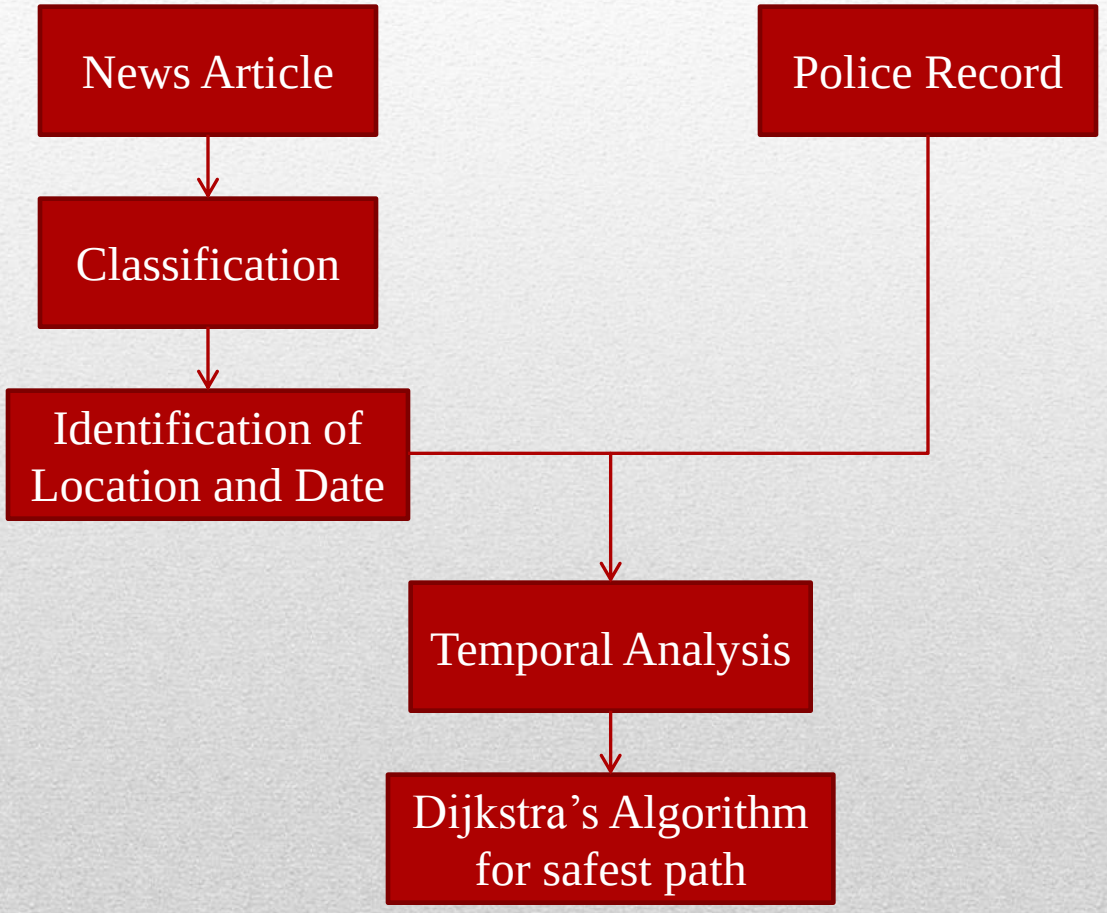
Identification of the safest path using spatio-temporal analysis

Puneet Singh (10548)
Priyanka Harlalka (11542)

Motivation

- In today's society criminal activities are on the rise
- We intend to come up with a way by which one can ensure that he travels from one place to the other by the safest route possible
- Governments all over the world are spending millions trying to curb this menace

Approach


News Article
Police Record
Classification
Identification of Location and Date
Temporal Analysis
Dijkstra's Algorithm for safest path

Classification of articles

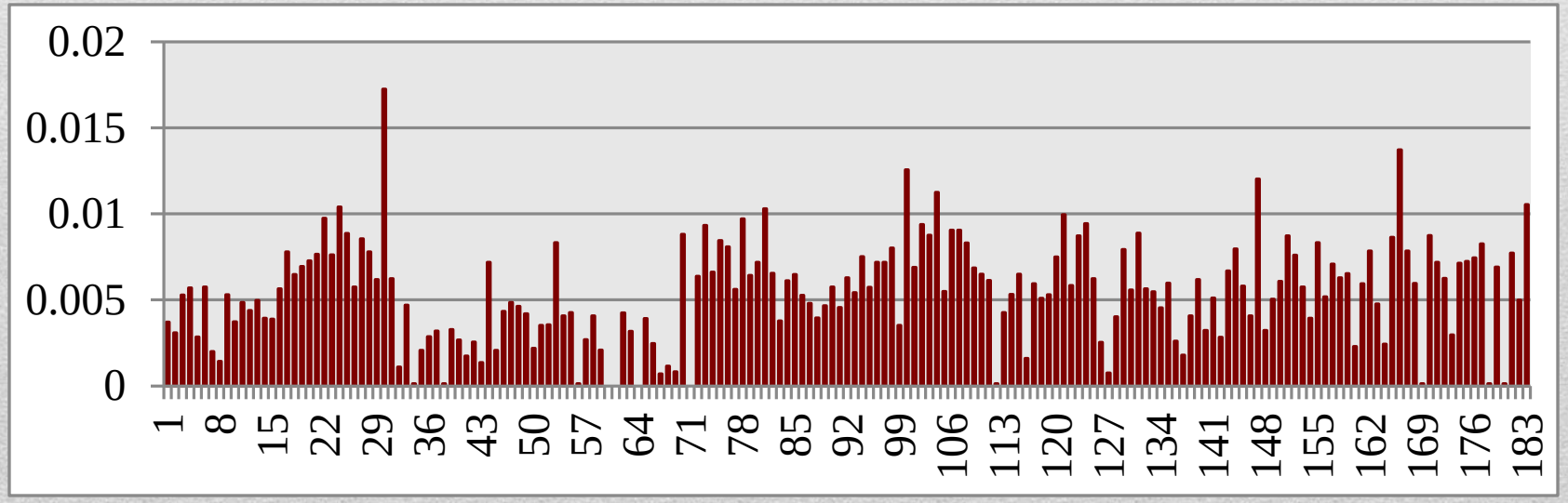
- We use the Latent Semantic Analysis[1] for classifying articles.
- LSA is essentially creating a vector representing a document.
  - Construct a term-document matrix of the corpus.

$$
\begin{array}{cccccc}
 & X & & U & \Sigma & V^T \\
 & (\mathbf{d}_j) & & & & (\hat{\mathbf{d}}_j) \\
 & \downarrow & & & & \downarrow \\
(\mathbf{t}_i^T) \rightarrow & \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} & = & (\hat{\mathbf{t}}_i^T) \rightarrow \begin{bmatrix} \begin{bmatrix} \\ \mathbf{u}_1 \\ \\ \end{bmatrix} & \dots & \begin{bmatrix} \\ \mathbf{u}_l \\ \\ \end{bmatrix} \end{bmatrix} & \cdot \begin{bmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_l \end{bmatrix} & \cdot \begin{bmatrix} [& \mathbf{v}_1 &] \\ & \vdots & \\ [& \mathbf{v}_l &] \end{bmatrix}
\end{array}
$$

- Single Value Decomposition (SVD) is then employed to reduce the dimensionality of the matrix.
- The LSA helps in grouping words with similar topics together.
- Classification using k-nearest neighbors with respect to cosine distances of the document vectors.

Identification of Location

- Statistical NER methods 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 POS tagger, and collect all word phrases consisting of proper nouns.
- We also apply NER to the article, and collect all phrases tagged as locations.
- For resolving the POS tags we use a number of heuristic rules.
- 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

Heuristic Rules

TABLE I

A SET OF HEURISTICS USED IN OUR TOPONYM RESOLUTION PROCESS.

	Heuristic	Description	Examples
$\mathcal{H}_1$	Dateline	Resolve dateline toponyms using: $\mathcal{H}_4$, $\mathcal{H}_5$, $\mathcal{H}_6$. Resolve other toponyms geographically proximate to resolved dateline.	**LONDON, Ont.** – A police... **Paris, TX** (AP) – New...
$\mathcal{H}_2$	Relative Geog.	Resolve anchor toponym using: $\mathcal{H}_1$, $\mathcal{H}_4$, $\mathcal{H}_5$, $\mathcal{H}_6$. Resolve other toponyms proximate to defined geographic point or region.	...**4 miles east of Athens, Texas.** ...lives **just outside of Lewistown**...
$\mathcal{H}_3$	Comma Group	Resolve toponym group using: $\mathcal{H}_6$, $\mathcal{H}_5$, Geographic Proximity.	...**California, Texas** and **Pennsylvania.**
$\mathcal{H}_4$	Loc/Container	Resolve toponym pairs with a hierarchical containment relationship.	...priority in **Jordan, Minn.**, ...
$\mathcal{H}_5$	Local Lexicon	Resolve toponyms geographically proximate to local lexicon centroid.	*(news source dependent)*
$\mathcal{H}_6$	Global Lexicon	Resolve toponyms found in a curated list of globally-known places.	...issues with **China**, knowing...
$\mathcal{H}_7$	One Sense	Resolve toponyms sharing names with earlier resolved toponyms.	*(article dependent)*

Source: M.D Liebermann et. al

Pseudo-Code

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

```
1:  procedure INFERLOCALLEXICON(A, D_max, S_min)
2:      G ← ∅
3:      L ← ∅
4:      for all a ∈ A do
5:          G ← G ∪ FUZZYGEOTAG(a)
6:      end for
7:      G ← ORDERBYWEIGHT(G)
8:      for i ∈ {1 ... |G|} do
9:          H ← CONVEXHULL(L ∪ G_i)
10:         if DIAMETER(H) > D_max then
11:             break
12:         end if
13:         L ← L ∪ G_i
14:     end for
15:     if |L| < S_min then
16:         L ← ∅
17:     end if
18:     return L
19: end procedure
```

Source: M.D Liebermann et. al

Temporal Analysis

- Extract the date of the news article/FIR through crawling
- We will use a hybridization of artificial neural networks and ARIMA models for time series forecasting[4].
- In an ARIMA (p, d, q) model, the future value of a variable is assumed to be a linear function of several past observations and random errors.

$$\phi(B)\nabla^d(y_t - \mu) = \theta(B)a_t$$

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

- The time series is considered as function of a linear and a nonlinear component. Thus, $y_t = f(L_t, N_t)$
- After performing ARIMA model at the first stage we assume that the residuals will contain a non-linear relationship.
- A multilayer perceptron is used to model the non-linear component existing in the residuals

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

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

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

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

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

- We will use simple Dijkstra's algorithm to find the "safest path" based on weights by temporal analysis

Dataset

1. Crime records have been extracted from the Delhi Police Website [5]
2. News articles (both crime and non crime) have been extracted from the Times Of India, Hindu etc. Website using a crawler.
3. ACE 2005 English SpatialML Annotations [6]

Result and validation

- The validation will be a three fold procedure

1. The accuracy for classification of an article as a crime/non-crime
2. Accuracy with which the location can be correctly specified on ACE 2005 dataset
3. Least Square residual for temporal analysis

Future Work

- Use actual road paths for mapping crime
- Include more sources of information for crime hotspot identification

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
6. I. Mani, J. Hitzeman, J. Richer, and D. Harris, *ACE 2005 English SpatialML Annotations*. Philadelphia, PA: Linguistic Data Consortium, 2008.

Questions/Suggestions