Analyzing Newspaper Crime Reports for Identification of Safe Transit Paths

Vasu Sharma[1] **Rajat Kulshreshtha**[2] **Puneet Singh**[1] **Nishant Agrawal**[3]
Akshay Kumar[4]

Indian Institute Of Technology,Kanpur[1] IIT,Guwahati[2] IIIT,Hyderabad[3]
VIT Chennai[4]

vasus@iitk.ac.in rk.kuls@gmail.com pun.singh92@gmail.com nash007@gmail.com
akshay.kumar2011@vit.ac.in

Abstract

In this paper, we propose a method to find the safest path between two locations, based on the geographical model of crime intensities. We consider the police records and news articles for finding crime density of different areas of the city. It is essential to consider news articles as there is a significant delay in updating police 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. Our model would also inform the user of crime sprees in a particular area thereby ensuring that user avoids these crime hot spots.

Keywords: Crime detection, Hotspot identification, Safest Path, Topic Modeling, Latent Dirichlet Allocation, Latent Semantic Analysis, Natural Language Processing.

1 Introduction

In today's society, reports of criminal activity are on the rise. Newspapers each day are replete with news articles about incidents of crime from different parts of our cities. Crime is not spread evenly across a city, the level of criminal activity varies with region. In traveling from one spot to another within a city, people naturally desire not to be a victim to criminal activity. In general, the likelihood of falling victim to criminal activity is greater in areas with elevated crime levels, hence the path one travels must preferentially avoid areas with higher levels of crime.

Our objective in this paper, is to find the safest possible path between any two points on the street map, based on actual or inferred knowledge of prior criminal activity. The map may be viewed as a graph, where junctions are vertices in the graph, and streets are edges. The problem of finding a path from an origin to a destination is simply that of finding a path between the corresponding vertices in the graph. For the purpose of this paper we have focused on the city of New Delhi, India, a city which has recently gained notoriety as being particularly unsafe for commuters especially women.

We can now cast our "safest-path" problem as a graph search problem. Each vertex and edge in the graph can be assigned a risk. The safest path between junction A and junction B is the least risky path, or, assuming the risk to be a cost, the least-cost path between the graph vertices a and b. Thus now we can restate the problem as finding the least-cost path between vertices.

Given a graph, the algorithm for finding the least-cost path between vertices is well known. We use the well known Dijkstra's algorithm (Dijkstra, 1959). The greater challenge now is that of *specifying* the graph. The structure of the graph, as mentioned earlier, is simply the street map of the city. The real challenge becomes that of assigning costs to the vertices and edges, which reflect the risk of crime in the junctions and streets they represent. We will do so by assigning the cumulative count of the number of instances of crime that were reported at any street or junction as the cost of the corresponding edge.

We do not have a direct way of assigning these costs, since detailed, updated crime information is

Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW), pages 17–24,
Denver, Colorado, June 1, 2015.

generally not available for the city. So we will try to infer this information using a variety of sources. We will use police records to assign costs based on historical data, and to compute *a priori* information for further inference. For more updated scores, we mine the newspaper reports. However mining newspaper articles is not easy, since the articles are in natural language. Moreover, they are often imprecise in locating the reported crimes and don't specify the roads or the junctions. So, we use a Bayesian formalism to determine the locations from the article.

Following the above mentioned steps, we can assign costs to our graph and thus find the safest path between any two locations. However, for simplicity we have not considered the actual road networks for finding the path, but do so based on neighborhoods, which we then map on to the road network. Our results show that we are able to infer location from newspapers reports with relatively high accuracy, and that moreover, the hypothesized paths are highly plausible.

The paper is organized as follows. Related literature is reviewed in Section 2. Section 3 presents our data collection strategy. Sections 4-7 present detailed methodology. Results and discussion are presented in Section 8. Our conclusions are presented in Section 9.

2 Literature Review

The majority of the literature on crime-data mining focuses on analyzing data and crime records to identify patterns and predict crime. (Chen et al. , 2004) propose a generic machine learning framework for various clustering and inference tasks that may be used to detect or predict crime based on observed activity. Other 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. Another approach to detect patterns within police records was presented in (Sukanya et al. , 2012), where an expert based semi-supervised learning method was used to cluster police crime records based on their attributes. Weights were introduced to the attributes, and various patterns were identified from subsets of attributes. A step further in this direction is crime forecasting, which was presented in (Yu et al. , 2011), which 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 are used 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.

For our purpose, it is sufficient to identify news articles that pertain to relevant criminal activity, and find the distribution of such crimes across the city. Our challenge, then, is to automatically identify news articles that relate to specific types of crime, and to automatically locate the crime that is reported with sufficient specificity that we can build a "path-safety map". As it turns out, little of the literature on crime-data mining actually directly relates to this task. The task of identifying news reports has its closest analog in the literature on document classification (Sebastiani , 2002), although we have not specifically encountered many that relate in particular to crime data.(Chau et al. , 2002) report on the use of machine learning algorithms to derive named entities from formal police reports, but do not specify the techniques used. As we see later from our work, we do not require sophisticated algorithms; simple classifiers can do this particular task quite effectively.

The current literature on crime mapping, e.g. (Maltz et al. , 2000) , (Leong et al. , 2000) does not significantly address the issue of generating maps from free-form text report. 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 (Mahendiran et al. , 2011), where the problem of identifying patterns in combined data sources is approached by inferring clusters from spatial and temporal distribution. Bayesian Belief Networks are used to find a probabilistic relation between crime, time and location. Creation of a "heat map" to represent unsafe areas was suggested but not part of this report. The group has collated crime reports from various websites. The distinguishing feature of our model is that we have combined crime reports and news feeds, and that we mapped our crime distribution into a graph with edges weighted according to crime intensities.

3 Data Collection

For the experiments reported in this paper we focus on Delhi, India, a city which has recently acquired some notoriety because of the spate of crimes reported from there, particularly against women. This recent notoriety particularly motivates people to try to find safe routes from origin to destination, making the solution reported in this paper especially relevant there.

In our proposed system, we gather data from disparate sources such as the reports from the Delhi Police Website [1]. Here we have ignored the gravity of crime and used only the number of crimes for allocating a cost to a location. We have used 42768 police crime records over a period of 3 years for the state of Delhi to form our historical prior. We parse the records and extract the location and type of crime from the records. We now tag the records to their nearest police station and maintain counts of the number of crimes committed in the jurisdiction area of every police station. This count is what we have considered as 'crime intensity' for that area. These are used to derive the *a priori* probability distribution of crime in the various precincts. A total of 162 locations were considered, one for each police station in Delhi.

We used a web crawler to obtain news articles from various news paper websites[2] to get crime related news articles. A total of 32000 news articles were obtained using the crawler out of which half were crime related and the other half were not crime related. These articles formed the prior for our k-nearest neighbor and LDA based approach used for classification as crime/non-crime and location identification described in the later sections.

4 Classification of Article as Crime or Non Crime

The news articles picked from news paper websites are not annotated. Besides, we are concerned only with crimes which affect safety of a person traveling through that region. For example cyber crimes, suicides , etc., do not affect the safety of a person traveling through a region and should not be classified as commuter affecting crimes by the model.

Therefore, in order to proceed with the "safety-map" generation, we must first classify the news articles as "crime" or "non-crime". We find that the language used to refer to such crime in the news articles is diverse, ranging from direct to oblique references. Even among the direct references, a variety of different vocabularies and constructs may be employed. Direct analysis of language and vocabulary may consequently require complicated classification schemes to account for all possible variations.

Instead, we work on a simpler hypothesis – we hypothesize that regardless of the manner in which the crimes are being referred to, there exist underlying *semantic* levels at which they are all similar, and that by expressing the documents in terms of their representation within these levels, we must be able to perform the requisite classification relatively simply.

Uncovering the underlying semantic structure must be performed in an unsupervised manner. A variety of statistical models such as latent semantic analysis, probabilistic latent semantic analysis (Hoffmann , 1999), latent Dirichlet allocation (Blei et al. , 2003) etc. have been proposed for this purpose. We employ a relatively lightweight, simple algorithm, latent semantic analysis (LSA)(Dumais , 2004). LSA is a singular-value decomposition (SVD) (Kumar , 2009) based statistical model of word usage that attempts to recover abstract equivalents of semantic structure from the co-occurrence statistics of words in documents. Given a collection of documents, it first composes a term-count matrix, where each column represents a document, and each row represents a particular word. The total number of columns represents the number of documents in the collection being analyzed, and the total number of rows represents the "vocabulary" of words being considered. The $(i, j)^{\text{th}}$ entry of the

[1] The police recoreds were obtained from : `http://delhipolice.serverpeople.com/firwebtemp/Index.aspx`

[2] The newspaper articles were obtained from:

- `http://timesofindia.indiatimes.com/` (Times of India online portal)
- `http://indiatoday.intoday.in/` (India Today news portal)
- `http://www.ndtv.com` (NDTV news portal)

term-count matrix represents the number of times the i^{th} word in the vocabulary occurs in the j^{th} document. The term-count matrix is decomposed using SVD. The M most significant left singular vectors recovered, corresponding to the M highest singular values, are assumed to represent the M directions of the underlying latent semantic space. Any document can be represented in this space as the projection of the term-count vector of the document (comprising a vector of counts of the words from the vocabulary in the document) onto the set of M singular vectors. The projection is assumed to exist in the corresponding semantic space.

To compute our model, we first stem our corpus, and eliminate all stop word such as "a", "an', "the', etc. We compose a term-document matrix from the documents, and employ LSA to reduce the dimensionality of the data. All documents are represented in the lower dimensional semantic space.

We annotate our training data instances to identify if they belong to the "crime" category or not. Subsequently, given any test document, we use a k-nearest neighbor classifier to classify it: we identify the k closest training instances, where closeness is computed based on cosine distance. If the majority of the k instances are crime-related, we classify the article as a crime article, otherwise we classify it as non-crime.

5 Identification of Location of the Article

After identifying crime-related articles, we must next identify the location where the reported crime occurred. Again, we observe that newspaper articles often do not make explicit identification of the location of the crime, often not providing more than city-level information explicitly. The exact location must be inferred from the text used to describe the area, and sometimes from other incidental information that the articles may contain. Identification of the location thus becomes a challenging problem. Unlike the problem of identifying that the article refers to a crime, this is a closed-set problem in that the reported crime has indeed occurred, and hence *must* have occurred in one of the areas of the city. Thus, we only need to identify which of the various locations in the city was the spot of occurrence of the crime. We do so by a combination of methods. In the First, we employ a named-entity extractor to identify potential location-related words from the document, in case the location may be inferred from direct references. Then we use a Naive Bayes classifier based on a representation derived from latent Dirichlet allocation analysis (Blei et al. , 2003) of the articles to identify the location. We describe both below.

5.1 Named Entity Recognition

Named Entity Recognition (Klein et al. , 2003) is a Natural Language Processing technique which can identify named entities like names, locations, organizations etc. from text. Specifically, we use the technique described in the aforementioned work, to identify locations from articles. It uses decision trees and Conditional Random Fields(CRF's) (Wallach , 2004) to identify named entities. Conditional random fields (CRFs) are a class of statistical modeling method often applied in pattern recognition and machine learning, where they are used for structured prediction. Whereas an ordinary classifier predicts a label for a single sample without regard to "neighboring" samples, a CRF can take context into account; e.g., the linear chain CRF popular in natural language processing predicts sequences of labels for sequences of input samples. Given the nature of our problem we determined this technique to be most appropriate for our data.

5.2 LDA-based Naive Bayes for Location Determination

Named entity recognition cannot pull up location information when it is not actually specified in the article. Even when it is mentioned, the reference may not be unambiguous. In order to deal with such articles we use a simple Naive Bayes classifier based on features derived using Latent Dirichlet Allocation (LDA) (Blei et al. , 2003).

LDA is a well-known document-analysis technique which assumes a 'latent' or underlying pattern in the pattern of words in it. The model assumes that documents are composed of topics. Topics are distinguished by the probability distributions of words associated with the topic – different topics have different distributions over words. For instance, a sports-related topic may have a higher prevalence of sports-related words, while a politics-related topic will have a higher prevalence of politics-related words. The generative model for LDA assume that in order to compose a document, for each word in the document a topic is selected according to a document-specific probability distribution over topics, and subsequently a word is drawn from the topic. Mathematically, the collection of words $\{w \in D\}$ in any article A are assumed to have been drawn from a distribution $P(w|t;\theta)P(t|A)$, where $P(t|A)$ represents the probability distribution over topics t within article, and $P(w|t)$ is the probability distribution of words within topic t. The probability distributions $P(t|w)$ are learned from training data. The probability distribution $P(t|w)$ of topics within any document is also drawn from an *a priori* Dirichlet distribution, the parameters of which are also learned from training data.

We employ the distribution over topics as the fundamental characterization of documents. We derive a set of T topics from a training corpus comprising crime-related news reports. Every article A is now decomposed into these topics. The probability distribution $P(t|A)$ of topics t in the article, which is derived using LDA, is now used as a representation for the documents.

We view each document as bag of topics, and $P(t|A)$ as a normalized count of the number of times the topic appears in the document. Now we cast the location classification problem as follows.

We associate locations with police stations. The city is partitioned into regions, one corresponding to the jurisdiction of each station. We tag a number of training articles with the location of the crime they report. We ensure that every station is adequately represented in the training set. Each article is now decomposed into a topic histogram $P(t|A)$.

We now compute a probability distribution of topics with respect to each location to be identified using the following maximum likelihood estimator:

$$P(t|L) = \frac{1}{|\{A \in L\}|} \sum_{A \in L} P(t|A)$$

where $A \in L$ represents the set of all training articles that refer to crimes in location L.

In order to appropriately represent the natural bias of crime in the city, we derive *a priori* probability distribution of crime in the various precincts, $P(L)$ from historical police FIR records as

$$P(L) = \frac{|C \in L|}{\sum_L |C \in L|}$$

where $C \in L$ represents the set of all FIR records of crimes reported at location L.

We can now apply the following Bayesian classifier to identify the location $\hat{L}(A)$ of the crime reported in any article A:

$$\hat{L}(A) = \arg\max_A P(L|A) \tag{1}$$

In other words, we are assigning the crime to the location that is most probable *a posteriori*, given the information in the article.

Using the usual modification of the above equation, the classification reduces to

$$\hat{L}(A) = \arg\max_A P(A|L)P(L) \tag{2}$$

and working in the log domain, taking into account the monotonicity of the $\log$ function:

$$\hat{L}(A) = \arg\max_A \log P(A|L) + \log P(L) \tag{3}$$

$\log p(L)$ in the above equation is directly obtained from the *a priori* probability distribution $P(L)$. We only need to compute $\log P(A|L)$ to perform the computation in Equation 3. To do so, we assume that the article being classified has been obtained by drawing topics from the location specific topic distribution $P(t|L)$ repeatedly. This leads us to the following equation for $P(A|L)$.

$$p(A|L) = \prod_t P(t|L)^{\lambda P(t|A)}$$

$$\log P(A|L) = \lambda \sum_t P(t|A) \log P(t|L)$$

where, as mentioned earlier, $P(t|A)$ is the normalized count of the times topic t in the article A, as computed using LDA. The term λ is required because we only know the *normalized* count of topic occurrence; this must be scaled to obtain the true counts. The overall classification rule thus becomes

$$\hat{L}(A) = \arg\max_A \lambda \sum_t P(t|A) \log P(t|L) + \log P(L) \tag{4}$$

In principle λ is article specific. In practice, we derive a global value of λ by optimizing over a development training set.

6 Mapping Crime Intensities

We apply the combination of the document-identification and location-detection algorithms to news articles and use it to generate a “heat map” of crime for the city. Every new incoming article that has been classified as relating to crime, and assigned to any police station, is used to increment the crime count for that station. In our work we have worked with a fixed number of articles, resulting in a fixed heat map; in practice, to prevent the entire map from being saturated, a forgetting factor must be employed to assign greater weight to more recent crimes. We associate the total crime count for any station with every junction in its jurisdiction. Crime counts for junctions that span multiple jurisdictions accumulate the counts of all the stations that cover them. This results in a crime-weighted street map that we can now use to find the safest path between locations.

7 Identifying Safest Path

Once the safety map showing the crime intensities is known, we can convert the safest path problem to a shortest path problem by modeling the edge weights as the sum of crime frequencies of the two connecting nodes. Now that we have a graph with well defined positive edge weights, we can apply Dijkstra’s algorithm(Dijkstra, 1959) to identify the shortest path which is the safest path here.

8 Results and Validation

The validation of the model is two-fold.In the first step we check the effectiveness of the classification of the article as crime or non crime. Then we check how well does the model identify the location of the article.

8.1 Result of Crime/Non Crime Classification

The test for crime/non-crime classification was done on 5000 articles (3000 crime and 2000 non-crime articles were taken) and various values of k were experimented with. The results of which are as follows:

Value of k	Accuracy	F-score
1	82.14%	0.78
3	84.86%	0.81
5	86.52%	0.82
7	87.94%	0.83
9	**89.36%**	**0.84**
11	87.60%	0.82

Table 1: Results of Classifying articles into Crime/Non-crime categories

As the experiments demonstrated the most suitable value for k was found to be 9.

8.2 Result of Identification of location

Method Used	Accuracy	F-score
NER	81.48%	0.78
LDA	79.38%	0.75
LDA+NER	**83.64%**	**0.81**

Table 2: Location Identification results

Clearly the combination of LDA and NER techniques yields the best results.

8.3 Result for Safest Path search

We did a survey for 1200 commuters to use our model for finding the safest transit path between two locations and rate the path suggested by our model on a scale of 1 to 10 based on their prior experience of commuting between these locations. We received an average rating of 8.75/10 from the 1200 users.

9 Conclusions

The model is able to predict the safest path between 2 locations to a very high degree of accuracy. The accuracy of the model depends on the correct classification of the article as crime/non crime and on the correct identification of crime’s location from article. Clearly the model achieves both of these with very high degrees of accuracy as can be seen from Tables 1 and 2. The model also maps this safest path correctly on the map and informs the user of the route he should opt for to avoid crime prone regions.

10 Assumptions used and Future Work

Our model presently doesn't take into account the actual road networks and instead gives the path from one region(represented by that region's police station) to the other based on the assumption that a region is connected directly only to it's nearest neighbors.

In the near future we plan to do away with this assumption by incorporating the actual road network in our model.

Other future work includes identifying safest paths which also take into account the time of the day and the traffic density of various routes.We also plan to identify the exact type of crime and assign different weights to different kinds of crimes in the near future.

11 Acknowledgments

We would like to acknowledge the efforts of Dr. Bhiksha Raj and Dr. Rita Singh of Carnegie Mellon University without whose constant support and guidance this project would not have been possible.

References

[Klein et al. 2003] D. Klein and J. Smarr and H. Nguyen and C.D. Manning 2003. *Named Entity Recognition with Character-Level Model. Proceedings the Seventh Conference on Natural Language Learning*.

[Dumais 2004] Susan T. Dumais. 2004. *Latent Semantic Analysis*. *Annual Review of Information Science and Technology*

[Hoffmann 1999] Thomas Hoffmann. 1999. *Probabilistic Latent Semantic Analysis Uncertainty in Artificial Intelligence*

[Blei et al. 2003] David M. Blei and Andrew Y. Ng and Michael I. Jordan 2003. *Latent Dirichlet Allocation*. *The Journal of Machine Learning Research*,

[Maltz et al. 2000] Michael D. Maltz and Andrew C. Gordon and Warren Friedman 2000. *Mapping Crime in Its Community Setting: Event Geography Analysis*. Springer Verlag.

[Sebastiani 2002] Fabrizio Sebastian 2002. *Machine learning in automated text categorization*. ACM Computing Surveys.

[Leong et al. 2000] Kelvin Leong and Stephen Chan. 2000. *A content analysis of web-based crime mapping in the world's top 100 highest GDP cities. Mapping Crime in Its Community Setting: Event Geography Analysis. Springer Verlag*

[Ku et al. 2011] Chih-Hao Ku and Gondy Leroy. 2011. A crime reports analysis system to identify related crimes . *Journal of the American Society for Information Science and Technology.*

[Deerwester et al. 1990] S. Deerwester and S. T Dumais and G W Furnas and T K Landauer and R. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science.*

[Kumar 2009] Ch. Aswani Kumar. 2009. Analysis of Unsupervised Dimensionality Reduction Techniques . *COMSIS.*

[Wallach 2004] Hanna M. Wallach. 2004. Conditional Random Fields: An Introduction . *University of Pennsylvania CIS Technical Report MS-CIS-04-21.*

[BeyondNormality 2013] BeyondNormality. 2013. Wikipedia Entry .

[Dijkstra1959] Dijkstra's, E.W. 1959. A note on two problems in connexion with graphs . *Numerische Mathematik.*

[Chau et al. 2002] Michael Chau and Jennifer J. Zu and Hisnchun Chen. 2002. Extracting meaningful entities from police narrative reports . *Proceedings of the 2002 annual national conference on Digital government research.*

[Zhang et al. 2010] Yin Zhang,Rong Zim,Zhi Hua Zhou. 2010. Understanding Bag-of-Words Model: A Statistical Framework . *International Journal of Machine Learning and Computing.*

[Wang et al. 2004] Tong Wang and Cynthia Rudin and Daniel Wagner and Rich Sevieri. 2004. Learning to Detect Patterns of Crime . *Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg*

[Hu 2013] Ruijuan Hu. 2013. Data Mining in the Application of Criminal Cases Based on Decision Tree . *International Journal of Engineering Sciences.*

[Yu et al. 2011] C.H. Yu and Max W. Ward and M. Morabito and W. Ding. 2011. Crime Forecasting Using Data Mining Techniques . *IEEE 11th International Conference on Data Mining Workshops.*

[Mahendiran et al. 2011] Aravindan Mahendiran and Michael Shuffett and Sathappan Muthiah and Rimy Malla and Gaoqiang Zhang. 2011. Forecasting Crime Incidents using Cluster Analysis and Bayesian Belief Networks .

[Nath 2006] Shyam Varan Nath. 2006. Crime Pattern Detection Using Data Mining . *Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference.*

[Bajpai 2012] Devesh Bajpai. 2012. Emerging Trends in Utilization of Data Mining in Criminal Investigation: An Overview . *Springer.*

[Sukanya et al. 2012] Sukanya, M. and Kalaikumaran, T.

and Karthik, S.. 2012. Criminals and crime hotspot detection using data mining algorithms: clustering and classification . *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET).*

[Chen et al. 2004] Chen, H. and Chung, W. and Xu, J.J. and Wang, G. and Qin, Y. and Chau, M.. 2004. Crime data mining: A general framework and some examples *. IEEE Computer.*