Identification of Safest Path Using Spatio-Temporal Analysis

Mentor: Prof. Amitabha Mukherjee

Priyanka Harlalka [1] *, Puneet Singh [1] †

1 Indian Institute Of Technology, Kanpur

Abstract: In the present situation, where every day we come across crimes, it is of prime importance to ensure ones safety and any measure which could ensure ones safety is certain to pay for itself. In this project, we propose a method to find the safest path between two locations, based on geographical models of crime intensities. We consider the police records and news articles as the basis for our calculations. It is essential to consider news articles as there is a significant delay in updating crime records. We address this problem by updating the crime intensities based on current news feeds. Based on the updated crime intensities, we identify the safest path. It is this real time updation of crime intensities which makes our model way better than the models that are presently in use.

1. Introduction

In todays society criminal activities are on the rise. Newspapers each day are full of news articles shrieking about crime incidences from all corners of the world. Governments all over the world are spending millions trying to curb this menace. We have come up with a way by which one can ensure that he travels from one place to the other by the safest route possible. This data mining paradigm we have developed mines essential information from police FIR records and Newspaper articles to provide the user with the most up to date information of the safest path and crime hotspots of any region. We use the police FIR records to obtain slightly outdated prior information of crime intensities of a particular region. Then we take the news articles. The first challenge is to identify whether the given news article is a crime related article or not. This is achieved by using the technique of Latent Semantic Analysis[1].Once that is done the next challenge we face is to identify the location the crime occurred at from the news article. This is achieved by using Geofuzzy tagging[4].Then we use the information so obtained to update the crime intensities use temporal analysis for predicting the future crime intensities at a particular location.

* *E-mail: priyah@iitk.ac.in, Roll: 11542*

† *E-mail: puneets@iitk.ac.in, Roll: 10548*

2. Literature Review

In this work, we have defined crime to be any activity that affects the safety of a person traveling through an area. For our purpose, it is sufficient to find articles that pertain to relevant criminal activity (as opposed to irrelevant instances of crimes like cyber crime), and find the distribution of such crimes across the city. Traditional approaches to crime data mining focus on finding relations between attributes of the crimes, or finding hot spots from a set of crime incidents.

In [2] the group has created a mining model to find a decision tree based on attributes from criminal cases. This binary tree is built by selecting the attribute which maximizes knowledge gain at each node by choosing children with minimum entropy. It can be used to find relevance of suspects with cases, by making decision rules from this tree. In [6], the group has proposed an algorithm, Series Finder, to detect patterns of crime committed by the same individual(s) within a database, by iteratively adding crimes to pattern 'seeds' based on a similarity criterion over attributes.

All these methods, while effective in finding patterns within records, are not targeted at predicting trends from past occurrences. A step further in this direction is crime forecasting, which was presented in [7]. The group developed a model in collaboration with police, aiming at predicting the location, time and likelihood of future residential burglary based on temporal and spacial (over grids) information taken from police records. Various classification techniques were used by the authors to develop a model that best relates attributes to crimes. This is pertinent to our model, as the group has investigated data mining techniques to forecast crime. They have focused on residential burglary in this work, and have planned to extend to other types of crime in the future. An interesting idea is division of the city into a grid, which is an intuitive method of quantizing the locations. In our model, we have assumed police stations to be a strong indicator of population (and consequently crime) density, and have mapped each locality to it's police station. Perhaps the most relevant work is done in [5], where the group approaches the problem of identifying patterns in combined data sources by inferring clusters from spatial and temporal distribution.

3. Classification of Articles

The news articles picked from newspaper websites are hardly annotated. Besides even if they are annotated we are concerned only with crimes which affect safety of a person travelling through that region. For example cybercrimes, suicides etc. do not affect the safety of a person travelling through a region and should not be classified as crime articles by the model. Hence to classify the news article as crime or non-crime we first build a training corpus of news articles and manually annotate them as crime and non-crime. We then apply stemming on these articles to remove the stop words like a, an, the etc. We then convert these articles to a term-document matrix i.e. for every document we record the frequencies of the words occurring in these documents. This matrix is too large to handle conveniently so we apply Singular Value Decomposition [11] to reduce the dimensionality of the matrix. Besides reducing the dimensionality of the matrix it also ensures that similar meaning words get combined under a common term. Now we can represent each of the documents in our corpus in this reduced dimensional vector space. We repeat the same process for our test document. Then we identify the k-nearest neighbors this test document based on cosine distances. If the majority of these k neighbors are crime related then we classify the article as crime article else we classify it as a non-crime article.

4. Spatio Analysis

Geotagging is the process of identifying the geographic location of the given data. The process can be broadly classified into two categories:

1. Toponym recognition: All toponyms are identified for example, Dwarka, Chandni Chowk. This step involves the knowledge of natural language.

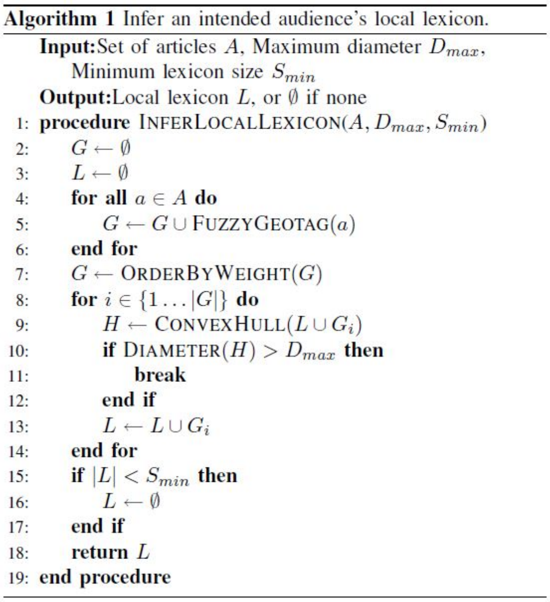
2. Toponym resolution: All the identified toponyms are assigned to correct geographic location. Documents content is vital to execute this step.

A Geotagger must be able to resolve the geo/non-geo ambiguity i.e it should be able to identify whether ”Dwarka” refers to a location or some other entity such as name of an organization or a person or it should be able to resolve whether ”Rajendra Nagar” refers to old Rajendra Nagar or new one.
We have used NER and POS tagging to tag locations and proper nouns respectively. We used stanford ner and not nltk for NER because

1. It has better result

2. Nltk writes the article into a file and runs Stanford NER command line tool to parse that file and parses the output back to python. So, overhead of loading classifiers everytime is unavoidable.

And, TreeTagger is used for POS tagging. Extracting all possible toponyms decreases true negatives and increases false positives. So, to decrease false positives, we used google api as our gazetteer database to detect possible locations or toponyms from ner and pos. This decreases the false positives and also resolves the geo/non-geo ambiguity. Now, to resolve the geo/geo ambiguity, we used fuzzy geotagging to determine sets of possible interpretations of all toponyms as propsed by Lieberman[4]. Then, assigned each interpreted police station, weights propotional to its prior probability.



```
Algorithm 1 Infer an intended audience's local lexicon.
  Input: Set of articles $A$, Maximum diameter $D_{max}$,
      Minimum lexicon size $S_{min}$
  Output: Local lexicon $L$, or $\emptyset$ if none
  procedure INFERLOCALLEXICON($A, D_{max}, S_{min}$)
      $G \leftarrow \emptyset$
      $L \leftarrow \emptyset$
      for all $a \in A$ do
          $G \leftarrow G \cup$ FUZZYGEOTAG($a$)
      end for
      $G \leftarrow$ ORDERBYWEIGHT($G$)
      for $i \in \{1 \dots |G|\}$ do
          $H \leftarrow$ CONVEXHULL($L \cup G_i$)
          if DIAMETER($H$) $> D_{max}$ then
              break
          end if
          $L \leftarrow L \cup G_i$
      end for
      if $|L| < S_{min}$ then
          $L \leftarrow \emptyset$
      end if
      return $L$
  end procedure
```

Figure 1: Algorithm used for fuzzy geotagging, Source: Leibermann et al.[4]

5. Temporal Analysis

Having identified the location of the articles, our next task is to predict the crime rate at time t+1 given the crime rate from 0 to t. In the recent past time series forecasting has attracted researchers all over the globe. In the past, the forecasted values were considered a linear combination of the previous values and the random errors associated with them. Hence, the forecasting was done on the basis of ARIMA (Autoregressive Integrated Moving Average) models. As time progressed researchers realized that the assumption made in the ARIMA models is not practical in many cases. To overcome the shortcomings of the ARIMA models Artificial Neural Networks were introduced to predict the future values of the time series. ANNs do not consider the linearity assumption and hence were able to forecast better when the future values were a non-linear function of the past values. However, the ANNs are not able to outperform the ARIMA models when the forecasted values are a linear function of the observed values. Experimental findings as well as well theoretical proofs suggest that combination of linear models and non-linear models yield effective results. We have used the hybridized model proposed in 2011 [3], which helps in exploring the linear structure present in the existing data and the Artificial Neural Networks are used to identify the underlying non-linear component of the data. The ARIMA (p,d,q) model can be written as:

$$\phi(B)\nabla^d y_t = \theta(B)\epsilon_t$$

Artificial Neural Networks do away with the linearity assumption of the ARIMA models as the significant advantage that they have is that they can approximate large number of functions with a relatively high degree of accuracy. For the artificial neural networks the mathematical relationship between the forecasted value and the observed value is as follows:

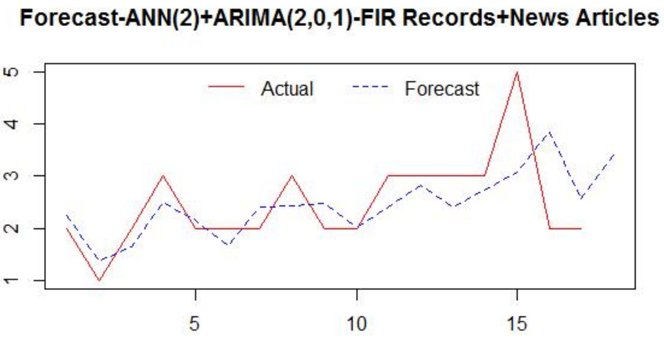
$$y_t = w_0 + \sum_{j=1}^{Q} w_j g(w_{0j} + \sum_{i=1}^{P} w_{i,j} y_{t-i}) + \epsilon_t$$

In this proposed approach of [3] the forecasted values are considered to be composed of a linear part and a non-linear part. For the linear part of the value ARIMA modeling is done and ANNs are applied to the non-linear part. The non-linear part is further assumed to be composed of the following non-linear variables

$$N_{1_t} = f^1(\epsilon_{t-1}, \ldots, \epsilon_{t-n})$$
$$N_{2_t} = f^2(z_{t-1}, \ldots, z_{t-n})$$
$$N_t = f(N_{1_t}, N_{2_t})$$

where f^1, f^2 are the nonlinear functions determined by the neural network.

$$y = f(N_{1_t}, L_t, N_{2_t}) = f(\epsilon_{t-1}, \ldots, \epsilon_{t-n}, L_t, z_{t-1}, \ldots, z_{t-n})$$



6. Dataset

In our proposed system, we gather data from disparate sources such as the reports from http://delhipolice.serverpeople.com/firwebtemp/Index.aspx. This data source is very heterogeneous and does not have a common structure. Getting all related data into one homogeneous data structure is the rst task of any data mining or visualization undertaking. Missing values need to be estimated or ignored depending upon their signicance so that the effect on the models is minimized.Also the gravity of crime was ignored and just the number of crimes was used as an estimate of crime intensity of the region. Once the data from all these sources are combined into one database to have consistent semantics across all records, we then proceed to apply our machine learning and data mining algorithms to nd patterns in crime.

News articles were obtained from the various News papers websites using web crawlers. A web crawler was used on http://timesofindia.indiatimes.com/ (Times of India online portal) , http://indiatoday.intoday.in/ (India Today news portal)and http://www.ndtv.com (NDTV news portal) to get crime re- lated news articles.

7. Results

Classification	Number of articles	Accuracy
Crime	70	98.57%
Non-crime	30	73.33%

Location	Number of articles	Accuracy
163 Police Stations	110	84.54%

Temporal Analysis MSE	0.579

Number of FIRs analyzed for prior information: 7,304

1. Articles giving expected output - (1) Badarpur5.txt (2) Saket0.txt (3) Kashmere-Gate3.txt
2. Articles giving wrong output - (1) Amar-Colony5.txt (2) Fatehpur-Beri2.txt (3) GK-I0.txt
3. Articles giving expected because of POS tagging i.e. these articles would have give no output if we would have use only NER - (1) Jamia-Nagar2.txt (2) Saket2.txt (3) Saket3.txt

Refer to 'articles' folder in code directory to view the above mentioned articles.

(a)Shortest path

(b)Safest path

Figure 2: Path from Dwarka to N.F. Colony

8. Conclusion

1. In our project we plot the crime hotspots on the map of Delhi using crime intensities.
2. Geofuzzy tagging efficiently identifies the articles location and is used to update the crime intensity of a given location.

3. Temporal Analysis efficiently predicts the future crime rate of a location

4. Future work would include incorporating the actual road network to find the safest path between the locations A and B.

9. Acknowledgement

We would like to acknowledge the efforts of Dr. Amitabha Mukherjee as without his constant support and guidance this project would not have been possible.

References

[1] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[2] Ruijuan Hu. Data mining in the application of criminal cases based on decision tree. *International Journal of Engineering Sciences*, 2:24–27, 2013.

[3] Mehdi Khashei and Mehdi Bijari. A novel hybridization of artificial neural networks and arima models for time series forecasting. *Applied Soft Computing*, 11:26642675, 2011.

[4] Michael D Lieberman, Hanan Samet, and Jagan Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pages 201–212. IEEE, 2010.

[5] Aravindan Mahendiran, Michael Shuffett, Sathappan Muthiah, Rimy Malla, and Gaoqiang Zhang. Forecasting crime incidents using cluster analysis and bayesian belief networks. 2011.

[6] Tong Wang1, Cynthia Rudin1, Daniel Wagner, and Rich Sevieri. Learning to detect patterns of crime. *JECET*, 1:124–131, 2012.

[7] Chung-Hsien Yu, Max W. Ward, Melissa Morabito, and Wei Ding. Crime forecasting using data mining techniques. *IEEE 11th International Conference on Data Mining Workshops*, pages 779–786, 2011.