and methods for identifying the criminals, the methods handled by the criminals, the security measures available etc. This massive data collection can be made through different mediums. But it is an easiest way to collect them through the means of cloud computing. As we said earlier it will provide the data, software, hardware and also their other essential needed services.

### *A. Intelligence Agencies in Crime Detection*

Intelligence agencies are actively collecting and analyzing information to investigate terrorist' activities [3]. Local law enforcement agencies have also become more alert to criminal activities in their own jurisdictions. When the local criminals are identified properly and restricted from their crimes, then it is possible to considerably reduce the crime rate of national crimes. There exists a large volume of data over these criminals and it is identified as a massive collection. There are different levels of crimes to be identified and to concentrate for the welfare of the nation. Refer Table I.

### *B. Different Techniques in Crime Detection*

Entity extraction has been used to automatically identify person, address, vechile, narcotic drug, and personal properties from police narrative reports. Clustering techniques have been used to automatically associate different objects (such as persons, organizations, vehicles) in crime records. Deviation detection has been applied in fraud detection, network intrusion detection and other crime analyses that involve tracing abnormal activities. Classifications has been used to detect email spamming and find authors who send out unsolicited emails. String comparator has been used to detect deceptive information in criminal records. Social network analysis has been used to analyze criminals' roles and association among entities in a criminal network[6].

## IV FILTERING OF DATA - AN APPROACH

The increasing adoption of the Intelligence – Led Policing model plus analysis at the heart of operational, tactical and strategic decision-making. In this model, intelligence serves as a guide to operations, rather than the reverse. Therefore, it is now more improvement than ever to find out how data mining can help create better understanding and predictions.

### *A. Traditional Approach in Analyzing Data on Crimes*

Traditionally, police systems focus on small parts of the available data (e.g., year, month, type of crime) for a specific purpose (e.g. monitoring crime rate for strategy). Without data mining, the amount of data used in analysis is limited by the time that analyses have to go through it, step by step. It is simply unfeasible to analyze all the potentially useful data by hand. However, for many policing problems it is important to use as much data as possible, to be able to explain, understand, link and predict, since the explanation of a phenomenon typically lies in small details. By using more data, patterns offer more contextual information and help analysis reach the right conclusions. So, to understand crime, data is needed that goes beyond simple aspects of an incident or person. Linking crimes by similarity also benefits from rich data for the same reasons. Finding links can help to detect crime series, connect cases and solve them. This is where need for data mining comes in and plays an eminent role for the data analyzing and eminent one for decision makings

.

### *B. Automated Pattern Recognition*

It is necessary to turn the data overload into a manageable flow of information that matters. Apart from handling the volume of data, data mining techniques also help to deal with the dynamic nature and complexity of criminal behavior. This can be seen apart from the data volume subject discussed above, simply because complex patterns can be present in just a few data elements. For example, the following question is



Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 2, No 2, 2014. ISSN: 2348-6600
Reference ID: IJCS-059 PAGE NO: 343-347.

very hard to answer using conventional techniques: where and when do crimes take place during the week looking at X – Y co-ordinates of the incidents, the time of day and the day of the week. A proven method to answer this question is to use clever clustering techniques.

It is a common misconception that data mining requires large data volumes in order to add value. On the contrary, it is our experience that when implementing data mining it is best to start with one, maybe two data sources to get acquainted with the possibilities and to manage expectations. Many organizations are surprised by how much information they can gain from data mining on just a small part of their data. Nevertheless, more data is always better for providing more depth and more contexts.

Data mining allows the police to better understand and predict crime because many data sources can be analyzed and complex patterns can be found. In 2004[8], an extensive study by the program bureau of the Dutch Police (ABRIO) concluded that 'data mining enables more effective and goal driven decision making on strategic, tactical and operational levels.

## V THE K-MEANS ALGORITHM – IN CRIME DETECTION

The k-means algorithm randomly selects an object from the available objects, each of which initially represents a cluster mean[11]. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges.

The k-mean algorithm takes the input parameter k and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high but the inter-cluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's centre of gravity.

The k-mean algorithm proceeds as follows. First, it randomly selects k of the objects. The square error criterion is used and defines as

$$E = \sum_{I=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

Where E is the sum of the square error for all objects in the data set; p is the point in space representing a given object, and $m_i$ is the mean of cluster $C_i$ (both p and $m_i$ are multidimensional). In other words for each object in each cluster, the distance from the object to its cluster centre is squared, and the distance are summed.

Basic K-mean Algorithm for finding k Clusters,

Step: 1 Select K points as the initial centroids,

Step: 2 Assign all points to the closed centroid.

Step: 3 Recomputed the centroid of each cluster, and

Step: 4 Repeat steps 2 and 3 until the centroids do not change.

There are many variants of the k-means method. They may be different in the selection of the initial centroids, the calculation of dissimilarity and the methods for calculating cluster. There are many existing clustering algorithms which work more efficiently to cluster the large datasets. Cluster a large datasets using k-mean algorithm with relational database seems to be more iterative and require more time. But using SQL the massive datasets can be clustered very effectively. In document clustering, one of the variants, bi-secting k-means, outperforms the basic k-means as well as the agglomerative approach in terms of accuracy and efficiency. The bi-secting k-means algorithm is a divisible hierarchical clustering method.

### *A. Basic Bi-Secting K-means Algorithm*

The bi-secting k-means is actually a divisive clustering algorithm that achieves a hierarchy of clusters by repeatedly applying the basic k-means algorithm. In each step of bi-secting k-means a cluster is selected to be split and it is split into two by applying basic k-means for k = 2. The largest cluster that is the cluster containing the maximum number of documents can be chosen to be split.

Step: 1 Pick a cluster to split

Step: 2 Find 2 sub-clusters using the basic of k-means algorithm.(Bi-secting step),

Step: 3 Repeat step 2, the bi-secting step, for iterative times and take the split that produces the clustering with the highest overall similarity, and

Step: 4 Repeat step 1, 2 and 3 until the desired number of clusters is reached.

The continuous proceedings of the above said steps lead to the process of grouping related data together.

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 2, No 2, 2014. ISSN: 2348-6600
Reference ID: IJCS-059 PAGE NO: 343-347.

*Evaluation of Algorithms*

Both the basic and bi-secting k-means algorithms are relatively efficient and scalable. The complexity of both algorithms is linear in the number of documents. In addition, they are so easy to implement that they are widely used in different clustering applications.

A major disadvantage of k-means is that it requires the user to specify k, the number of clusters, in advance, which may be impossible to estimate in some cases. Incorrect estimation of k may lead to poor clustering accuracy. Also, it is not suitable for discovering clusters of totally different size which is very common in document clustering. Moreover, the k-means algorithm is sensitive to noise and outlier data objects as they may substantially influence the mean value, which in turn lower the clustering accuracy.

## VI. CRIMINAL NETWORK ANALYSIS

Criminals often develop networks in which they form groups or teams to carry out various illegal activities. Our data mining task consisted of identifying subgroups and key members in such networks and then studying interaction patterns to develop effective strategies for disrupting the networks. Our data came from various Police Department incident summaries involving 164 criminals committed from 1985 through May 2003[14]. It is used with a concept-space approach to extract criminal relations from the incident summaries and create a likely network of suspects. Co-occurrence weight measured the relational strength between two criminals by computing how frequently they were identified in the same incident.

### *A. Hierarchical Clustering in Partitioning Crime Groups*

Hierarchical clustering is used to partition the network into subgroups and the block-modeling approach to identify interaction patterns between these subgroups. It is also calculated centrally measures – degree, difference and closeness – to detect key members in each group, such as leaders and gatekeepers. In Figure 2, the circles represent subgroups the system found, and they bear the labels of their leaders' names [14]. A circle's size is proportional to the number of members in that subgroup. The thick-ness of straight lines connecting circles indicates the strength of relationships between subgroups. Refer Figure 1 & 2.

## VII. DESIGN MODEL SUPPORTING DATA PREDICTION & PRECAUTION MEASURES

The proposed model is designed with an approach towards a supportive aspect for various national security departments and intelligent agents. The technologies of cloud are highly implied for the massive collection of data related to the crime over different locations and over different times. The collected data are used to go with an analysis stage for splitting of data into different crime types. In that process the data are made to flow through k-means algorithm to identify the groups in an efficient, speedy manner. The analyzed data are charted based on the groups and are highly useful to crime departments for making prior decisions over the security measure. The table below is the structure of proposed model. Refer table 2.

## CONCLUSION

The patterns, associations or relationships among all this data can provide information. Information can be converted into knowledge for behavior. Various data mining techniques are used to identify the patterns and relationships from the data. Several data mining techniques were applied to discover new and interesting pattern relationships from large spatial data. This paper contributes a major support in the crime pattern discovers or detection through the technique of cloud computing and the usage of k-mean algorithm for clustering of identified data. It is shown that how data mining and clustering algorithms can be used to identify crimes and how it is helpful in reducing and preventing criminal activities.

## REFERENCES

[1]. U,M. Fayyad and R. Uthurusamy, "Evolving Data Mining into Solutions for Insight," Comm, ACM, Aug, 2002, pp. 28-31.

[2]. W. Chang et al., "An International Perspective on Fighting Cybercrime," Proc, Ist NSF/NIJ Symp, Intelligence and Security Informatics, LNCS 2665, Springer-Verlg, 2003

[3]. Dr. Wendy A. Warr, Wendy Warr & Associates (wendy@warr.com, http://www.warr.com), November 2009.

[4]. J.Han a dM.Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001.

[5]. H. Karupta, K. Liu, and Ryan, "Privacy-Sensitive Informatics, NSF/NIJ Symp, Intelligence and Security Informatics, LNCS 2665, Springer-Verlg, 2003

[6]. M. Chau, J.J Xu, H. Chen, "Extracting Meaningful Entities from Police Authorship Analysis Techniques to Computer Programs," Conf. Digital Government Research, Digital Government Research Centre, 2002

[7]. A Gray, P. Sallis, and S.MacDonell, "Software Forensics: Proc, 3rd Biannual conf, Int'l Assoc, Forensic Linguistics.

[8]. R.V. Hauck et al., "Using Coplink to Analyze Criminal-Justice Data," Computer, Mar. 2002, pp. 30-37.



Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 2, Issue 2, No 2, 2014. ISSN: 2348-6600
Reference ID: IJCS-059 PAGE NO: 343-347.

[9]. T. Senator et al, "The FinCEN Artifical Intelligence System: Identifying Potential Money Laundering from Reports of Large Cash Transactions," AI Magazines, vol.16.

[10]. W. Lee. S.J Stolfo, and W. Mok "A Data Mining Framework for Building Intrusion Detection Models," Proc 1999 IEEE Symp., Security and Privacy, IEEE CS Press, 1999

[11]. C MeCue, "Using Data Mining to Predict and Prevent Crimes," available at: http:/ /www. Spss .com/ dirvideo / richmond.htm?source=dmpage&zone=rtsidebar.

[12]. O. de Vel et al., "Mining E-Mail Content for Author Identification Forensics," SIGMOD Record, vol. 30, no.4.

[13]. G. Wang, H.Chen, "Automatically Detecting Deceptive Criminal Identities," Comm, ACM, Mar 2004.

[14]. S .Wasserman and K. Faust, Social Network Analysis: Methods and Applications, Cambridge University, Press.

[15]. Huber, F. Matthes, L., Vollhardt, K. Ulbrich D(2006) "Centre of Market-Oriented Product and Production Management Mainz ISBN:3-938879-13.

[16]. Chau. M, Xu, Chen.H(2002) Extracting meaningful entities from police narrative reports. In Proceedings of the National Conference for Digital Government Research.

[17]. Chen. H., Lynch., "Automatic construction of networks of concepts characterization of documents databases, IEEE Transactions on Systems, Man,Cybernetics, 22(5).



Published by SK Research Group of Companies (SKRGC).