MARSHALLING EVIDENCE THROUGH DATA MINING

Daniel Barbará
James J. Nolan
David Schum
Arun Sood

George Mason University
Problem

♦ Lots of disparate, heterogeneous pieces of evidence.
♦ How do we make sense of it all?
♦ How do people investigate: making hypothesis. Two solutions:
  − A system that can make automatic hypothesis, through the use of models.
  − A system that supports hypothesis ``testing.''

Trifles?

You know my method. It is founded upon the observation of trifles...

Sherlock Holmes

``The Boscombe Valley Mystery``
Two approaches

- Automatic generation of hypothesis:
  + Less labor intensive
  - Rigid (constrained by the previously-built models. E.g.: Bayesian Networks)
  ⇒ Fails to adapt to new situations
- Human-in-the-loop (generating hypothesis)
  + Humans have great capacity for discovering new patterns
  - Laborious

Our approach: human-in-the-loop + Heavy support for hypothesis testing.
Hypothesis testing?

- By supporting:
  - Query answering
  - Linkage of evidence by data mining methods.
The architecture

Intelligence Trifles

Images
Signals
Human Intelligence
Open Source

New and Uncorrelated I-XML pages

Linked I-XML pages
Not Linked to Hypothesis

Query Processing

Linked I-XML pages
Linked to Other I-XML pages

I-XML pages
Linked to Existing Hypothesis

Weak Linkage to Hypothesis

I-XML pages
Strong Linkage to Hypothesis

Strong Linkage to Hypothesis

Add, Delete, Alter

Build I-XML pages

Build New Hypothesis

Existing Hypothesis

Build I-XML pages

Hypothesis Creation, Justification, Negation

Human Interaction
<table>
<thead>
<tr>
<th>Trifle address</th>
<th>Source</th>
<th>Date</th>
<th>Time</th>
<th>Location</th>
<th>Individuals</th>
<th>Assesment</th>
<th>Text</th>
<th>Image characteristics</th>
<th>Some tags</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Queries, queries everywhere…

♦ Durability:
  – Standing (continuous) queries
  – Ad-hoc queries

♦ Complexity
  – Unformatted (query-by-example: take a trifle and use it as a query)
  – Formatted: list-of-keywords (or I-XML tags)
  – Richer queries (data mining): e.g., “is there a change in the trend of money transfers to a certain group of individuals?” HYPOTHESIS can be formulated by richer queries.
Richer queries

♦ Only limited by the implemented tasks in our system. E.g.,:
  – Frequent episodes,
  – Time-series outliers
  – Trend shifts
Grouping and querying trifles

♦ Supervised learning (text, and other media classification)
  – Trifle parsing into I-XML
  – Dimensionality reduction
  – Classifier building:
    • Unlabeled sets: clustering
    • Record linkage
    • T. Mitchell’s work
Trifles grouped by class

**SQ1:** Query (formatted) Iraq AND Weapons of mass destruction

1. If $C_I$: Belongs to $I$?
   - Yes: New trifle $T_n$
   - No: $C_W$: Belongs to $W$?
     - Yes: $C_S$: Belongs to $S$?
       - Yes: $T_n$
       - No: $T_n$
     - No: $T_n$

2. If $C_W$: Belongs to $W$?
   - Yes: $T_n$
   - No: $T_n$

$I$: Iraq $\cdots T_i, \cdots T_j, \cdots$

$W$: Weapons of mass destruction $\cdots T_i, \cdots T_k, \cdots$

$C_S$: Belongs to $S$?

$T_n$

$S$: Somalia $\cdots T_j, \cdots T_s, \cdots$
Clustering

SQ (unformatted)

C1

T1

C2

C3
Challenges

♦ Unsupervised learning
  – Absence or limited availability of a training corpus of trifles
  – Dynamic nature of the trifles
  – Large volume of trifles

♦ Clustering
  – Large dimensional space
  – Lots of missing values
  – Large volume of trifles

♦ Richer queries
  – Scalable methods in a distributed environment
Extra links

♦ Some important trifles may be missed by similarity comparison

♦ Linking trifles is a way to avoid this. E.g.:
  – Ti and Tj are target for SQ1, Tj and Tk are target for SQ2: Ti and Tk may be linked
  – Ti and Tj have words in common, but they are not classified (or clustered) together.

♦ The “small worlds” principle (the Kevin Bacon Game)
The value of “extra links”

Trifle: Strange shootings in backyard of house in WA (early in the investigation; regarded low priority)

Trifle: Gunman leaves note, instructing to wire $10M to acc xxx (Saturday, Oct. 19)

Trifle: ATM card of acc xxx had been used recently in WA

Trifle: House tenants Ids.
On the shoulders of AIGA

http://aiga.cs.gmu.edu/

♦ Agent-based Imagery and Geospatial Architecture (AIGA)

♦ To-date Achievements
  – AIGA Architecture
  – Publications
  – Prototype
Agent Architecture

♦ Agents
  – Perform specified function
  – Imagery, geospatial, Info. Retrieval (Google), Natural Language Processing (Annie), Data Mining.

♦ Locations
  – Provide places for agents to execute

♦ Communication Space
  – Allow agents to pass messages, data, objects to one another
  – Asynchronous communication
  – Knowledge repository

♦ Data Repositories
  – Provide access to imagery, geospatial, and other data
Architectural View

I-XML Page Space

- Imagery Agent
- Geospatial Agent
- Loc.
- Clustering Agent
- Data Agent
- Info. Retrieval Agent
- Client
- Geo Library
- Name Server
- Image Library
- Data Agent
- Geo Library
- Imagery Agent
- Client
- Data Agent
- Classf. Agent
- Loc.
- Geo Library
- Name Server
- Image Library
- Data Agent
Example scenario

1. **Trifle 1:** Shoulder fired anti-aircraft missiles stolen from a US Army base by member of a militia group in USA
2. **Hypothesis:** The militia group has shoulder-fired AA missiles
3. **Trifle 2:** Phone conversation between X and a member of the militia group.
4. **Hypothesis:** The conversation involved the sale of stolen weapons.
5. **Trifle 3:** X is a Saudi national known to have been in Afghanistan in January, 2001
6. **Hypothesis:** X has obtained shoulder-fired AA missiles from the militia group.
7. **Hypothesis:** The shoulder-fired AA missiles that will be delivered to the L. A. area.
8. **Trifle 4:** Photo of X at an ATM at LAX.
9. **Hypothesis:** X is now in LA area.
10. **Hypothesis:** Al Qaeda sleepers in the USA are planning attacks on civilian airliners landing and taking off at LAX.
Example scenario

**HUMINT**
- **Evidence 1:** Shoulder fired anti-aircraft missiles stolen from a US Army base by members of a militia group here in the USA.
- **Evidence 2:** Recorded phone conversation between member of militia group and a man identified as X.
- **Evidence 3:** The person identified as X is a Saudi national known to have been in Afghanistan in January of 2001.

**SIGINT**
- **Evidence 2:** Recorded phone conversation between member of militia group and a man identified as X.

**IMINT**
- **Evidence 4:** Video tape of person identified as X at an ATM at LAX Airport.

**NEW AND UNCORRELATED XML PAGES**

**CLUSTER AND LINK ANALYSIS:**
Helps to generate lines of argument on hypotheses.

NEW HYPOTHESIS:
Al Qaeda Sleepers in the USA planning attacks on civilian airliners landing and taking off at LAX.
Summary

- A flexible architecture to support hypothesis testing via query evaluation
- AIGA provides the distributed agents framework
- System can be incrementally enriched by adding query capabilities (through agents)
- A test bed for intelligence management techniques