CRIME INCIDENTS EMBEDDING USING RESTRICTED BOLTZMANN MACHINES

SHIXIANG ZHU, YAO XIE
GEORGIA INSTITUTE OF TECHNOLOGY
ATLANTA, GA, USA
SHIXIANG.ZHU@GATECH.EDU

CREATING THE NEXT®

INTRODUCTION

WHAT IS CRIME ANALYSIS?

Georgia Tech

CREATING THE NEXT®

A fundamental and one of the most challenging tasks in crime analysis is to find **related crime series**, which are committed by the same individual or group.

Crime incident reports (a.k.a. police reports) are a large source of data that contains rich information for detecting related crime series.

According to the crime analysts, the free-text narrative contains the **most useful information** form their investigation.

Georgia Tech

Inspired by the idea of *word2vec* (word embeddings).

Find a **distributed representations** for the free-text narratives of the police reports, which map key information into a feature vector space that automatically captures the similarity of incidents.
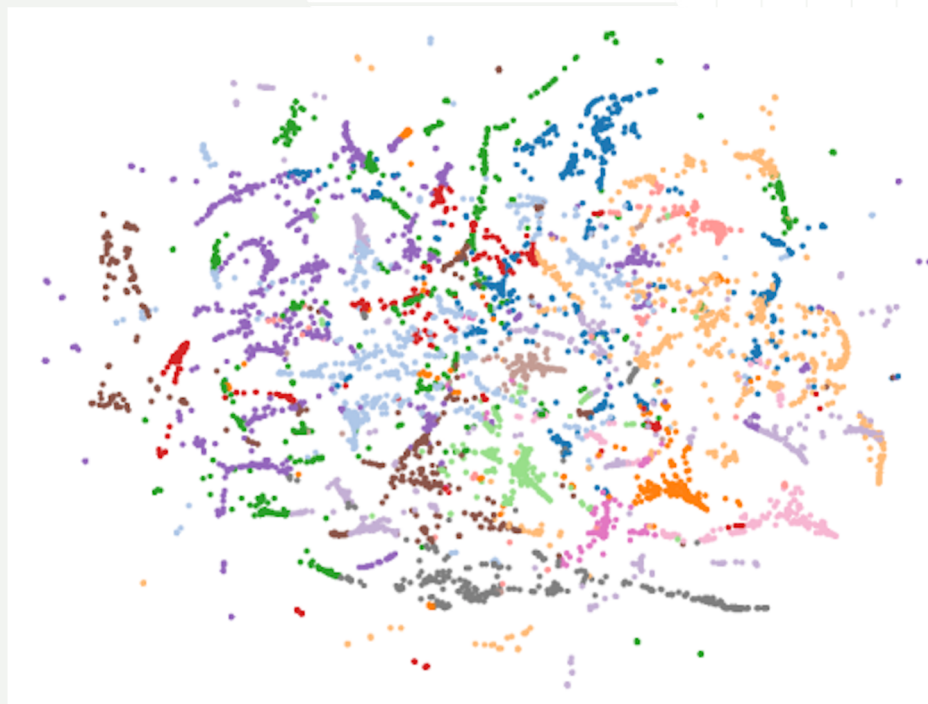
METHOD

Questioning the victims or the witnesses

Narratives of the crime report

Representations of the Bag of Words

GBRBMs feature encoder

Visible Units

Hidden Units

Projections on Feature Space

Feature Embeddings

Projections of tfidf or other conventional methods on Feature Space

Terms	tfidf	Counts
Stolen	0.00345	3
Gun	0.00121	1
Break in	0.00048	1
Jewelry	0.00725	5
... ...	... ...	... ...

| 0.23175 | 0.98147 | 0.45462 | 0.12759 |

Georgia Tech

Data Cleaning → **Tokenization (tri-gram)** → **Bag-of-Words** → **TF-IDF**

Tri-Gram Term

Denoised text: The | jewelry box | contained | mostly costume jewelry

Tokenizations:

| The jewelry box | jewelry box contained | ...... | mostly costume jewelry |

Tri-gram term i   Tri-gram term i+1   Tri-gram term i+4

What is TF-IDF weighting

In information retrieval, TF-IDF, short for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

$$\text{TF-IDF}_{t,d} = \text{TF}_{t,d} \times \text{IDF}_t \qquad (1)$$

$$\text{TF}_{t,d} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \qquad \text{IDF}_t = log\,\frac{N}{|\{d \in D : t \in d\}|}$$

Burglary happened in Buckhead Area

- mrs (1.869054)
- home (1.614854)
- drawers (1.589252)
- jewelry (1.491565)
- house (1.321998)
- residence (1.198386)
- door (1.101752)
- drawer (1.079033)
- burglary (1.010357)
- silver (0.990514)
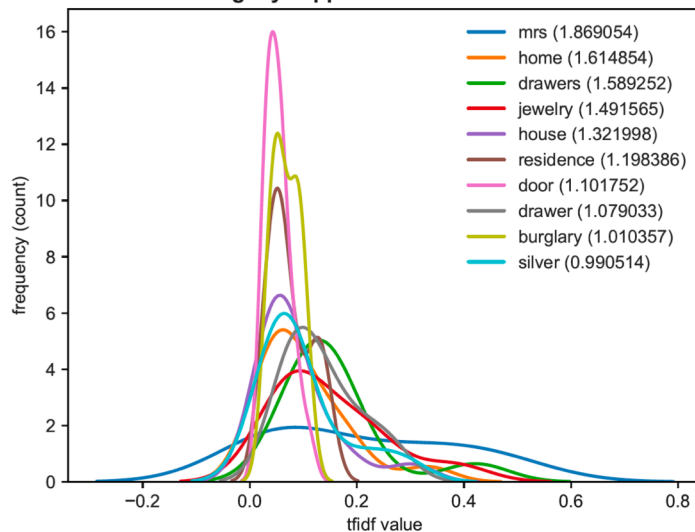
Pedestrian Robbery Committed by Suspect A

- pizza (1.583869)
- delivery (0.779159)
- black (0.643182)
- males (0.611581)
- male (0.528808)
- vehicle (0.473574)
- suspect (0.458568)
- ms (0.449708)
- scene (0.432278)
- give (0.404787)

Georgia Tech
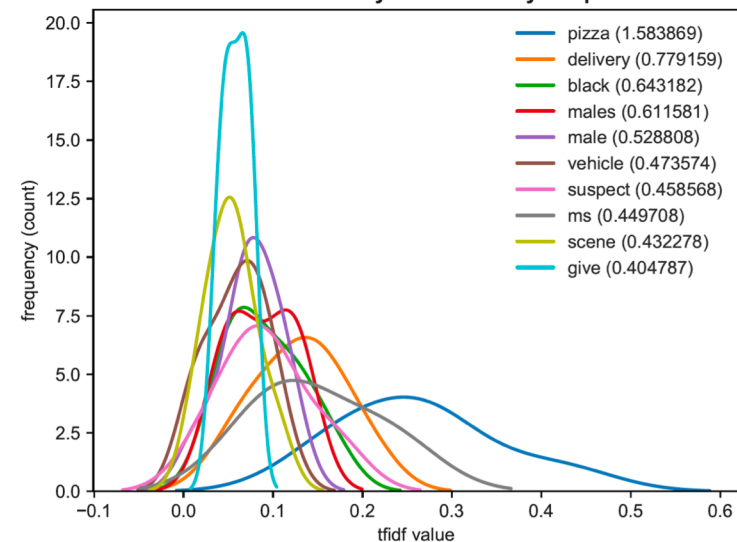
Energy Function

$$E(v, h) = -\sum_{i,j} w_{ij} h_j \frac{v_i}{\sigma_j} - \sum_i \frac{(v_i - b_i)^2}{2\sigma^2} - \sum_j h_j c_j.$$

Independency

$$p(v|h) = \prod_{i=1}^{m} p(v_i|h), \quad p(h|v) = \prod_{j=1}^{n} p(h_j|v).$$

$$p(h_j = 1|v) = \sigma(\sum_{i=1}^{m} w_{ij} \frac{v_i}{\sigma_i} + c_j)$$

Bernoulli

Hidden Layer h_1 h_2 h_3 h_4

Visible Layer v_1 v_1 v_2 u_1 u_2 u_3 u_4

Binomial   Bernoulli   Gaussian
(i.i.d. Bernoulli replicas)
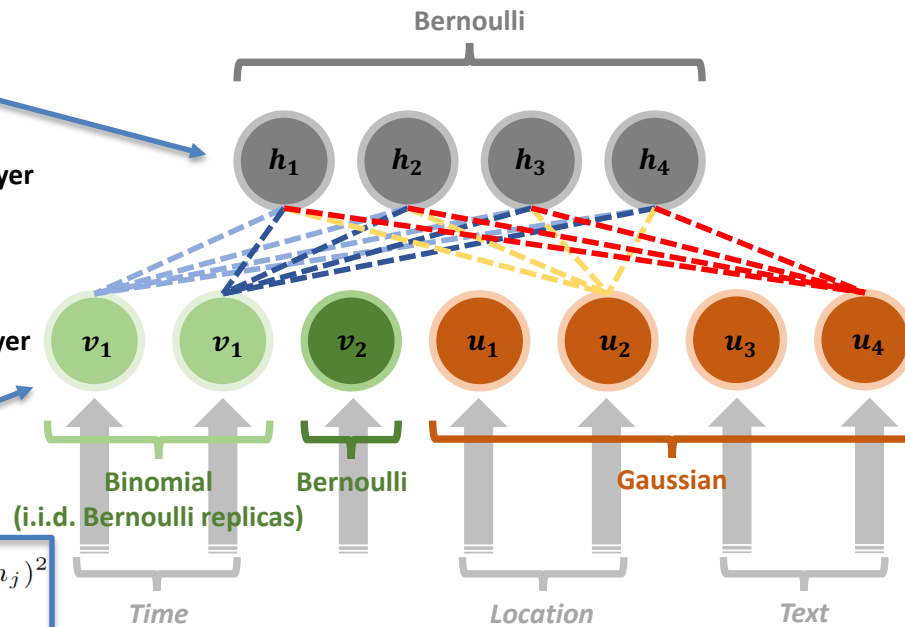
$$p(v_i = v|h) = \frac{1}{\sigma_i \sqrt{2\pi}} \cdot e^{-\frac{1}{2\sigma_i^2}(v_i - b_i - \sigma_i \sum_{j=1}^{n} w_{ij} h_j)^2}$$

Time   *Location*   *Text*

Likelihood

$$p(v) = \sum_h p(v, h) = \frac{\sum_h e^{-E(v,h)}}{\sum_{v,h} e^{-E(v,h)}}.$$

Contrastive Divergence

$$\mathrm{CD}_k(\theta, v^{(0)}) = -\sum_h p(h|v^{(0)}) \frac{\partial E(v^{(0)}, h)}{\partial \theta}$$
$$+ \sum_h p(h|v^{(k)}) \frac{\partial E(v^{(k)}, h)}{\partial \theta}.$$

A. Fischer and C. Igel, "An Introduction to Restricted Boltzmann Machines," Lecture Notes in Computer Science: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications , vol. 7441, pp. 14–36, 2012.

T. Tieleman, "Training restricted Boltzmann machines using approximations to the likelihood gradient," Proceedings of the 25th international conference on Machine learning - ICML '08 , pp. 1064–1071, 2008.

G. E. Hinton, "Training products of experts by minimizing constractive divergence," Neural Computation , 2002.

EXPERIMENTS

Table 1. Details of the test data

Id	Number	Category
Crime Series 1	8	*Robbery at Residence*
Crime Series 2	7	*Robbery at Gas Station*
Crime Series 3	4	*Pedestrian Robbery*
Crime Series 4	15	*Attempt Auto Theft*
Crime Series 5	22	*Burglary*
Random Cases	441	*Over 89 Categories*
Total	497	

The dataset contains **five hand-labeled crime series** that were identified as committed by five individual arrestees, and **441 randomly selected irrelevant crime cases**.

Table 2. The comparison of the training time.

Methods	Training Time
GBRBMs with 1000 units	*∼ 2 mins*
GBRBMs with 2000 units	*∼ 3 mins*
GBRBMs with 5000 units	*∼ 7 mins*
LDA with 1000 topics	*∼ 5 mins*
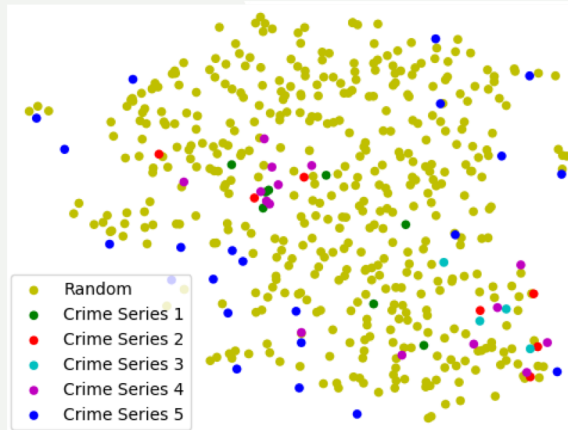
LDA (1000 topics)
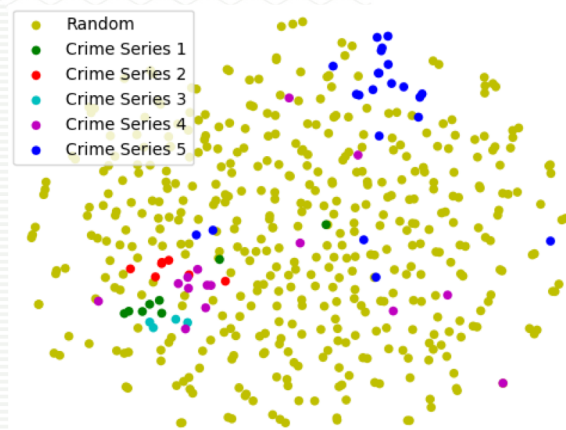
GBRBM (1000 nodes)
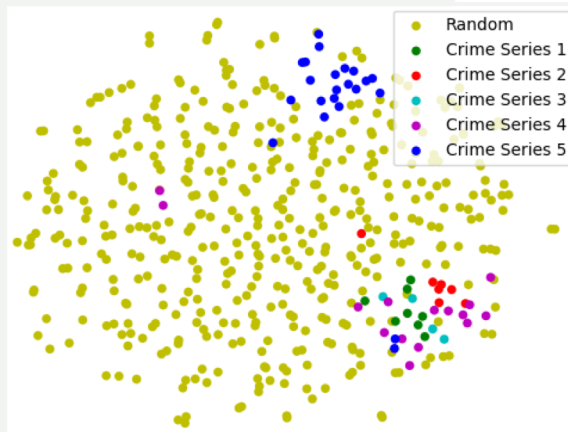
GBRBM (2000 nodes)

GBRBM (5000 nodes)

Georgia Tech
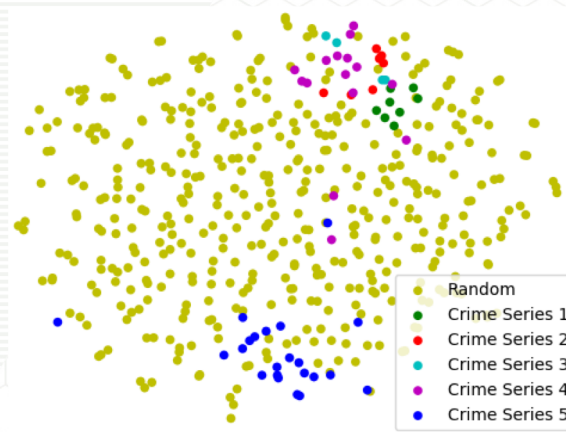
➤ We have presented a novel approach for detecting crime series that are related, using embedding found by the **Gaussian-Bernoulli Restricted Boltzmann Machine** (GBRBM).

➤ The GBRBM tends to map related cases (that share certain correlation in the raw feature) into features that are in the **vicinity in the Euclidean space**.

➤ Our methods demonstrate very promising results on **real police data** and demonstrated that the feature embeddings can have advantages over the conventional text processing methods on detecting crime series in certain cases.

➤ Ongoing work is to develop an **online** crime series detection algorithm on a **larger** and **self-increasing** real dataset.

Thank You