Advanced data mining, link discovery and visual correlation for data and image analysis

Prof. Boris Kovalerchuk

Borisk@cwu.edu

Tutorial Notes

International Conference on Intelligence Analysis (IA ‘05)
Sheraton Premiere at Tysons Corner
McLean, VA
2 May 2005

http://analysis.mitre.org
# Agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:00 pm</td>
<td>Introduction (motivation, concepts, process)</td>
</tr>
<tr>
<td>1:30 pm</td>
<td>Numeric data mining: capabilities and limitations for IA</td>
</tr>
<tr>
<td>2:00 pm</td>
<td>Relational data mining: new opportunities for IA</td>
</tr>
<tr>
<td>2:30 pm</td>
<td>Data mining and people: roles and expectations</td>
</tr>
<tr>
<td>3:00 pm</td>
<td>Break</td>
</tr>
<tr>
<td>3:30 pm</td>
<td>Relational methods for financial fraud/deception analysis</td>
</tr>
<tr>
<td>4:00 pm</td>
<td>Visual correlation/link discovery for terrorism analysis</td>
</tr>
<tr>
<td>4:30 pm</td>
<td>Relational methods for imagery analysis and integration</td>
</tr>
<tr>
<td>5:00 pm</td>
<td>End</td>
</tr>
</tbody>
</table>
Abstract

Purpose. Why is the topic important?

Data mining and link discovery have obvious value and great potential for intelligent analysis (IA). It is important that abilities as well as limitations of these methods will be well understood in the Intelligence Community. Traditional numeric statistical data mining methods have relatively limited applicability in IA, because data are often not numeric and have a very asymmetric pattern representation. For instance, there are only a few terrorism messages in the stream of normal ones. New relational data mining and link discovery have significant potential to address these challenges. Also, relational methods have important advantages over traditional methods for linking, integrating, and conflating images of different resolutions, sensor modalities, viewing angles, and geometric projections.

Learning objectives for the tutorial. What will participants learn?

This tutorial will focus on presenting relational data mining and link discovery methods for data and images. Illustrative examples of how such methods can be used for problems related to terrorism and financial support of terrorism will be provided. Another set of examples will be provided to illustrate the usefulness of relational methods for geospatial data and imagery analysis, integration and linking. We will give a brief mathematical background of these methods. We will compare them with more traditional data mining methods focusing on applicability issues. Some historical perspective will also be provided. Historically, relational methods have been more intensively developed in Europe, but more recently significant interest has also been generated in the US, with DARPA and other agencies support.

Audience

The target audience for this tutorial includes all source analysts, imagery analysts, researchers, and research managers. There is no formal prerequisite knowledge required although general knowledge of information technology, statistics, data, and image processing will be helpful while terms used will be explained.
Introduction

● This tutorial discusses
  - capabilities and limitations of numeric data mining for IA and
  - new opportunities that relational data mining provides for IA.

● The tutorial presents
  - Data mining methods conceptually and
  - demonstrate them
    • for financial fraud/deception analysis,
    • visual correlation/link discovery for terrorism analysis and
    • imagery and geospatial data integration/fusion.
Numeric Data Mining

- The state of the art in numeric data mining
  - Capabilities and Limitations for IA
- How to make these methods beneficial for IA?
- How to avoid unrealistic expectations?
- Numeric Data Mining methods
  - Time Series analysis,
  - Neural Networks,
  - Decision Trees,
  - Discriminant Analysis, and
  - Bayesian methods

- How to evaluate the results?
Numeric Data Mining

- Comparative performance of numeric methods
- Data
  - highly noisy real data
  - the International CoIL data mining competition [Putten, Someren, 2004].
- CoIL participants
  - data mining professionals
  - university students
    - CWU
    - Free University in Amsterdam.
- Surprising results
Numeric Data Mining

- CoIL surprising results [Putten, Someren, 2004].

*Figure 1.* Histogram of prediction task performance for CoIL Challenge participants and two reference groups of students (bucket size is 5).
Relational Data Mining and “Expert Mining”

- Relational data mining is
  - the data mining technology process based on relations between pieces of data and/or objects.
- Knowledge and patterns based on relations naturally capture the structure of data that lead to deeper understanding of data to be analyzed.
- Two types of relational methods based on
  - propositional logic and
  - the first order logic
    - much higher expressive power to discovering complex structural patterns.
- The difference between two types of relational methods
Relational Data Mining and “Expert Mining”

- Ideas behind relational methods
  - Inductive and Stochastic Logic Programming
  - Propositional vs. First order logic Bayesian networks
  - Semantic Probabilistic Reasoning, MMDR
  - Propositional Decision Trees vs. First order logic
    Decision trees
  - FOIL and
  - Propositional vs. First order logic association rules.

- Comparison of relational methods with numeric methods
  - Applicability of
    - relational methods and
    - numeric methods
Relational Data Mining and “Expert Mining”

- Traditional data mining techniques
  - Written data
    - already separated from a human that provided data
    - mining for patterns based on written data
- A new emerging challenge for data mining in IA
  - no written data
  - mining for data and patterns directly from humans, experts – “Expert mining”
- Historical perspective on developments of relational methods
  - Europe
  - USA
- Current trend
Financial Fraud and Deception

- The relational approach is naturally applicable for discovering financial fraud and deception
  - in traditional business and
  - in businesses that support terrorism.
- A new conceptual approach [Kovalerchuk, Vityaev; 2003]
- How to overcome difficulties of traditional data mining in discovering knowledge patterns?
  - a few relevant events
  - the ocean of irrelevant data
  - insufficient statistics of relevant data.
- An example
Financial Fraud and Deception

1.1. Generate predicates Q_1, Q_2, ..., Q_n and first order logic sentences A_1 \& A_2 \& ... \& A_{n-1}, A_n

1.2. Compute a probability P that pattern 
A_1 \& A_2 \& ... \& A_{n-1} \Rightarrow A_n is true on a given database.

1.3. Compare P(A_1 \& A_2 \& ... \& A_{n-1} \Rightarrow A_n) with a threshold T, say T=0.9.

   If P(A_1 \& A_2 \& ... \& A_{n-1} \Rightarrow A_n) > T then the pattern is “normal” for the database. (Note, DB may not be “normal”)

1.4. Test statistical significance of

   P(A_1 \& A_2 \& ... \& A_{n-1} \Rightarrow A_n).

   Mathematical detail on testing technique [Kovalerchuk, Vityaev, 2000]

1.5. Negate “normal pattern” to get suspicious pattern
# Financial Fraud and Deception

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Probability $P(A_1 &amp; A_2 &amp; \ldots &amp; A_{n-1} \Rightarrow A_n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In database without suspicious cases</td>
</tr>
<tr>
<td>Normal pattern, MBP $\Rightarrow$ SR</td>
<td>&gt; 0.95</td>
</tr>
<tr>
<td>Negated $\neg(MBP \Rightarrow SR)$, MBP $\Rightarrow \neg(SR)$</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Normal pattern CBP$\Rightarrow$ SR</td>
<td>&gt;0.95</td>
</tr>
<tr>
<td>Negated pattern $\neg(CBP \Rightarrow SR)$, CBP $\Rightarrow \neg(SR)$</td>
<td>&lt;0.05</td>
</tr>
</tbody>
</table>
# Discovered regularity

<table>
<thead>
<tr>
<th>#</th>
<th>Rule</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IF Seller_type = Manufacturing AND Buyer__type = Manufacturing THEN New_Item_type = product</td>
<td>72 / (6 + 72) = 0.923077</td>
</tr>
<tr>
<td>2</td>
<td>IF Seller_type = Manufacturing AND New_Buyer__type = Manufacturing THEN New_Item_type = product</td>
<td>72 / (6 + 72) = 0.923077</td>
</tr>
<tr>
<td>3</td>
<td>IF Seller_type = Manufacturing AND Item_type = precursor THEN New_Item_type = product</td>
<td>152 / (59 + 152) = 0.720379</td>
</tr>
<tr>
<td>4</td>
<td>IF Seller_type = Manufacturing AND Price_Compare = 1 AND New_Buyer__type = Trading THEN New_Item_type = product</td>
<td>47 / (2 + 47) = 0.959184</td>
</tr>
<tr>
<td>5</td>
<td>IF Seller_type = Manufacturing AND Price_Compare = 1 AND Item_type = precursor THEN New_Item_type = product</td>
<td>79 / (5 + 79) = 0.940476</td>
</tr>
</tbody>
</table>
Visual Correlation and Terrorism

- Visual analysis, data mining, link discovery and correlation are gaining momentum in IA
  - ARDA/NGA GI2VIs program.
- How visual relational methods contribute to this work?
- Data -- MUC3/4 competition data on terrorism
- A visual correlation system
  - Demo
- Visual Reasoning

exact correlation

Visual correlation using intermediate object \( B' \)
# Visual Correlation and Terrorism

## MUC raw text corpus description

<table>
<thead>
<tr>
<th><strong>Characteristic</strong></th>
<th><strong>Description</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Data sources</td>
<td>The Foreign Broadcast Information Service.</td>
</tr>
<tr>
<td>2. Text types</td>
<td>Newspaper and newswire stories, radio and TV broadcasts, interviews, and rebel communiqués summary reports, transcripts from speeches and interviews</td>
</tr>
<tr>
<td>3. Location</td>
<td>Latin America</td>
</tr>
<tr>
<td>4. Original language</td>
<td>Spanish</td>
</tr>
<tr>
<td>5. Text grammar</td>
<td>Well-formed sentences, all are in upper case</td>
</tr>
<tr>
<td>6. Number of texts</td>
<td>1300 texts</td>
</tr>
<tr>
<td>7. Individual text size</td>
<td>Average size is 12 sentences (~0.5 a page), smallest text -one paragraph, largest text -two pages</td>
</tr>
<tr>
<td>8. Number of sentences</td>
<td>15,600 sentences</td>
</tr>
<tr>
<td>9. Number of unique lexical items</td>
<td>18,240 lexical units</td>
</tr>
<tr>
<td>10. Number of words</td>
<td>400,000 words</td>
</tr>
<tr>
<td>11. Number of events in a text</td>
<td>1-5 events per single text source</td>
</tr>
<tr>
<td>12. Average sentence length</td>
<td>27 words</td>
</tr>
<tr>
<td>13. Timeframe</td>
<td>1980s</td>
</tr>
<tr>
<td>14. Terrorism relevant texts, %</td>
<td>50%</td>
</tr>
</tbody>
</table>
Visual Correlation and Terrorism

A sample of raw text [MUC Data Sets, NIST, 2001]

DEV-MUC3-0008 (NOSC)
BOGOTA, 9 JAN 90 (EFE) -- [TEXT] RICARDO ALFONSO CASTELLAR, MAYOR OF ACHI, IN THE NORTHERN DEPARTMENT OF BOLIVAR, WHO WAS KIDNAPPED ON 5 JANUARY, APPARENTLY BY ARMY OF NATIONAL LIBERATION (ELN) GUERRILLAS, WAS FOUND DEAD TODAY, ACCORDING TO AUTHORITIES. CASTELLAR WAS KIDNAPPED ON 5 JANUARY ON THE OUTSKIRTS OF ACHI, ABOUT 850 KM NORTH OF BOGOTA, BY A GROUP OF ARMED MEN, WHO FORCED HIM TO ACCOMPANY THEM TO AN UNDISCLOSED LOCATION. POLICE SOURCES IN CARTAGENA REPORTED THAT CASTELLAR’S BODY SHOWED SIGNS OF TORTURE AND SEVERAL BULLET WOUNDS. CASTELLAR WAS KIDNAPPED BY ELN GUERRILLAS WHILE HE WAS TRAVELING IN A BOAT DOWN THE CAUCA RIVER TO THE TENCHE AREA, A REGION WITHIN HIS JURISDICTION. IN CARTAGENA IT WAS REPORTED THAT CASTELLAR FACED A ”REVOLUTIONARY TRIAL” BY THE ELN AND THAT HE WAS FOUND GUILTY AND EXECUTED. CASTELLAR IS THE SECOND MAYOR THAT HAS BEEN MURDERED IN COLOMBIA IN THE LAST 3 DAYS. ON 5 JANUARY, CARLOS JULIO TORRADO, MAYOR OF ABREGO IN THE NORTHEASTERN DEPARTMENT OF SANTANDER, WAS KILLED APPARENTLY BY ANOTHER GUERRILLA COLUMN, ALSO BELONGING TO THE ELN. TORRADO’S SON, WILLIAM; GUSTAVO JACOME QUINTERO, THE DEPARTMENTAL GOVERNMENT SECRETARY; AND BODYGUARD JAIRO ORTEGA, WERE ALSO KILLED. THE GROUP WAS TRAVELING IN A 4-WHEEL DRIVE VEHICLE BETWEEN CUCUTA AND THE RURAL AREA KNOWN AS CAMPANARIO WHEN THEIR VEHICLE WAS BLOWN UP BY FOUR EXPLOSIVE CHARGES THAT DETONATED ON THE HIGHWAY.
## Visual Correlation

### Iconic sentences that summarize drag trafficking records

<table>
<thead>
<tr>
<th>Date</th>
<th>Location</th>
<th>Offender</th>
<th>Delivery, storage</th>
<th>Transit point</th>
<th>Value</th>
<th>Witness</th>
<th>Legal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>🌴</td>
<td>🧥</td>
<td>🛡️</td>
<td>🏛️</td>
<td>$$$$</td>
<td>🦆</td>
<td>📑</td>
</tr>
<tr>
<td>1</td>
<td>🌍</td>
<td>🎤</td>
<td>🛩️</td>
<td>🏛️</td>
<td>$$</td>
<td>🦆</td>
<td>📑</td>
</tr>
<tr>
<td>1</td>
<td>🌴</td>
<td>🧥</td>
<td>🛡️</td>
<td>🏛️</td>
<td>$</td>
<td>–</td>
<td>📑</td>
</tr>
</tbody>
</table>
Visual Correlation

<table>
<thead>
<tr>
<th>Date</th>
<th>Location</th>
<th>Offender</th>
<th>Tools</th>
<th>Victim</th>
<th>Harm</th>
<th>Witness</th>
<th>Legal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>![Tree]</td>
<td>![File]</td>
<td>![Knife]</td>
<td>![Person]</td>
<td>--</td>
<td>![Person]</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>![World]</td>
<td>![File]</td>
<td>![Bulb]</td>
<td>![Person]</td>
<td>![Wheelchair]</td>
<td>![Person]</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>![Palm]</td>
<td>![Car]</td>
<td>![Gun]</td>
<td>![Person]</td>
<td>![Question Mark]</td>
<td>--</td>
<td></td>
</tr>
</tbody>
</table>

Iconic sentences that summarize criminal records
**Visual Correlation and Terrorism**

| Organization (a tree icon) of a medium size (encoded by green lines) and relatively high confidence (encoded by a yellow mark on red) | Several soldiers perpetrators encoded by the soldier icon, red modifier for perpetrators; 5 blue lines for several and a yellow mark for medium in the red confidence scale. | Terrorist act with dynamite and significant damage (red lines) |

---

Add your Tutorial Title Here
Visual Correlation and Terrorism

Bruegel System case studies.
Visual Correlation and Terrorism

<table>
<thead>
<tr>
<th></th>
<th>message</th>
<th>template</th>
<th>security</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>message</td>
<td>![icon]</td>
<td>![icon]</td>
<td>![icon]</td>
<td>![icon]</td>
</tr>
<tr>
<td>template</td>
<td>![icon]</td>
<td>![icon]</td>
<td>![icon]</td>
<td>![icon]</td>
</tr>
<tr>
<td>security</td>
<td>![icon]</td>
<td>![icon]</td>
<td>![icon]</td>
<td>![icon]</td>
</tr>
<tr>
<td>ID</td>
<td>![icon]</td>
<td>![icon]</td>
<td>![icon]</td>
<td>![icon]</td>
</tr>
</tbody>
</table>

Composite icon generation.
Visual Correlation and Terrorism

<table>
<thead>
<tr>
<th>Visual queries</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Person" /> <img src="image2" alt="Open Hand" /> <img src="image3" alt="Gun" /></td>
<td>Standard query: Find an adult male using his hand to operate a gun</td>
</tr>
<tr>
<td><img src="image4" alt="Up Arrow" /> <img src="image5" alt="Down Arrow" /> <img src="image6" alt="Flashlight" /></td>
<td>OR</td>
</tr>
<tr>
<td><img src="image1" alt="Person" /> <img src="image2" alt="Open Hand" /> <img src="image7" alt="Phone" /></td>
<td>Exception: “Find an adult male using his <em>hand</em> to operate a telephone”.</td>
</tr>
</tbody>
</table>
Visual Correlation and Terrorism

Dynamics of compression of iconic sentence.

10 times compression
Max possible for a given n=10

3.3 (10/3) times compression

2 times compression

Single icon for the event (pure Bruegel)

3 icons for the event

5 icons for the event

10 icons

Icons with user defined weighting.
Visual Correlation and Reasoning

Visual reasoning rules
Visual Correlation and Reasoning

Integrated visual evidentiary reasoning scheme
Visual Image Integration/Fusion

- Image integration including fusion/registration/conflation is becoming a major task for imagery analysts.
- Images are increasingly coming from disparate sources with:
  - different resolutions,
  - sensor modalities,
  - geometric projections and
  - uncontrolled noise.
- The fundamental role that relational methods can play in integrating images and geospatial data.
- Examples
  - NGA NURI projects
Visual Image Integration/Fusion

- Image integration/fusion
Visual Image Integration/Fusion

• Image integration/fusion

Lake shores extracted from the photographs of Sonkyl

Structural interpolations of a polyline.

Sections of the extracted shorelines with the first level BSD interpolation with $k=1$ and $n=2$.

Fragment of BSD level 2 for the two polylines.
Visual Image Integration/Fusion

- Image integration/fusion: relational approach

<table>
<thead>
<tr>
<th>Angle</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>L4</th>
<th>L5</th>
<th>L6</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>L2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>L4</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>L5</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>L6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

\(<A_1, P_1(L_i,L_j)>\) abstract feature: angle

<table>
<thead>
<tr>
<th>Edge</th>
<th>a1</th>
<th>a2</th>
<th>a3</th>
<th>a4</th>
<th>a5</th>
<th>a6</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>a2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>a3</td>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>a4</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>a5</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>a6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

\(<A_2, P_2(a_i,a_j)>\) abstract feature: edge length

\(<A_2, P_3(a_i,a_j)>\) abstract feature: edge intensity

\(<A_2, P_4(a_i,a_j)>\) abstract feature: edge spectral

Relational feature representation
Expanding abstract feature concept.
Visual Image Integration/Fusion

Original feature proportions

Correct match found by BSD method

Alternative matches found
Summary

- Data mining and link discovery have significant value and great potential for intelligent analysis.
- Relational data mining methods may play a special role in intelligent analysis because of their abilities to deal with non-numeric data.
- Relational methods may play special role in efficient integration of images and geospatial data because of their abilities to capture structural property of image features.
References


Tutorial Evaluation Form
(Use reverse side to provide detailed comments)

- Please characterize yourself (e.g., analyst/technologist/program manager/student; academia/industry/national lab/government)
- Why did you attend?

- How well did the course meet the stated objectives (not at all, partially, fully)? Why?

- How would you rate each of the major sections, both in terms of content and effectiveness of presentation (poor, good, very good, excellent)

- Was the overall course valuable for you?

- What did you like most/least about the course?

- Do you suggest any improvements (e.g., topics covered, presentation, instructor)?

- Your name, e-mail address (feel free to leave blank) __________
Instructor’s Biography

Dr. Boris Kovalerchuk is a professor of Computer Science at Central Washington University and Director of the Imaging Laboratory at the University (http://www.cwu.edu/~Imaglab). He holds a Ph.D. in Computer Science and Applied Math, and an M.S. in Mathematics. Dr. Kovalerchuk has 30 years of research and teaching experience, and published two books on Data Mining, and Visual and Spatial Analysis (Kluwer, 2000, Springer, 2005), 5 book chapters and more than 80 papers on the subjects of closely related to this tutorial. He is deeply involved in the fields of data mining, link discovery, image integration, and analysis. Dr. Kovalerchuk has organized special sessions, delivered tutorials, presented new research at major conferences, such as SPIE Conferences on Defense and Security. He is a principal investigator on three research projects funded by the Intelligence Community (ARDA/NGA, http://www.ic-arda.org/InfoExploit/aquaint/gi2vis /projects/cffveaia.html, http://www.cwu.edu/~Imaglab). In this research he works with partners from National Laboratories, the Aerospace industry and Universities. Dr. Kovalerchuk teaches data mining at the University using his book. This book is also used in other universities in the US and abroad.