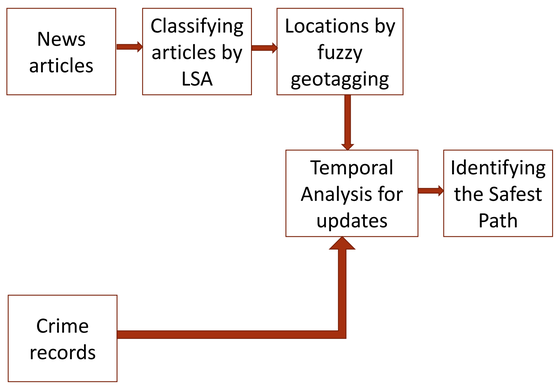
Identification of Safest Path using Spatio-Temporal Analysis

Puneet Singh[1], Priyanka Harlalka[1]
Under the guidance of: Dr. Amitabha Mukherjee[1], CSE Department
[1]Indian Institute of Technology, Kanpur

Abstract

In this project, we have proposed a method to find the safest path between two locations, based on geographical models of crime intensities. We have used the spatio and temporal modeling techniques to identify the location and relevance(based on the date) of the crime. 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 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 using Djikstra's algorithm.

Introduction



❖ **Classification of news articles**

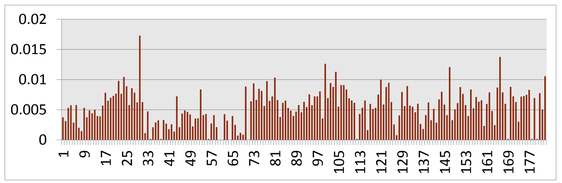
- We have used the Latent Semantic Analysis[1] for classifying articles.
- LSA essentially creates a vector representing a document.
- Classification using k-nearest neighbors with respect to cosine distances of the document vectors

$$(\mathbf{t}_i^T) \rightarrow \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix} = (\hat{\mathbf{t}}_i^T) \rightarrow \begin{bmatrix} [\ ] \\ \mathbf{u}_1 \\ [\ ] \end{bmatrix} \cdots \begin{bmatrix} [\ ] \\ \mathbf{u}_l \\ [\ ] \end{bmatrix} \cdot \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_l \end{bmatrix} \cdot \begin{bmatrix} [\ ] & \mathbf{v}_1 & [\ ] \\ & \vdots & \\ [\ ] & \mathbf{v}_l & [\ ] \end{bmatrix}$$

(with X ($\mathbf{d}_j$), U, Σ, V^T ($\hat{\mathbf{d}}_j$) above the respective matrices)

Spatio Analysis

- Statistical NER methods are not well-suited to the dynamic nature of news as noted by Stokes et.al [2]
- We use fuzzy geotagging [3] to resolve the bootstrapping problem associated with the traditional method
- In fuzzy geotagging a toponym recognition system first finds the toponyms T in an article a.



- Given a news article, we tag each word with its part of speech, using the Tree tagger, and collect all word phrases consisting of proper nouns.
- We also apply NER to the article using Stanford NER, and collect all phrases tagged as locations.
- Database of geographic locations, is then used to associate each $t \in T$ with the set of all possible interpretations R_t
- For each t and $r \in R_t$, a weight w_r is assigned to r using default sense heuristics

Algorithm 1 Infer an intended audience's local lexicon.

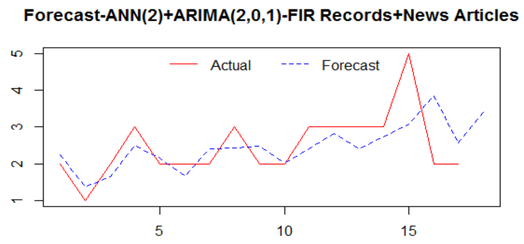
Input: Set of articles A, Maximum diameter D_{max}, Minimum lexicon size S_{min}
Output: Local lexicon L, or $\emptyset$ if none

```
procedure InferLocalLexicon(A, D_max, S_min)
    G ← ∅
    L ← ∅
    for all a ∈ A do
        G ← G ∪ FuzzyGeotag(a)
    end for
    G ← OrderByWeight(G)
    for i ∈ {1 ... |G|} do
        H ← ConvexHull(L ∪ G_i)
        if Diameter(H) > D_max then
            break
        end if
        L ← L ∪ G_i
    end for
    if |L| < S_min then
        L ← ∅
    end if
    return L
end procedure
```

Source: M.D Liebermann et. al

Temporal Analysis

- We have used a hybridization of ANNs and ARIMA models for time series forecasting[4].



- The parameters are estimated such that an overall measure of errors is minimized

$$y_t = f(L_t, N_t)$$

- A multilayer perceptron is used to model the non-linear component existing in the residuals

$$N_{1_t} = f^1(e_{t-1}, \ldots, e_{t-n})$$

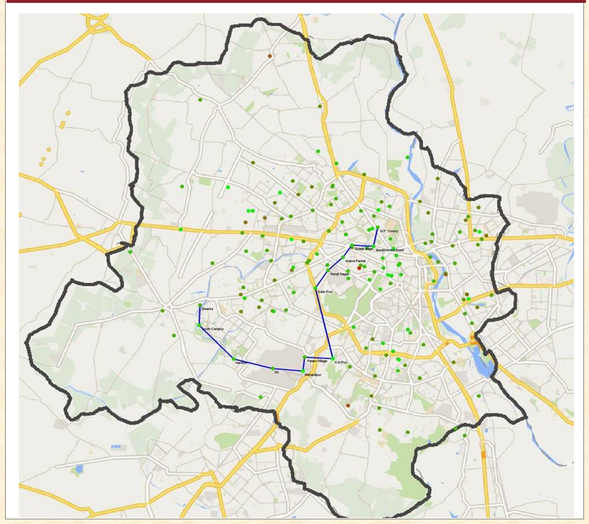
$$N_{2_t} = f^2(z_{t-1}, \ldots, z_{t-n})$$

$$N_t = f(N_{1_t}, N_{2_t})$$

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

$$y_t = f(N_{1_t}, \check{L}_t, N_{2_t}t) = f(e_{t-1}, \ldots, e_{t-n_1}, \check{L}_t, z_{t-1}, \ldots, z_{t-m_1})$$

Mapping Crime



Dataset and Results

- Delhi Police Official Website for FIR updates [5]
- News articles from NDTV , The Hindu official websites

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

Conclusion

- In our project we plot the crime hotspots on the map of Delhi using crime intensities.
- Gefuzzy tagging efficiently identifies the article's location and is used to update the crime intensity for the identified region.
- Temporal Analysis efficiently predicts the future crime rate of a location
- Future work would include incorporating the actual road network to find the safest path between two locations.

References

[1] S. T. Dumais. "Latent Semantic Anlaysis". In: Annual Review of Information Science and Technology vol. 38 (2004), pp. 188{230.
[2] N. Stokes, Y. Li, A. Moffat, and J. Rong, "An empirical study of the effects of NLP components on geographic IR performance," IJGIS, vol. 22(3), 247–264, Mar. 2008
[3] M.D Liebermann, H. Samet, J. Sankaranarayanan "Geotagging with Local Lexicons to Build Indexes for Textually-Specified Spatial Data", ICDE Conference 2010, pp: 201 – 212
[4] M. Khashei, M. Bijari, A novel hybridization of artificial neural networks and ARIMA models for time series forecasting, Applied Soft Computing (2011), pp: 2664-2675
[5] http://delhipolice.serverpeople.com