Crime Analysis and Prediction Using Data Mining

Shiju Sathyadevan, Devan M.S
Amrita Center for Cyber Security
Amrita Vishwa Vidyapeetham,
Amritapuri, Kerala, India
shiju.s@am.amrita.edu, devanms@am.amrita.edu

Surya Gangadharan. S
Amrita Vishwa Vidyapeetham
Amritapuri, Kerala, India
suryagangadharan@yahoo.com

***Abstract*—Crime analysis and prevention is a systematic approach for identifying and analyzing patterns and trends in crime. Our system can predict regions which have high probability for crime occurrence and can visualize crime prone areas. With the increasing advent of computerized systems, crime data analysts can help the Law enforcement officers to speed up the process of solving crimes. Using the concept of data mining we can extract previously unknown, useful information from an unstructured data. Here we have an approach between computer science and criminal justice to develop a data mining procedure that can help solve crimes faster. Instead of focusing on causes of crime occurrence like criminal background of offender, political enmity etc we are focusing mainly on crime factors of each day.**

Keywords—***Naive Bayes; Apriori algorithm; Decision tree; NER; Mongo DB; Neo4j; GraphDB;***

I. INTRODUCTION

Day by day the crime rate is increasing considerably. Crime cannot be predicted since it is neither systematic nor random. Also the modern technologies and hi-tech methods help criminals in achieving their misdeeds. According to Crime Records Bureau crimes like burglary, arson etc have been decreased while crimes like murder, sex abuse, gang rape etc have been increased. Even though we cannot predict who all may be the victims of crime but can predict the place that has probability for its occurrence.

The predicted results cannot be assured of 100% accuracy but the results shows that our application helps in reducing crime rate to a certain extent by providing security in crime sensitive areas. So for building such a powerful crime analytics tool we have to collect crime records and evaluate it [1].

It is only within the last few decades that the technology made spatial data mining a practical solution for wide audiences of Law enforcement officials which is affordable and available. Since the availability of criminal data or records is limited we are collecting crime data from various sources like web sites, news sites, blogs, social media, RSS feeds etc. This huge data is used as a record for creating a crime record database. So the main challenge in front of us is developing a better, efficient crime pattern detection tool to identify crime patterns effectively. The main challenges we are facing are:

- Increase in crime information that has to be stored and analyzed.
- Analysis of data is difficult since data is incomplete and inconsistent.
- Limitation in getting crime data records from Law Enforcement department.
- Accuracy of the program depends on accuracy of the training set.

Finding the patterns and trends in crime is a challenging factor. To identify a pattern, crime analysts takes a lot of time, scanning through data to find whether a particular crime fits into a known pattern. If it does not fit into an existing pattern then the data must be classified as a new pattern. After detecting a pattern, it can be used to predict, anticipate and prevent crime.

Before this clustering algorithms have been used for crime analysis. For instance, one site it is revealed that suspect has black hair and from next site/witness it is revealed that suspect is youth and from third one reveals that the offender has tattoo on his left arm etc. By describing the offender details it gives a complete picture from different crime incidents. Today most of it is manually done with the help of multiple reports that the detectives usually get from the computer data analysts and their own crime logs.

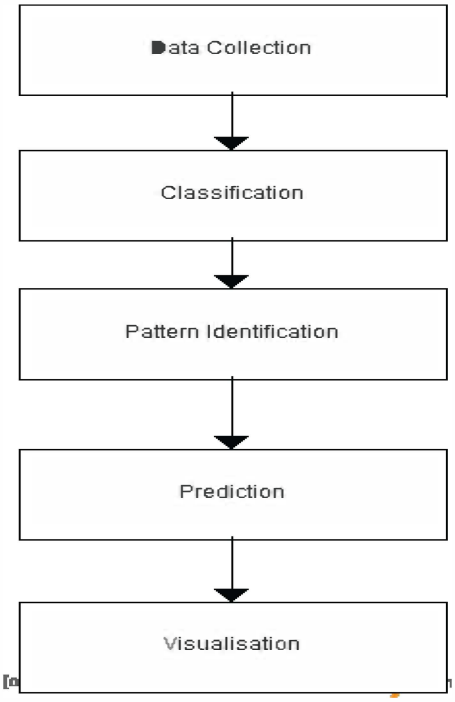
The reason for choosing this method is that we have only data about the known crimes we will get the crime pattern for a particular place. Therefore, classification technique that will rely on the existing and known solved crimes, will not give good predictive quality for future crimes. Also nature of crimes change over time, so in order to be able to detect newer and unknown patterns in future, clustering techniques work better.

There are steps in doing Crime Analysis:

1) Data Collection
2) Classification
3) Pattern Identification
4) Prediction
5) Visualization

978-1-4799-3486-7/14/$31.00©2014 IEEE

Fig. 1. Steps in Crime analysis



II. Related Work

In countries like England, Cambridge Police Department have done a similar one named Series Finder for finding the patterns in burglary. For achieving this they used the modus operandi of offender and they extracted some crime patterns which were followed by offender. The algorithm constructs modus operandi of the offender. The M.O. is a set of habits of a criminal and is a type of behaviour used to characterize a pattern.

The data included means of entry (front door, window, etc.), day of the week, characteristics of the property (apartment, house), and geographic proximity to other break-ins. Using nine known crime series of burglaries Series Finder recovered most of the crimes within these patterns and also identified nine additional crimes. The predicted result showed more than 80% accuracy. So the same concept we are applying here i.e. find unknown patterns from known data and facts [5]. It's the first mathematically principled approach to the automated learning of crime series.

III. Methodology

A. Data Collection

In data collection step we are collecting data from different web sites like news sites, blogs, social media, RSS feeds etc. The collected data is stored into database for further process. Since the collected data is unstructured data we use Mongo DB. Crime data is an unstructured data since the no of field, content, and size of the document can differ from one document to another the better option is to have a schema less database. Also the absence of joins reduces the complexity. Other benefits of using an unstructured database is that:

- Large volumes of structured, semi-structured, and unstructured data.
- Object-oriented programming that is easy to use and flexible.

The advantage of NoSQL database over SQL database is that it allows insertion of data without a predefined schema. Unlike SQL database it not need to know what we are storing in advance, specify its size etc.

B. Classification

For classification we are using an algorithm called Naïve Bayes which is a supervised learning method as well as a statistical method for classification. Naive Bayes classifier is a probabilistic classifier which when given an input gives a probability distribution of set of all classes rather than providing a single output. The algorithm classifies a news article into a crime type to which it fits the best. From classification what we get is "What is the probability that a crime document D belongs to a given class C?" [2].

The advantage of using Naive Bayes Classifier is that it is simple, and converges quicker than logistic regression. Compared to other algorithms like SVM (Support Vector Machine) which takes lot of memory the easiness for implementation and high performance makes it different from other algorithms. Also in case of SVM as size of training set increases the speed of execution decreases.

Using Naive Bayes algorithm we create a model by training crime data related to vandalism, murder, robbery, burglary, sex abuse, gang rape, arson, armed robbery, highway robbery, snatching etc. By training means we have to teach them on particular inputs such that we can test them for unknown inputs. For testing the accuracy of the model we apply test data. Unlike SVM as the size of training data increases accuracy of test set also increases. Another advantage of Naïve Bayes is that it works well for small amount of training to calculate the classification parameters.

Also it fixes the Zero-frequency problem i.e. while estimating probability sometimes while checking a probability P(A) * P(B/D) * P(C/D) * P(E/D) where P(C/D)=0. So the estimated probability result always give zero which leads to uncertainty in results. To avoid this condition we add +1 to the count of every zero value classes to achieve uniform distribution.

Test results shows that Naive Bayes shows more than 90% accuracy since it categorise each words as tokens and removing frequent words like "the", "and", "of" etc which improves accuracy. A word is automatically terminated if it occurred fewer times or less than 3 times. Figure 2 shows a sample pseudo code of Naïve Bayes algorithm.

We are also integrating the concept of Named Entity Recognition(NER) in the crime articles. NER also known as Entity Extraction finds and classify elements in text into pre-defined categories such as the person names, organizations,

locations, date, time etc[11]. So by using this concept in crime article we can get more details related to crime like victim and offender names, location of crime, date, time etc [6]. A sample result of NER is shown in Figure 3.

Fig. 2. Pseudo code of Naïve Bayes

Algorithm 1 Pseudocode

1. Given training data set D which consists of documents belonging to different class say class A and B.
2. Calculate the prior probability of class A=number of objects of class A / total number of objects
   Calculate the prior probability of class B=number of objects of class B / total number of objects
3. Find ni, the total number of word frequency of each class.
   na= the total number of word frequency of class A.
   nb= the total number of word frequency of class B.
4. Find conditional probability of keyword occurrence given a class.
   P(word1 / class A) = wordcount / ni(A)
   P(word1 / class B) =wordcount / ni(B)
   P(word2 / class A) = wordcount / ni(A)
   P(word2 / class B) =wordcount / ni(B)
   ….....
   .....
   P(wordn / class B) =wordcount / ni(B)
5. Avoid zero frequency problems by applying uniform distribution.
6. Classify a new document C based on the probability P(C /W).
   a) Find P(A / W) = P(A) * P(word1/ class A) * P(word2/ class A) ……* P(wordn / class A).
   b) Find P(B / W) = P(B) * P(word1 / class B) * P(word2/ class B) ……* P(wordn / class B).
7. Assign document to class that has higher probability.

Also related to crimes like burglary we can extract the list of weapons offender used while committing the crime. We have included a concept called Coreference Resolution to find the referenced entities in a text. In linguistics, Coreference occurs when two or more expressions in a text refer to the same person or thing i.e. if they have the same referent[12].

1) Input: NAVI MUMBAI: The bike borne chain snatchers targeted two women pedestrians in Sanpada and Panvel on May 6, 2014, Tuesday and robbed their gold ornaments. While, 60-year-old woman's gold chain worth Rs 20,000 was snatched by the bike's pillion rider around 3.45 pm, while she was walking on the street near HDFC bank in sector-14, Sanpada, yet another woman from Khalapur was targeted by the pillion rider while she was walking along the road near old Thane naka in Panvel. The thief snatched away her gold necklace worth Rs 67,500. In both the incidents, robbery case under Section 392 and 34 has been registered at Turbhe and Panvel police stations respectively.

Fig. 3. A sample output of NER


```
{
  "nerList": [
    {
      "location": "Vashi"
    },
    {
      "location": "MUMBAI"
    },
    {
      "location": "Sanpada"
    },
    {
      "location": "Sanpada"
    },
    {
      "location": "Panvel"
    },
    {
      "date": "May 6,2014"
    },
    {
      "date": "Tuesday"
    }
  ]
}
```


For example : Seema said she would come i.e. here "she" refers to person "Seema". Likewise we are extracting all referenced entities in a text. Below example shows the working of Coreference concept. A sample is shown below in Fig 4.

2) Input: E.g.: A pillion bike rider snatched away a gold mangalsutra worth Rs 85,000 of a 60-year-old woman pedestrian in sector 19, Kharghar on Friday. The victim, Shakuntala Mande, was walking towards a vegetable outlet around 9.40am, when a bike came close to her and the pillion rider snatched her mangalsutra. A robbery case has been registered at Kharghar police station.

Fig 4 Sample output of Coreference Resolution


```
["The victim" -> "Shakuntala Mande"
 "her mangalsutra" -> "a gold mangalsutra"
 "Kharghar" -> "Kharghar on Friday"
 "her" -> "Shakuntala Mande"
 "the pillion rider" -> "A pillion bike rider"]
```


C. Pattern Identification

Third phase is the pattern identification phase where we have to identify trends and patterns in crime. For finding crime pattern that occurs frequently we are using Apriori algorithm. Apriori can be used to determine association rules which highlight general trends in the

database. The result of this phase is the crime pattern for a particular place. Here corresponding to each location we take the attributes of that place like VIP presence, weather attributes, area sensitivity, notable event, presence of criminal groups etc. After getting a general crime pattern for a place, when a new case arrives and if it follows the same crime pattern then we can say that the area has a chance for crime occurrence. Information regarding patterns helps police officials to facilitate resources in an effective manner. They can avoid crime occurrence by providing security/ patrolling in crime prone areas, fixing burglar alarms / CCTV etc.

Take a sample list of 100 news for a place and apply Apriori algorithm in it. It will mine the frequent crime patterns for a place. So if there is a pattern in which crime occurred then we assume that if again that pattern occurs in a place then there is probability for crime occurrence in that place. We are considering several attributes for crime pattern detection.

E.g.: For a place Meerut the pattern after mining will be:

- attribute 1, attribute 2, attribute 3, attribute 4
- attribute 1, attribute 3, attribute 4, attribute 5

So the above will be the crime pattern for Meerut. So crime occurs only if the above patterns occur on a day. If any of these patterns occur then only we can say that there is probability for crime occurrence.

D. Prediction

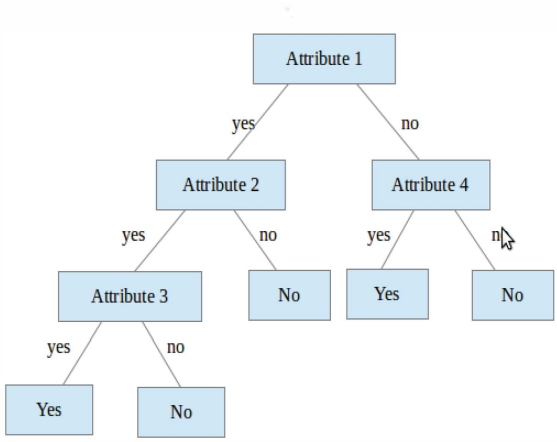
For prediction we are using the decision tree concept. A decision tree is similar to a graph in which internal node represents test on an attribute, and each branch represents outcome of a test. The main advantage of using decision tree is that it is simple to understand and interpret. The other advantages include its robust nature and also it works well with large data sets. This feature helps the algorithms to make better decisions about variables [4].

Corresponding to each place we build a model. So for getting the crime prone areas we pass current date and current attributes into the prediction software. The result is shown using some visualization mechanisms. Fig 5 shows the example of a decision tree model. Below shown is the example of decision trees of two different places Meerut and Delhi.

TABLE I. DECISION TREE FOR DELHI

Area sensitivity	**Notable event**	**VIP presence**	**Criminal group**	**crime**
Yes	Yes	Yes	No	yes
Yes	Yes	No	Yes	no
No	No	No	Yes	no
Yes	No	No	No	no
Yes	Yes	Yes	yes	yes
No	Yes	No	No	no

Fig. 5. E.g. of a decision tree



The working of decision tree seems to be little confusing but it's really easy. Consider a variety of plant species. We classify them according to order, genus, species etc. Instead we have to classify them into a common category as shrubs and trees. If a new species is identified then we have to classify this into any of the two categories. Basically we categorize it based on its characteristics i.e. we have a set of questions to check whether it satisfies the conditions. If first condition is satisfied then we check the next case and if the first condition itself is not satisfied then there is no need to check the rest. So the series of questions and their answers can be organized in the form of a decision tree. The tree has three types of nodes:

- A Root node, that has incoming edges and zero or more outgoing edges.
- Internal nodes, each of which has one incoming edge and two or more outgoing edges.
- Leaf node or end node, each of which has exactly one incoming edge and no outgoing edges. [8]

This supervised machine learning technique builds a decision tree from a set of class labeled training samples and by using this tree, tests the new samples [4] .It is a predictive model which uses a set of binary rules to calculate the class value. The tree determines:

- Which variable to split at a node.
- Decision to stop or split.
- Assign terminal nodes [9].

E. Visualization

The crime prone areas can be graphically represented using a heat map which indicates level of activity, usually darker colors to indicate low activity and brighter colors to indicate high activity. Below figure is an example of a heat map.

Below Figure 7 shows the regions that has high probability for crime occurrence. The advantages of using heat maps over other representational mechanisms are:

- Numeric and category based color images.
- Gradient color range.
- Analyze only the data we want.
- Out of range data is automatically discarded.

So by knowing about the probable regions we can prevent crimes by taking preventive mechanisms like night patrolling, fixing burglar alarms, fixing CCTV camera etc.

Fig 7 Map showing crime prone areas

Figure 8 shows the statistical data. In the x-axis all main locations in India are plotted whereas in y-axis the crime rate is plotted. The graph shows the regions which has maximum crime rate. It is calculated based on the crime rate in crime records. From the graph plotted based on historical data, it is clear that Delhi has maximum crime. The data plotted here is based on the historical records. The graph changes over time when we add more data into the records.

Figure 9 shows the rate/percentage of crime occurrence in places like airport, temples, bus station, railway stations, bank, casino, jewelry shops, bar, ATM, airport, bus station, highways etc. In the x axis the main spots like temple, bank, bus station, railway station, ATM etc are plotted while in y-axis the rate of crime is plotted.

This representation is strictly based on the historic crime records in the database. So our results shows that crimes like robbery, murder, highway robbery and burglary is higher in regions which lacks proper security and also less inhabited whereas crimes like arson, vandalism occurs when there is any notable event happening or VIP presence. Crimes can be solved to a great extend by fixing burglar alarms, providing proper security in less inhabited/ crime prone areas, increasing night patrolling and fixing CCTV's in sensitive areas.

Fig 8 Statistical Data

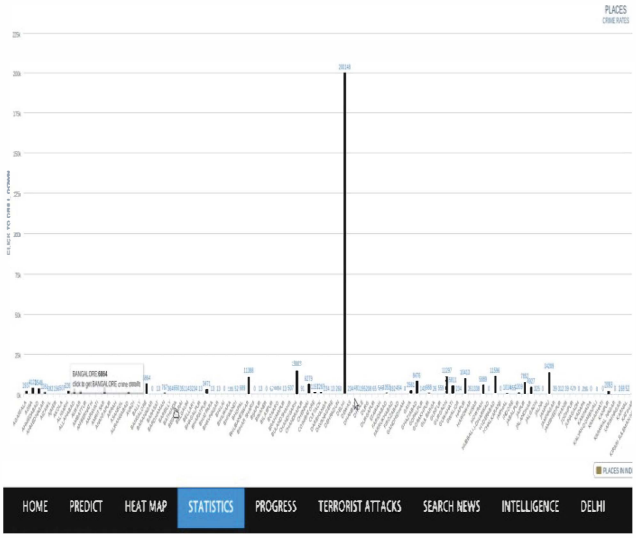
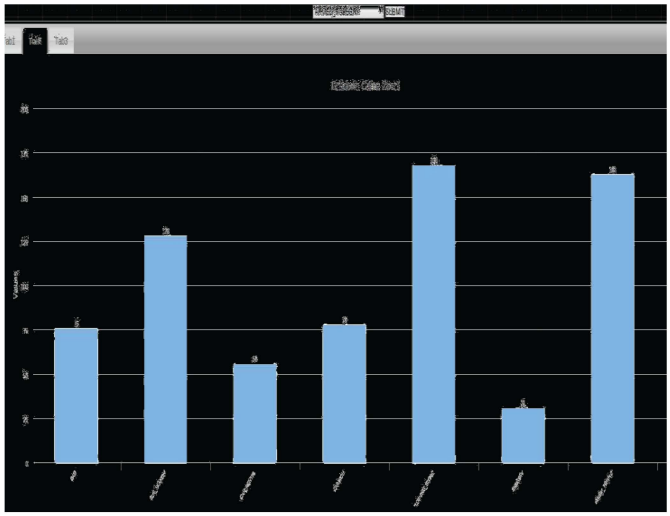


Fig. 9. Statistical data



IV. Future Work

A. Criminal Profiling

In addition to this a new concept called Criminal profiling which helps the crime investigators to record the characteristics of criminals. It is a very accurate tool for profiling the characteristics or details of offenders is a behavioural and investigative tool that is intended to help investigators to accurately predict and profile the characteristics of unknown criminal subjects or offenders. The main goal of doing criminal profiling is that:

- to provide crime investigators with a social and psychological assessment of the offender;
- to evaluate belongings found in the possession of the offender.

For doing this we have to analyse the criminal backgrounds and criminal records for collecting the maximum criminal data. So the maximum details of each criminals is collected from criminal records. i.e. when crimes like burglary occurs in a certain place then from reports like FIR we get the offender details and their modus operandi(mode of operation). After getting these details we can know about the criminals with these behaviour. So sifting through each crime record after a particular crime occurrence is tedious task. So instead we can use some visualisation mechanisms to represent the criminal details in a human understandable form.

For representing criminal data we use a graph database called Neo4j. We propose a data model and query language that integrates an explicit modelling and querying of graphs smoothly into a standard database environment. It allows representation of partitioning object classes into simple classes, link classes and path classes whose object can be viewed as nodes and edges. [3]. Using this we can represent each criminals and their attributes in a flexible node format. Below is a sample example of criminal data represented using Neo4j. Here only certain attributes of criminals like name, hair-colour, eye-colour, nationality, blood group, age, marital status, whether member of any criminal groups etc. Fig 9 shows an example of criminal records represented using GraphDB.

B. *Snatching*

We are concentrating more on crimes like snatching to get more details related to it like crime location, time, date, crime type(which type of snatching), victim and offender names etc. Currently we are getting crime details like:

1) Name of person(victims, offenders)
2) Location
3) Organization
4) Type of crime(whether murder, robbery)
5) Subcategories of crime type(for snatching there are other categories like chain snatching, purse snatching etc)
6) Type of vehicle offender used.
7) Whether any weapons used.
8) Time of incident
9) Date
10) Incident summary
11) Criminal groups involved

V. CONCLUSION

In this paper we have tested the accuracy of classification and prediction based on different test sets. Classification is done based on the Bayes theorem which showed more than 90% accuracy. Using this algorithm we trained numerous news articles and build a model. For testing we are inputting some test data into the model which shows better results. Our system takes factors/attributes of a place and Apriori algorithm gives the frequent patterns of that place. The pattern is used for building a model for decision tree.

Corresponding to each place we build a model by training on these frequent patterns. Crime patterns cannot be static since patterns change over time. By training means we are teaching the system based on some particular inputs. So the system automatically learns the changing patterns in crime by examining the crime patterns. Also the crime factors change over time. By sifting through the crime data we have to identify new factors that lead to crime. Since we are considering only some limited factors full accuracy cannot be achieved. For getting better results in prediction we have to find more crime attributes of places instead of fixing certain attributes. Till now we trained our system using certain attributes but we are planning to include more factors to improve accuracy.

Our software predicts crime prone regions in India on a particular day. It will be more accurate if we consider a particular state/region. Also another problem is that we are not predicting the time in which the crime is happening. Since time is an important factor in crime we have to predict not only the crime prone regions but also the proper time.

ACKNOWLEDGEMENT

We thank our college Amrita School of Engineering, Amritapuri and Amrita Center of Cyber Security, Amritapuri for giving us an opportunity to be a part of the internship program, that leads to the development of this work. Many thanks to Shiju Sathyadevan and Devan M.S for countless discussions and feedback that help me to complete the work successfully.

REFERENCES

[1] Malathi. A and Dr. S. Santhosh Baboo. Article:an enhanced algorithm to predict a future crime using data mining. International Journal of Computer Applications, 21(1):1–6, May 2011. Published by Foundation of Computer Science.

[2] Eibe Frank and Remco R. Bouckaert. Naive bayes for text classification with unbalanced classes. In Proceedings of the 10th European Conference on Principle and Practice of Knowledge Discovery in Databases, PKDD'06, pages 503–510, Berlin, Heidelberg, 2006. Springer-Verlag.

[3] Ralf Hartmut Güting. Graphdb: Modeling and querying graphs in databases. In Proceedings of the 20th International Conference on Very Large Data Bases, VLDB '94, pages 297–308, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc

[4] Lior Rokach and Oded Maimon. Decision trees. In Oded Maimon and Lior Rokach, editors, The Data Mining and Knowledge Discovery Handbook, pages 165–192. Springer, 2005.

[5] Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. Detecting patterns of crime with series finder. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2013), 2013.

[6] Li Zhang, Yue Pan, and Tong Zhang. Focused named entity recognition using machine learning. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information

Retrieval, SIGIR '04, pages 281–288, New York, NY, USA, 2004. ACM.

[7] http://www.cs.sunysb.edu/~cse634/lecture_notes/07apriori.pdf. Last accessed:12-April-2014, 10:00 AM

[8] http://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf. Last accessed:14-April-2014, 1:00 PM.

[9] Ned Horning . Introduction to Decision trees and Random Forests, American Museum of Natural History's Center for Biodiversity and Conservation.

[10] Lafferty, McCallum, and Pereira (2001); Sutton and McCallum(2010)."http://aliasi.com/lingpipe/demos/tutorial/classify/read-me.html [2010].

[11] http://nlp.stanford.edu/software/jenny-ner-2007.pdf. Last accessed :24-Feb-2014.

[12] Wikipedia contributors.(9 July 2013), Stanford NLP. [Online].Available :http://www-nlp.stanford.edu/software/dcoref.shtml. Last accessed: 24-Feb-2014, 10:00 AM.

[13] Wikipedia contributors.(12 May 2014 at 19:05.), Series Finder. [Online].Available:http://en.wikipedia.org/wiki/Crime_analysis, Last accessed: 12-Feb-2014, 12:00 PM.