DM Research Map in India

IIT Bombay
CSE, Chemical
Sunita Sarawagi [IR]
Soumen Chakrabarti [Web]
Saketa Nath [ML]
Pramod Wangikar [Bio]

IIT Bangalore
Srinath Srinivas [Graph, Doc]

IISc Bangalore
CSA, EE, SERC
M Narasimha Murthy [ML, ARM]
C Bhattacharyya [ML, Bio, Hand]
Shirish Shevade [SVM]
P S Sastry [Temporal, DecTrees]
Jayant Haritsa [ARM, Privacy]

IIT Delhi
CSE
S K Gupta [ARM, Security]

IIT Kanpur
CSE
Arnab Bhattacharya [Spatial, Temporal, Bio]

IIT Kharagpur
CSE
Pabitra Mitra [ML]

IIT Hyderabad
Kamal Karlapalem [Ecom, Cluster]
Vikram Pudi [ARM, Multimedia]
P Krishna Reddy [ARM]

IIT Madras
CSE
P Sreenivasa Kumar [Text, XML]
D Janaki Ram [S/W Engg]
B Ravindran [Text, Tutoring, Motion]
Temporal Data Mining

[P S Sastry]
A temporal data mining framework

Data is a sequence of events. Each event has a type and a time of occurrence

‘Patterns’ in the formalism are episodes – partially ordered sets of event types.

Episode occurrence: events in the data that conform to the partial order
Algorithms are developed to discover frequent episodes of different structures.

Can handle additional temporal constraints.

Statistical theory developed to assess significance under various null hypothesis models.

Unified view of counting-based CS and model-based Stats approaches [equivalence between frequent episodes and HMMs]
Example Data Sequences

- Fault report logs in manufacturing plants: root cause diagnostics
- Multi-Neuronal Spike train data: discovering microcircuits or groups of neurons with ‘strong’ interactions
- Web navigation data: Prediction of user behaviour
An Example Application

- **Data**: Sequence of fault codes in an assembly plant
- **Event types**: fault codes, date and time
- **Episodes**: Causative (?) Chains
- Rules derived from episodes provide help in root cause diagnosis
- Currently in use in some engine assembly plants of General Motors
Application: Status logs from Assembly Plants

Fault Correlations

Root cause diagnostics

Fault Logs

<table>
<thead>
<tr>
<th>Station</th>
<th>Error</th>
<th>Date/Time</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>OP 063</td>
<td>600-42</td>
<td>4/5/2004 4:42</td>
<td>0:01:00</td>
</tr>
<tr>
<td>OP 050</td>
<td>650-47</td>
<td>4/5/2004 5:06</td>
<td>0:02:00</td>
</tr>
<tr>
<td>OP 063</td>
<td>650-294</td>
<td>4/5/2004 5:38</td>
<td>0:01:00</td>
</tr>
<tr>
<td>OP 063</td>
<td>650-234</td>
<td>4/5/2004 5:50</td>
<td>0:01:00</td>
</tr>
<tr>
<td>OP 014</td>
<td>650-87</td>
<td>4/5/2004 6:01</td>
<td>0:02:00</td>
</tr>
<tr>
<td>OP 014</td>
<td>650-87</td>
<td>4/5/2004 6:04</td>
<td>0:02:00</td>
</tr>
<tr>
<td>OP 109</td>
<td>650-24</td>
<td>4/5/2004 6:25</td>
<td>0:07:00</td>
</tr>
<tr>
<td>OP 063</td>
<td>650-294</td>
<td>4/5/2004 6:38</td>
<td>0:01:00</td>
</tr>
<tr>
<td>OP 063</td>
<td>650-232</td>
<td>4/5/2004 6:40</td>
<td>0:01:00</td>
</tr>
</tbody>
</table>

June 2010

Indo-Spain ICT Workshop
Associated Publications


Online Generation of Web Tables
[Prof. Sunita Sarawagi]
# The semi-structured web

## Table

<table>
<thead>
<tr>
<th>Country</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ouagadougou</td>
<td>Burkina Faso</td>
</tr>
<tr>
<td>Pago Pago</td>
<td>American Samoa</td>
</tr>
<tr>
<td>Palikir</td>
<td>Federated States of Micronesia</td>
</tr>
<tr>
<td>Panama City</td>
<td>Panama</td>
</tr>
<tr>
<td>Papeete</td>
<td>French Polynesia</td>
</tr>
<tr>
<td>Paramaribo</td>
<td>Suriname</td>
</tr>
<tr>
<td>Paris</td>
<td>France</td>
</tr>
</tbody>
</table>

## Regular page

### Airports in Germany

- **Berlin-Tegel Airport**
  - Berliner Flughafen-Gesellschaft mbH, Flughafen Tegel, 13405 Berlin, Germany
  - Airport Code: TXL

- **Cologne - Bonn Airport**
  - Postfach 98 01 20, 50129 Cologne, Germany
  - Airport Code: CGN

- **Munich Airport**
  - PO Box 23 17 55, 85326 Munich, Germany
  - Airport Code: MUC

- **Hamburg Airport**
  - Flughafensstrasse 1-3, 22335 Hamburg, Germany
  - Airport Code: HAM

## List

2. Ixtoc I oil well, Gulf of Mexico, June 3, 1979
3. Norweroi oil field, Persian Gulf, February, 1983
4. Atlantic Empress and Aegean Captain collision, off Trinidad and Tobago, 1989
5. Castillo de Bellver, off Cape Town, South Africa, August 6, 1983
6. Amoco Cadiz (BP/Amoco, USA) - Brittany, France, March 16, 1979
7. Torrey Canyon, South England, March 18, 1967
8. Sea Star, Gulf of Oman, December 19, 1972
9. Urquiola, La Coruna, Spain, May 12, 1976
11. Othello, Tralhavet Bay, Sweden, March 20, 1970

## Formatted list

- **Braer - Shetland Islands**, January 5, 1993
- **Prestige - Galicia, Spain**, November 13, 2002
- **Aegean Sea**, off N Spain, December 3, 1992
- **Sea Empress - Wales**, February 15, 1996
- **World Glory**, off South Africa, June 13, 1968
- **Corinthos Delaware River, Marcus Hook, Pennsylvania**, January 31, 1975
- **Burrnagh Aggie Galveston Bay, Texas**, November 1, 1979
- **Exxon Valdez (Exxon, USA)** - Prince William Sound, Alaska, March 24, 1989
- **Ko - off MA, November 5, 1959**
- **Storage Tank, Seavent NJ**, November 4, 1969
- **Ekofisk oil field, North Sea**, April 22, 1977
- **Frika - Bay of Biscay**, December 12, 1995
Queries in WWT

- Query by example
  
<table>
<thead>
<tr>
<th>Inventor</th>
<th>Computer science concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alan Turing</td>
<td>Turing Machine</td>
</tr>
<tr>
<td>E. F. Codd</td>
<td>Relational Databases</td>
</tr>
</tbody>
</table>

- Query by description
<table>
<thead>
<tr>
<th>Inventor</th>
<th>Computer Science Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alan Turing</td>
<td>Turing Machine</td>
</tr>
<tr>
<td>Seymour Cray</td>
<td>Supercomputer</td>
</tr>
<tr>
<td>E. F. Codd</td>
<td>Relational Databases</td>
</tr>
<tr>
<td>Tim Berners-Lee</td>
<td>WWW</td>
</tr>
<tr>
<td>Charles Babbage</td>
<td>Babbage Engine</td>
</tr>
</tbody>
</table>
Experiments

- Aim: Reconstruct Wikipedia tables from only a few sample rows.
- Sample queries
  - TV Series: Character name, Actor name, Season
  - Oil spills: Tanker, Region, Time
  - Golden Globe Awards: Actor, Movie, Year
  - Dadasaheb Phalke Awards: Person, Year
  - Parrots: common name, scientific name, family
Accuracy of joint labeling

- Dataset
  - Manually labeled
    - 450 tables spanning general Web and Wikipedia
  - Automatically labeled
    - 650 tables from Wikipedia where cells have entity links

![Accuracy Graph]

- F1 Accuracy
- LCA
- Majority
- Ours

<table>
<thead>
<tr>
<th></th>
<th>Manual</th>
<th>Automatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>Types</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>relations</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>Entity</td>
<td>0.8</td>
<td>0.9</td>
</tr>
</tbody>
</table>
Summary

- Amazing amount of quality information on the semi-structured web.

- WWT
  - Online: structure interpretation at query time
  - Domain-independent methods for extraction and coreference resolution
  - Relies heavily on unsupervised statistical learning
    - Joint training for exploiting overlap during extraction
    - Graphical model for table annotation
    - Collective column labeling for descriptive queries
    - Bayesian network for consolidation
    - Page rank + confidence from a probabilistic extractor for ranking
Further Details

- Gupta and Sarawagi, VLDB 2009
Association Rule Mining

[Jayant Haritsa]
Data Layouts for AR Mining

- **Table Organization**
  - Horizontal (transactions / rows)
  - Vertical (items / columns)

- **Data Representation**
  - Value List (only presence)  
  - Bit Vector (presence and absence)  

```
1  2  3  4  5
1  1  0  0  1
```
Data Layout Combinations

Horizontal Item Vector (HIV)  Horizontal Item List (HIL)

Vertical Tid Vector (VTID)  Vertical Tid List (VTID)

Our Approach  Apriori
Why Vertical Organization?

- “Natural” for association rule mining’s goal of discovering correlated items (columns)
- Support counting simple (set intersections)
- No excess baggage from disk (automatic and immediate reduction of database after each scan)
- Ideal for parallel implementations (asynchronous, not level-wise, computation)
  - counting of AB can start before item C has been fully counted
Why Bit-Vector Representation?

- Tremendous scope for compression, especially since databases are typically sparse
- Vertical orientation offers better compression than horizontal since column lengths are proportional to size of database whereas row lengths are proportional to size of schema
- In fact, compressed VTV occupies much less space than HIL
Contributions

• Compressed VTV data layout — “Snake”
• VIPER snake mining algorithm
  – (Vertical Itemset Partitioning for Efficient Rule-extraction)
  – several optimizations for snake generation, intersection, counting and storage
  – “general-purpose” (prior vertical algorithms have restrictions on DB size, shape, contents, mining process)
  – substantial performance improvement (response time, disk space, disk traffic)
  – in some cases, beats “optimal” horizontal!

• Presented in ACM SIGMOD 2000
Privacy Preserving Mining
A Typical Web-Service Form (e.g. Amazon.com)
State-of-the-Art

• To cater to such privacy concerns, users are forced to resort to falsifying their data
  - (E.g.: There appear to be numerous grandmothers from Bangalore downloading rock music, because the actual clients — young male IISc students — have falsified their age and gender)

• But then, the models become meaningless …
  – Garbage in, Garbage out
Design Wish-list

- **High Privacy**
  - User-visible (i.e. should be done at user site)

- **Highly accurate models**
  - Association Rules correctly identified

- **Efficiency**
  - Data collection / Mining-process
The Digital Divide

Desire “Globally accurate, locally private”, but

Data Privacy  Accurate Models

Versus
Bridging the Divide

User Data

Data Distortion Procedure

Distribution Reconstruction Procedure

Mining Algorithm

Accurate Models
Optimal Distortion Matrix

\[
x = \begin{bmatrix}
\gamma & 1 & 1 & \cdots \\
1 & \gamma & 1 & \cdots \\
1 & 1 & \gamma & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{bmatrix}
\]

\[\gamma \text{ determines privacy level}\]

Where \[x = \frac{1}{\gamma + |S_U| - 1}\]

- Symmetric positive-definite Toeplitz matrix, with Gamma diagonal
- Results in matrix with lowest “condition number” (ratio of maximum and minimum eigen-values)
  - makes reconstruction least sensitive to the variance in the distribution of the distorted database
  - order-of-magnitude accuracy improvements
- Dependent column perturbation
Summary

• By careful mathematical design, it is possible to simultaneously achieve user data privacy, accurate statistical models, and good runtime efficiency.
Further Details

• “Maintaining Data Privacy in Association Rule Mining” [VLDB 2002]
• “On Addressing Efficiency Concerns in Privacy-Preserving Mining” [DASFAA 2004]
• “A Framework for High-Accuracy Privacy-Preserving Mining” [ICDE 2005]

• All available at http://dsl.serc.iisc.ernet.in
Questions?
Experiments

- Aim: Reconstruct Wikipedia tables from only a few sample rows.
- Sample queries
  - TV Series: Character name, Actor name, Season
  - Oil spills: Tanker, Region, Time
  - Golden Globe Awards: Actor, Movie, Year
  - Dadasaheb Phalke Awards: Person, Year
  - Parrots: common name, scientific name, family