Analysis of Crime Data using MongoDB

Ila Savant[1], Mayur Gade[2], Rohan Kalap[3], Ajay Kamble[4], Kamran Khan[5]

Professor, Computer Engineering, MMCOE, PUNE, India [1]
Student, Computer Engineering, MMCOE, PUNE, India [2]
Student, Computer Engineering, MMCOE, PUNE, India [3]
Student, Computer Engineering, MMCOE, PUNE, India [4]
Student, Computer Engineering, MMCOE, PUNE, India [5]

Abstract: Crime analysis is a systematic approach of the intelligent and law enforcement organizations for identifying and analyzing trends and patterns in crime. Generally they collect crime related data to prevent future attacks and take necessary action accordingly. A major challenge is efficiently and accurately analyzing the growing volumes of crime related data. This paper presents a crime analysis system which is designed to overcome the above mentioned problem. The proposed system comprises of crime analysis techniques such as crime comparison and crime pattern visualization. The proposed system consists of an easy and simplified environment for performing crime analysis effectively.

Keywords: Crime analysis, RSS news feeds, crime patterns, document classification, entity extraction, duplicate detection.

I. INTRODUCTION

Day by day the crime rate is increasing rapidly. Crime cannot be predicted because it is not systematic or random. The predicted results are not 100% correct but our system helps in reducing crime rate to a certain extent. The availability of criminal data or records is limited so that we are collecting crime data from RSS feeds. So the main challenge is developing a better and simpler crime pattern detection tool to identity crime patterns. The main challenges that we are facing are:

A. Increase in crime information that is to be stored and analyzed.
B. Since the data is incomplete and inconsistent analysis of data is difficult.
C. Limitation in getting crime related data from law enforcement department.
D. Accuracy of predicted results.

This paper presents a crime analysis system which makes use of newspaper articles to perform crime analysis for crime comparison and crime pattern visualization. The newspaper articles are downloaded from RSS feeds, required entities are extracted from articles and duplicate detection is performed. By using preprocessed data crime analysis is performed and visual reports are displayed.

When newspapers are considered, they contain articles only for a subset of total crime incidence. However they contain major crimes that take place in India. Most of the time police are more concerned about major crime incidents then taking decisions. Therefore crime analysis results based on newspaper will be helpful to identify pattern and relate a crime to another one, and will be useful for researchers, investigators for their respective task.

II. BIG DATA

Big Data is the broader term for Data sets which are so large or complex that traditional data processing application and methodologies are inadequate for processing [8]. The challenges in processing such large data sets include capture, data acquisition, storage, transfer, analysis, visualization, and information privacy. The term usually refers to simply to the use of predictive analysis of data or other certain advanced methodologies to extract value from data and seldom to particular size of data set. The most confident decision making in Big Data is known by its accuracy. And better accuracy means reduced risk, cost reduction and more operational efficiency [8]. The data set may contain structured database, unstructured database or semi structured database. According to gather 5V's definition Bid Data has five characteristics: volume, variety, velocity, variability or value as show in fig. 1.

A.   Volume (Large amount of data):
The amount of data set is usually referred as volume. In case of Big Data this dataset is huge in size. This data can be generated every second Ex. Image, video, audio, emails, sensor data, log reports, news feeds etc. This data can be in terabytes, petabytes or zettabytes.

B.   Velocity(fast processing velocity):
It means the speed of production of data set and the time they take to move around.

C.   Variety(Many type of data and source):
This term usually refers to the different types of data sets such as structured, unstructured or semi-structured data. This data is usually generated from emails, audio files, documents, video, images, log files, click streams, call records, or financial transactions. Many different attributes in multiple dimensions in the database increases more and more complexity for the traditional database management tools and applications to handle.

Fig. 1. Big Data V's Model

D.   Variability(Correct-meaning useful data):
This term refers to the correctness and trust worthiness of the data. With many forms excellence and accurateness of data are less controllable. Ex. Facebook posts with asterisk (*), hash tag (#), underscore (_), tiled (~),smiles, strikers, abbreviation, typos and colloquial speech. Big data analytics tools and technology now allow us to work with these types data. The huge volumes often make up for the lack of excellence and accurateness.

E.   Value(low-density data value):
Big data tends to have a relatively low value density, as compared to the data we manage in traditional system. For Example logistics industry have best mode to transport for goods based on weight and value or ratio of business relevance to the size of the data.

III. RELATED WORK

Crime data mining is one of the applications of data mining technique [3]. There are number of researches are carried out in this domain and few of them are given in [2], [1]. An intelligent crime analysis system is explained in [4] which is used to predict suspect. Message, gateway, prisoner, criminal, evidence is the five agents they have used.

A system was developed in [5] to collect and perform analysis on drug crime data. Rule based extraction tool is used to find where and how drug smugglers hide drugs, identify types of drug and nationality of drug dealers.

Crime detection system in [6] was developed to discover relationship between offenders and communities and visualize the results to help crime data analysis.

IV. METHDOLOGY AND PROPOSED SYSTEM

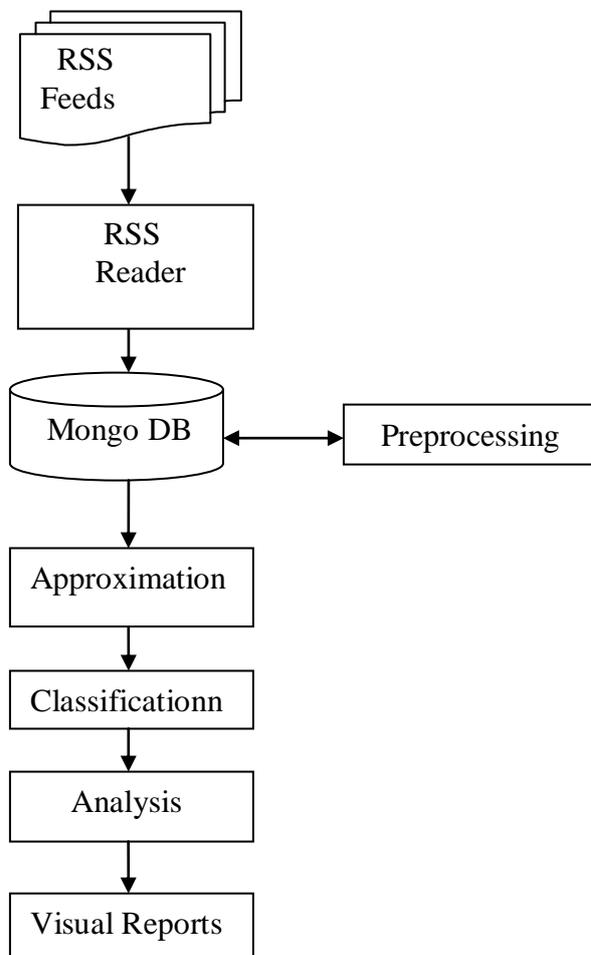The Flow chart of proposed methodology is illustrated in following fig. 2.

Fig. 2. Proposed system Architecture.

Input: - Live Data from RSS feeds.
Output: - Filtered Data in key value pair.

A. Data Acquisition and Filtration Algorithm
Steps:
1. Not all data collected from RSS feed is relevant to crime. It may also about other daily news. The data of our interest i.e. Data of crime is selected for further processing and all other unnecessary data is discarded.
2. The selected data is then divided into appropriate Key Value pair and given to the next step.
3. Assign and transmit each distinct block of the processed data to various processing steps in Data Processing Unit.

47

This algorithm takes live RSS Feed Data and then filters and divides them into segments and performs approximation.

In Step I, related details filtered out.

In Step 2, filtered data are the association of different key value pairs and each pair is different numbers of sample, which results in formation of data block.

In Next Steps, these blocks are forwarded to Data Processing Unit.

B. Processing and calculation Algorithm

Input: Filtered Data

Output: Normalized News data in Numerical comparable form along with Historical Values.

Steps:
1. For each data block, relevant Historical Data is extracted.
2. This data is then normalized for all the live Feeds received.

```
┌─────────────────────────────┐
│   News Data Acquisition     │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│      Data Filtration        │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│     Entity Extraction       │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│    Duplicate Detection      │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│      Crime Analysis         │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│      Visual Reports         │
└─────────────────────────────┘
```
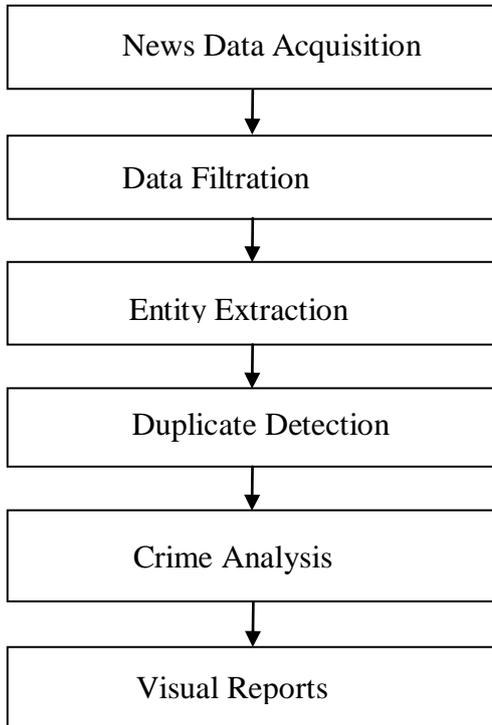
Fig. 3. Flow chart of the methodology

C. Visualization

In this module analyzed crime data results are shown in the form of pie charts which is used for crime comparison. For Example,
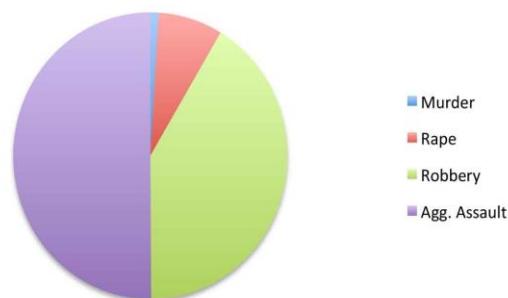

Fig. 4. Crime Comparison

V. CONCLUSION

Retrieving relevant news from the large volume of data is very tedious task. Our system will help the end user to retrieve relevant news from RSS feeds. The proposed system performs crime analysis operation such as crime comparison and pattern visualization. Graphical user interface of system generates graphs, pie charts to display analysed results which make analysis task simpler. Then police officers and other interested users will be able to use this system effectively and efficiently. An efficient method for crime prediction can be developed in the future based on the results of the crime analysis system.

ACKNOWLEDGEMENT

REFERENCES

[1] H. Chen, W. Chung, J. Xu, G. Wang, Y. Qin and M. Chau, "Crime data mining: a general framework and some examples." IEEE Explore- Computer, vol. 37, no 4, pp. 50-56, 2004.

[2] P. Chamikara, D. Yapa, R.Kodituwakku and J. Gunathilake, "SL-SecureNet : intelligent policing using data mining techniques," International Journal of Soft Computing and Engineering, vol. 2, no. 1, pp. 175-180,2012

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

[4] S. Adhikari and S. Kumar, "Intelligent criminal identification system," in the 8th International Conference on Computer Science and Education, Colombo, Sri Lanka, 2013, pp. 633-638

[5] K. R. Rahem, and N. Omar, "Drug-Related crime information extraction and analysis," International Conference on Information Technology and Multimedia(ICIMU), Malaysia, IEEE, November 2014, pp.250 – 254.

[6] N. T. Elyezjy, M. A. Elhaless, "Investigating crime using text mining and network analysis ," International Journal of Computer Applications, 2015, Vol. 126, No.8, pp. 19 -25.

[7] V. Sharma, R. Kulshreshtha, P. Singh, N. Agarwal, A. Kumar   "Analyzing newspaper crime Reports for identification of safe transit paths," NAACL-HLT 2015 Student Research Workshop (SRW), pp 17-24.

[8] https://en.wikipedia.org/wiki/Big_data.