UNCOVERING THE PLOT: DETECTING SURPRISING COALITIONS OF ENTITIES IN MULTI-RELATIONAL SCHEMAS

Hao Wu
Jilles Vreeken
Nikolaj Tatti
Naren Ramakrishnan
**Motivation**

Knowledge discovery from multi-relational data

Intelligence Analysis

- Persons
- Locations
- Phone Numbers
**Motivation**

Knowledge discovery from multi-relational data

Biological knowledge discovery

DNA

Proteins

Pathways

...
MOTIVATION

Automatically discover *surprising* multi-relational “3C” (coalitions, connections, & chains) patterns.
STRUCTURED AND UNSTRUCTURED DATA

We consider **two** types of input data, or ‘pattern spaces’
STRUCTURED AND UNSTRUCTURED DATA

We consider **two** types of input data, or ‘pattern spaces’ *by using a trick*
PATTERNS

**Bicluster:**
connected entity set
Patterns

Bicluster:
connected entity sets

Redescription:
bicluster pair identifying
(roughly) the same
entities for shared domain
**Patterns**

**Bicluster:**
connected entity sets

**Redescription:**
bicluster pair identifying (roughly) the same entities for shared domain

**Bicluster Chain:**
A chain of redescriptions
Bicluster Chains
SURPRISING PATTERNS

‘Just mine biclusters!’ – nope.

‘Just mine redescriptions!’ – better, but still nope.

We are after chains of biclusters, such that plots in the data are revealed

and, we want only those chains that stand out from what we already know
**RELATED**

Maximal Completely Connected Subgraphs

- Spyropoulou & De Bie (2011)

multi-relational database → \( \rightarrow \) K-partite graph → \( \rightarrow \) MCCS *surprising* wrt margins
CONNECTING TO MCCS

We mine chains of *redescriptions*

probabilistic model of the data

entity-entity graph or document-entity db

Chain of *redescriptions* surprising wrt margins and all mined chains
**MOTIVATION**

Automatically discover *surprising* multi-relational “3C” (coalitions, connections, & chains) patterns.

[Diagram showing connections between various locations and contacts with phone numbers]
Iterative Mining

Knowledge *changes* during data analysis

- **interestingness** of chains changes depending on what results we study/reject

Static ranking of results is overly simplistic

- leads to redundancy – hides interesting results

*How can we score results based on (accumulated) background knowledge?*

*What prior should we use?*
Maximum Entropy Modelling

‘the best distribution $p^*$ satisfies the background knowledge, but makes no further assumptions’

very useful for data mining: unbiased measurement of subjective interestingness

(Jaynes 1957; De Bie 2009)
**MaxEnt For Binary Data**

Tiles
- A tuple of row IDs and column IDs from the given binary data matrix \( D \).
- Frequency of a Tile

\[
\gamma_T = fr(T; D) = \frac{1}{|\sigma(T)|} \sum_{(i,j)\in\sigma(T)} D(i, j)
\]

where \( D(i, j) \) represents the \((i, j)\) entry in \( D \), and \( \sigma(T) \) represents the set of all the entries in tile \( T \).
MaxEnt For Binary Data

Needed: MaxEnt model for tiles

- we use the model by Tatti & Vreeken (2011), De Bie (2011)

\[ p^*_T = \arg \max_{p \in \mathcal{P}} H(p) \]

where

\[ \mathcal{P} = \{ p \mid fr(T; p) = \gamma_T, \forall T \in \mathcal{T} \} \]

\[ H(p) = - \sum_{D \in \mathcal{D}} p(D) \log p(D) \]

\[ fr(T; p) = \frac{1}{|\sigma(T)|} \sum_{(i,j) \in \sigma(T)} p((i, j) = 1) \]
BACKGROUND KNOWLEDGE

Background information in terms of Tiles

- \( \mathcal{T}_{\text{col}} \): a set of column margin tiles
- \( \mathcal{T}_{\text{row}} \): a set of row margin tiles \textit{per entity domain}
- \( \mathcal{T}_{\text{dom}} \): a set of entity domain tiles

\[
\mathcal{T}_{\text{back}} = \mathcal{T}_{\text{row}} \cup \mathcal{T}_{\text{col}} \cup \mathcal{T}_{\text{dom}}
\]
MEASURING SURPRISINGNESS

Evaluating a bicluster chain

1) Convert the chain into a set of tiles (depends on data model, see paper)
2) Infer the MaxEnt model
3) Calculate surprisingness through divergence

\[ s_{global}(B) = KL(P_B \parallel P_{back}) \]

\[ s_{local}(B) = - \sum_{T \in T_B} \sum_{(i,j) \in \sigma(T)} \log p^*((i, j) = D(i, j)) \]
**GLOBAL VS LOCAL SCORE**

\[ s_{local}(B) = - \sum_{T \in T_B} \sum_{(i, j) \in \sigma(T)} \log p^*((i, j) = D(i, j)) \]

| .73 | .94 | .82 | .89 | .82 | .46 | .73 | .61 |
| .58 | .88 | .70 | .80 | .70 | .30 | .58 | .45 |
| .73 | .94 | .82 | .89 | .82 | .46 | .73 | .61 |
| .30 | .70 | .42 | .55 | .42 | .12 | .30 | .20 |
| .30 | .70 | .42 | .55 | .42 | .12 | .30 | .20 |
| .44 | .80 | .56 | .69 | .56 | .19 | .44 | .31 |
| .44 | .80 | .56 | .69 | .56 | .19 | .44 | .31 |
| .18 | .54 | .27 | .39 | .27 | .06 | .18 | .11 |
| .30 | .70 | .42 | .55 | .42 | .12 | .30 | .20 |

\[ s_{global}(B) = KL(P_B \parallel P_{back}) \]

| .86 | .98 | .92 | .85 | .77 | .40 | .67 | .55 |
| .75 | .96 | .85 | .73 | .62 | .24 | .50 | .37 |
| .86 | .98 | .92 | .85 | .77 | .40 | .67 | .55 |
| .44 | .85 | .60 | .42 | .30 | .08 | .21 | .13 |
| .20 | .63 | .31 | .61 | .48 | .15 | .36 | .25 |
| .30 | .76 | .45 | .74 | .63 | .25 | .50 | .37 |
| .30 | .76 | .45 | .74 | .63 | .25 | .50 | .37 |
| .11 | .47 | .19 | .45 | .32 | .09 | .22 | .15 |
| .20 | .63 | .31 | .61 | .48 | .15 | .36 | .25 |
SEARCHING GOOD CHAINS
Super Naïve Strategy:

1) Mine all the biclusters!
2) Construct all the chains!
3) Evaluate all subsets of $k$ chains!
4) Choose the most surprising set.
SEARCHING GOOD CHAINS
Slightly Less Naïve Strategy:

1) Mine all the biclusters!
2) Construct all the chains!
3) While not yet chosen $k$ chains:
   evaluate each chain $C$ against $P_{back}$
   greedily choose most surprising $C$
   $back \leftarrow back + C$, and infer $P_{back}$
SEARCHING GOOD CHAINS

Our strategy:

1) Mine all the biclusters!

2) while not yet mined $k$ chains:
   find most surprising bicluster $B_0$,
   while there is a redescription $B_i$ of $B_{i-1}$
   add most surprising $B_i$ to chain
   $back \leftarrow back + C$, and re-infer $P_{back}$
## EXPERIMENT RESULTS

### Datasets Statistics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Documents</th>
<th>Number of Entities</th>
<th>Doc–Entity</th>
<th>Entity–Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic 1k</td>
<td>1000</td>
<td>1000</td>
<td>0.01 — 0.05</td>
<td>0.01 — 0.05</td>
</tr>
<tr>
<td>Synthetic 2k</td>
<td>2000</td>
<td>2000</td>
<td>0.01 — 0.05</td>
<td>0.01 — 0.05</td>
</tr>
<tr>
<td>Synthetic 3k</td>
<td>3000</td>
<td>3000</td>
<td>0.01 — 0.05</td>
<td>0.01 — 0.05</td>
</tr>
<tr>
<td>Synthetic 5k</td>
<td>5000</td>
<td>5000</td>
<td>0.01 — 0.05</td>
<td>0.01 — 0.05</td>
</tr>
<tr>
<td>Synthetic 10k</td>
<td>10000</td>
<td>10000</td>
<td>0.01 — 0.05</td>
<td>0.01 — 0.05</td>
</tr>
<tr>
<td>Atlantic Storm</td>
<td>111</td>
<td>716</td>
<td>0.0179</td>
<td>0.0261</td>
</tr>
<tr>
<td>Crescent</td>
<td>41</td>
<td>284</td>
<td>0.0425</td>
<td>0.0357</td>
</tr>
<tr>
<td>Manpad</td>
<td>47</td>
<td>143</td>
<td>0.0299</td>
<td>0.0385</td>
</tr>
</tbody>
</table>

min %1s | max %1s | min %1s | max %1s
---|---|---|---
0.01 | 0.05 | 0.01 | 0.05
0.01 | 0.05 | 0.01 | 0.05
0.01 | 0.05 | 0.01 | 0.05
0.0179 | 0.0608 | 0.0357 | 0.136
0.0299 | 0.0714 | 0.0385 | 0.136
EXPERIMENT RESULTS

First things first: Synthetic Data
- can we uncover the plot?
Second things second: Synthetic Data
- can we tell when to stop?
EXPERIMENT RESULTS

Runtime Performance

Background model training time

Total time
EXPERIMENT RESULTS

- Global Score vs. Local Score

![Graph showing the relationship between global score and local score for different matrix sizes and values of \( \alpha \).]
EXPERIMENT RESULTS

Real Data

Intelligence Analysis Dataset: Crescent
EXPERIMENT RESULTS

Iterative Knowledge Discovery

\( C_1: \) Arnold C., Ryder, Kamel J., Mr. C. 
- Ralph T., who is a member of Aryan Militia, bought weapons and sells them to George W. (Muhammad J.) who is a member of Al-Qaeda.
- Ralph T. meets Kamel J. in Atlanta, Georgia, and Kamel J. drives a truck from Atlanta to St. Paul, Minnesota. He probably transports weapons.

\( C_2: \) Kamel J., Ralph T., George W., Mohammad H.
- Arnold C., Rudy C., Kamel J. to Mohammad H., U-Haul, State University.
- Atlanta, Georgia, NSA, FBI.
- Aryan Militia, FBI.
- Ralph T., FBI, George W.

\( C_3: \) ...
- Arnold C. (Abu H.), who was a suspect of the 9/11 attack and spent time in Afghanistan, rents a U-Haul truck and drives it from Boulder, Colorado to Los Angeles. He probably transports the weapons.
- Homer W., who is a member of Aryan Brotherhood of Colorado, sells the weapons to John H., who is a member of Al-Qaeda, in Colorado.
CONCLUSION

- Applicable to analyze multi-relational unstructured or discrete data

- Discover surprising entity coalitions with new data modeling primitives and algorithms

- Experiments on both synthetic and real datasets show that elaborate ‘plots’ can be detected

- Support human-in-loop iterative knowledge discovery
Thanks!

- Applicable to analyze multi-relational unstructured or discrete data

- Discover surprising entity coalitions with new data modeling primitives and algorithms

- Experiments on both synthetic and real datasets show that elaborate ‘plots’ can be detected

- Support human-in-loop iterative knowledge discovery