Data Mining: Efficiently Extracting Interpretable and Actionable Patterns

Tao Li
School of Computing and Information Sciences
Florida International University
Data Mining Characteristics

*We are drowning in information*  
*But starving for knowledge*  
------John Naisbett

- Data mining (knowledge discovery from data)
  - Efficiently extracting interpretable and actionable patterns/knowledge from large amount of data collections
  - Patterns are Interesting
    - non-trivial, implicit, previously unknown and potentially useful

- Data Mining
  - An Engineering Process
  - An Interdisciplinary Field
  - A Collection of Functionalities
  - A Combination of Theory and Application
Research Challenges

• Efficiently extracting interpretable and actionable patterns from large amount of data collections
  – Performance: efficiency, effectiveness, and scalability
    • Novel algorithms
      – High dimensional, large volume
      – Efficient data structures
    • Parallel, distributed and incremental mining methods
  – Interpretable and actionable patterns
    • Pattern evaluation, Incorporation of background knowledge,
    • Knowledge fusion: Integration of the discovered knowledge with existing one
    • User interaction
      – Expression and visualization of data mining results
      – Interactive mining of knowledge at multiple levels of abstraction
    • Complex structure mining (e.g., graph, relational data)
    • Mining from heterogeneous information sources
    • Mining in Social Network Setting
      – Linked data between emails, webpages, blogs, citations, sequences
      – Static and dynamic structural behavior
    • Security, Privacy and Data Integrity

• New Applications
  – Anti-terrorism, Search advertising, Recommendation systems, Computing system management, Personalized Search
My Research Directions

Research Interests

Data Mining, Machine Learning and Information Retrieval

Practical Problems

Generic Tools

Theory

Applications:
- Computing System Management
  - Help Desk Applications
  - Self-managing Systems
  - Security Management
- Music Information Retrieval
  - Personalized Music Service
- Adaptive Workflow Management
  - Process Mining
  - Workflow Optimization
- Bioinformatics
  - Microarray data analysis
  - Sequence Analysis

Tools and Libraries:
- Workflow and System Management
  - Event Mining
  - Log Parser
  - Visualization Tools
- Personal Music Assistant
- Tools for Bioinformatics
  - Gene Selection/Tissue Classification
  - Sequence analysis
  - Literature Assistant

Algorithm Issues:
- Matrix-based Learning Framework
- Semi-supervised Learning
- Learning from heterogeneous data types
- Learning from non-vectorial data
- Stream data mining
- Incremental learning
Some Projects in My Group

- Intelligent Help Desk
- Database Navigation
- Music Information Retrieval
- Computing System Management
- Business Intelligence
  - Leading KPI Discovery
- Bioinformatics
- Social Network Analysis
- Matrix-Based Algorithms for Data Mining
Example of Call Center Cases

- Adding more memory on XXXXXX
- Email (a wrong setting)
- Scanner Account
- Problem with encrypted telnet
- PDFs not opening in IE

S: What is your example?
C: An example: http://.../example.pdf
S: What is your environment?
C: Windows XP with Acrobat Standard
S: Running a repair on Windows
C: It doesn’t work.
S: Uninstalling Acrobat Standard and Reader and then reinstalling the Reader.
C: It doesn’t work either.
S: Install two Microsoft updates for IE.
C: It works.

the request q

Number of open requests: 2

<table>
<thead>
<tr>
<th>Req #</th>
<th>Owner</th>
<th>Age</th>
<th>Told</th>
<th>Status</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>28000</td>
<td>onece001</td>
<td>5 mth</td>
<td>resolve <a href="mailto:jrodr047@cs.fiu.edu">jrodr047@cs.fiu.edu</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Command:</td>
<td>Comment</td>
<td>Respond</td>
</tr>
<tr>
<td>27999</td>
<td>cleun001</td>
<td>5 mth</td>
<td>resolve <a href="mailto:raju@cs.fiu.edu">raju@cs.fiu.edu</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Subject: Please set up laptop &amp; projector in ECS 243 by 1:30pm.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Command:</td>
<td>Comment</td>
<td>Respond</td>
</tr>
<tr>
<td>27998</td>
<td></td>
<td>5 mth</td>
<td>resolve <a href="mailto:ocarbone@cis.fiu.edu">ocarbone@cis.fiu.edu</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Subject: Toner Low-Stapler</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Command:</td>
<td>Comment</td>
<td>Respond</td>
</tr>
<tr>
<td>27997</td>
<td>jrodr047</td>
<td>5 mth</td>
<td>resolve <a href="mailto:jrodr047@cs.fiu.edu">jrodr047@cs.fiu.edu</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Subject: Khalil vhl31</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Command:</td>
<td>Comment</td>
<td>Respond</td>
</tr>
</tbody>
</table>

Done
Motivation

• Problems
  – Symptoms must be manually described to help-desk
  – Help-desk needs to determine solution - difficult to compare with all previous problems
  – Difficult for one help-desk employee to gain from another's experience.
  – Difficult to determine if solution worked successfully - User feedback may not be accurate.

• How can data mining/machine learning help?
  – Malfunction diagnoses, call distribution, skill-based routing, outbound targeting, etc.
  – 90% of calls are due to the most frequent 10% of problems
  – Problem analysis based on known cases
  – Augment known cases when a new case is solved
  – Significantly reducing the cost and shortening time needed to diagnose malfunction issues, improve customer satisfaction
Two Different Approaches

• Semantic-based Case Retrieval
  – Using Semantic features to retrieval the most similar past cases

• Problem Probing
  – Find probe/question set which maximizes utility (benefit – cost)
Semantic Feature Extraction

➢ To detect sentence relations

**Semantic Roles**
For certain verbs of motion
- Arg0: causer of motion
- Arg1: thing in motion
- Arg2: distance moved
- Arg3: start point
- Arg4: end point
- Arg5: direction

-- e.g. [Arg0 I] [ArgM-MOD can][ArgM-NEG not]
[rel receive] [Arg1 emails]

Subtypes of ArgM modifier tag
- LOC: location
- EXT: extent
- DIS: discourse connectives
- ADV: general-purpose
- NEG: negation marker
- MOD: modal verb

**Word Relations**
WordNet 3.0 (147,306 words)
20 word relations, e.g. identical, synonyms, antonyms, attribute; hypernym and hyponym
-- A is a hypernym of B if B is a kind of A
-- A is a meronym of B if A is a part or member of B

**Features**
Arg0_identical,
Arg0_synonyms,
Arg0_hypernym,
Arg0_meronym,
rel_identical,
rel_synonyms,
rel_hypernym,
Rel_meronym,
Arg1_identical,
### Case Study – Online Help Desk Interface

**Can you switch these two computers?**

```plaintext
15496.txt
Subject: KVM switch
We would like to purchase:
- 1 eight-port KVM switch
- 6 VGA (ps/2) adapters
Could we get quotes please?
```

```plaintext
18697.txt
Subject: research 2-port KVM switch
When we deploy xx's new machines, another 2-port KVM switch will need to order cables.
```

```plaintext
3191.txt
Hi xx
I have an approval from xx for computer exchange. I need some time to make a backup of important information and then I will bring both computers on Tuesday.
```

### Most related emails

<table>
<thead>
<tr>
<th>15496.txt</th>
<th>18697.txt</th>
<th>3191.txt</th>
<th>11002.txt</th>
<th>6557.txt</th>
</tr>
</thead>
<tbody>
<tr>
<td>28600.txt</td>
<td>14363.txt</td>
<td>12904.txt</td>
<td>15496.txt</td>
<td>20440.txt</td>
</tr>
<tr>
<td>14363.txt</td>
<td>28600.txt</td>
<td>12904.txt</td>
<td>15496.txt</td>
<td>12544.txt</td>
</tr>
<tr>
<td>12904.txt</td>
<td>12904.txt</td>
<td>15496.txt</td>
<td>20440.txt</td>
<td>12810.txt</td>
</tr>
<tr>
<td>15496.txt</td>
<td>15496.txt</td>
<td>20440.txt</td>
<td>12810.txt</td>
<td>18697.txt</td>
</tr>
<tr>
<td>20440.txt</td>
<td>20440.txt</td>
<td>12810.txt</td>
<td>18697.txt</td>
<td>25355.txt</td>
</tr>
<tr>
<td>12810.txt</td>
<td>12810.txt</td>
<td>18697.txt</td>
<td>25355.txt</td>
<td>20391.txt</td>
</tr>
<tr>
<td>18697.txt</td>
<td>18697.txt</td>
<td>25355.txt</td>
<td>20391.txt</td>
<td>3191.txt</td>
</tr>
<tr>
<td>25355.txt</td>
<td>25355.txt</td>
<td>20391.txt</td>
<td>3191.txt</td>
<td>6557.txt</td>
</tr>
<tr>
<td>20391.txt</td>
<td>20391.txt</td>
<td>3191.txt</td>
<td>6557.txt</td>
<td>3191.txt</td>
</tr>
<tr>
<td>3191.txt</td>
<td>3191.txt</td>
<td>6557.txt</td>
<td>3191.txt</td>
<td>3191.txt</td>
</tr>
</tbody>
</table>
```

### Slides from Dingding Wang
Some Projects in My Group

- Intelligent Help Desk
- Database Navigation
- Music Information Retrieval
- Computing System Management
- Business Intelligence
  - Leading KPI Discovery
- Bioinformatics
- Social Network Analysis
- Matrix-Based Algorithms for Data Mining
SQL-Query-Result Navigation

What mutual funds to invest? There are thousands of them

High return? Low risk? Low expense? …

Information overloading problem: users often can not form a query that returns answers exactly matching their needs
Existing Solutions – Iterative Refinement

• Iterative refinement
  – Find all funds with 3 Year Return > 10%
  – Find all funds with 3 Year Return > 10% and 1 Year Return > 10%
  – Find all funds with 3 Year and 1 Year Return > 10% and expense < 2%
  – …

• Problem: too time consuming
Existing Solutions - Ranking

• Ranking
  – Similar to TF/IDF (Agrawal et al. CIDR 2003)
  – Probabilistic ranking (Chaudhuri et al. VLDB 2004)
  – Ordering attributes (Das et al. SIGMOD 2006)
  – Context-sensitive (Agrawal et al. SIGMOD 2006)
  – Based on object relationships (Chakrabarti et al. SIGMOD 2006)
  – A lot of work in IR field (e.g., Shen et al. SIGIR 2005)
Existing Solutions - Ranking

• Problem: the same ranking function for all users (except the context sensitive work)
  – User 1 prefers high returns
  – User 2 prefers lower risks (often means lower returns)
  – User 3 prefers a mix of high return funds and low risk funds
  – One ranking function can not fit all
Existing Solutions - Categorization

Chakrabarti et al. SIGMOD 2004:

- Divide query results recursively into categories (a tree structure)
- The splitting attributes are selected based on query history (more frequently asked attributes are typically put closer to root)
- User can navigate this tree and show records in a node
- Categorization is complementary to ranking (rank records in a node)

```
Root(193)
  ↓  ↓  ↓
0 <= 3 Yr return < 10% (93)  10% <= 3 Yr return < 20% (71)  3 Yr return >= 20% (29)
```
Existing Solutions - Categorization

Problem: differences between users are not considered

This tree does not help user 2 and 3 much

Root(193)

User 1, 3

3 Yr return >= 20% (29)

0 <= 3 Yr return < 10% (93)

10% <= 3 Yr return < 20% (71)

0 <= 1 Yr return < 10% (63)

10% <= 1 Yr return < 20% (55)

User 2, 3

User 2, 3

0% <= Std < 5% (38)

5% <= Std < 10% (7)

10% <= Std < 20% (48)

We try to address this issue for categorization
Address Diverse User Preferences (Chen & Li, SIGMOD 2007)

- **Offline**: infer a set of typical types of user preferences
  - Data are clustered using query history of all existing users
  - Each cluster represents a typical type of user preference (e.g., high return funds, low risk funds)

- **Runtime**:
  - A navigational tree is constructed over clusters in the query results
  - This tree highlights the differences between these clusters
  - User navigates this tree to select clusters that he is interested in
  - User can then rank or browse records in those clusters
Some Projects in My Group

- Intelligent Help Desk
- Database Navigation
- Music Information Retrieval
- Computing System Management
- Business Intelligence
  - Leading KPI Discovery
- Bioinformatics
- Social Network Analysis
- Matrix-Based Algorithms for Data Mining
System Log Data

- **Performance Metric**
  - CPU/Memory/Swap utilization, workload average, response time
- **Application-level Log**
  - Windows NT system log, Websphere application log, Oracle activity
- **Failure data**
  - System crash dumps, error reports
- **Network Traffic Data**
  - SNMP Link-level byte counts, Netflow Link-level flow statistics, PMA link-level packet header traces, TCPdump host-side packet traces
- **Request Data**
  - Apache and Microsoft IIS log
- **Reports from Operators**
  - Trouble ticket data: problems, reports, possible causes and symptoms for failures
- **Others**
  - Network-based alert logs, Probes
The Integrated Framework on Data-Driven Computing System Management

Component logs

Logs

Historical Data Collection

Realtime Analysis

Anomaly Detection

Fault Diagnosis

Problem Determination

Real Time Management

Active Data Collection

Planning/Actions

Correlation/Dependency Knowledge

Knowledge Base

Log Data Organization

Log Adapter

Situation Identification and Categorization

Summarization/Visualization

Rule Construction

Temporal Pattern Discovery

Knowledge Management

Offline Analysis
Pattern Mining

Raw data
-> Data table

Event Plots

Aggregation
summarization

Outages
Temporal patterns
Event bursts
Periodic patterns
Weekly rebooting
Morning rebooting
Periodic patterns
## A Case Study

<table>
<thead>
<tr>
<th>Time</th>
<th>EventSource</th>
<th>Log Message</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>1093359583</td>
<td>UPHClean</td>
<td>The following handles in user profile hive FIU-SCS \ lcruz01 (S-1-5-21-876885435-739206947-926709054-2036) have been closed because they were preventing the profile from unloading successfully wmiprvse.exe (2432)HKCU (0x228)</td>
<td>report</td>
</tr>
<tr>
<td>1093359583</td>
<td>UPHClean</td>
<td>User profile hive cleanup service version 1.5.5.21 started successfully.</td>
<td>start</td>
</tr>
<tr>
<td>1093359603</td>
<td>AutoEnrollment</td>
<td>Automatic certificate enrollment for local system failed to contact the active directory (0x8007054b). The specified domain either does not exist or could not be contacted. Enrollment will not be performed&lt;br&gt;.</td>
<td>dependency</td>
</tr>
<tr>
<td>1093359605</td>
<td>Inventory Scanner</td>
<td>Unable to open output file C:\invdelta.tmp.</td>
<td>dependency</td>
</tr>
<tr>
<td>1093359606</td>
<td>Inventory Scanner</td>
<td>LDISCN32: Can’t create temporary file.</td>
<td>create</td>
</tr>
<tr>
<td>1093359641</td>
<td>Inventory Scanner</td>
<td>Unable to open output file C:\invdelta.tmp.</td>
<td>dependency</td>
</tr>
<tr>
<td>1093359918</td>
<td>Inventory Scanner</td>
<td>LDISCN32: Can’t create temporary file.</td>
<td>create</td>
</tr>
<tr>
<td>1093359584</td>
<td>MsiInstaller</td>
<td>Product: WebFldrs XP – Configuration completed successfully.</td>
<td>dependency</td>
</tr>
</tbody>
</table>

**State:**
- start, stop, dependency, create, connection, report, request, configuration, other

**Plot:**
(a) The 2D plot of a part of temporal log data. X axis is the time. Y axis is the event type.

(b) The 2D plot of the dataset after the subset reflecting the dependency relationship from event 5 (dependency6) to event 6 (create6) being removed.

(c) The dataset after the subset reflecting the synchronization relationship between event 13 (request1) and event 14 (stop4) being removed.

(d) The dataset after the subset reflecting the dependency relationships from event 9 (connection6) to event 16 (dependency10) and from the event 16 to the event 2 (dependency7) being removed.

(e) The dataset after the subset reflecting the dependency relationship from event 4 (other2) to event 9 (connection6) being removed.

(f) The dataset after the subset reflecting the dependency relationships from event 17 (stop11) to event 2 (dependency7) and from event 3 (other5) to event 5 (dependency6) being removed.

(g) The dataset after the subset reflecting the dependency relationship from event 12 (stop8) to event 8 (configuration8) being removed.

(h) The dataset after the subset reflecting the concurrent relationship from event 9 (connection6) to event 2 (dependency7) and event 3 (other5) being removed.
ERN in the Case Study
Other Projects in My Group

• Intelligent Help Desk
• Database Navigation
• Music Information Retrieval
• Computing System Management
• Business Intelligence
  – Leading KPI Discovery
• Bioinformatics
• Social Network Analysis
• **Operational/business Intelligence**
  – Collecting and analyzing metrics (Key Performance Indicators) to support better decision-making
  • Focus: leading indicator discovery
  – Using operational metrics to adaptively manage & improve workflow efficiency
  • Focus: workflow exception detection and tracing
Examples of KPI Types

<table>
<thead>
<tr>
<th>Leading Indicators (Value Drivers)</th>
<th>Lagging Indicators (Outcomes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of clients that sales people meet face to face each week</td>
<td>Sales revenue</td>
</tr>
<tr>
<td>Complex repairs completed successfully during the first call or visit</td>
<td>Customer satisfaction</td>
</tr>
<tr>
<td>Number of parts for which orders exceed forecasts within 30 days of scheduled delivery</td>
<td>Per unit manufacturing costs</td>
</tr>
<tr>
<td>Number of customers who are delinquent paying their first bill</td>
<td>Customer churn</td>
</tr>
</tbody>
</table>
Methodology for Leading Indicator Learning

- **Similarity matrix** (or Distance matrix) using non-metric distance or Granger Causality
- **Reduced Time series**
  - correlation
  - dimensionality reduction
  - define KPIs
- **Raw data**
  - Production metrics
  - Workflow activity logs
  - Job/Device/Document events

**Selected Leading Indicators or Leading Indicator Combinations**

- **Clustering**
- **Sequence KPIs**
- The root KPI may be the key leading indicator

**Leading Indicator Selection based on Domain Knowledge**

- **New Business Concern(s)**
  - If Model Valid?
  - If no, update model
  - If yes
  - **Leading Indicators**
  - **Selected Leading Indicators or Leading Indicator Combinations**
Some Projects in My Group

- Intelligent Help Desk
- Database Navigation
- Music Information Retrieval
- Computing System Management
- Social Network Analysis
- Business Intelligence
  - Leading KPI Discovery
- Bioinformatics
- Social Network Analysis
- Matrix-Based Algorithms for Data Mining
Biological Data Sources

Sequence Data

Protein-protein Interactions

Gene Ontology

Gene Expression

Text Literature

Protein Expression

Transcription Factor Binding Sites

Clustering

Gene Clusters

Function

Co-regulation
Problem We are Studying

• Marker Gene Identification
  – 15 gene selection algorithms
    • ReliefF, F-statistic, A-optimality, D-Optimality, mRNR, ReliefF-mRNR, GSNR
    • RankGene 1.1

• How to perform clustering using data from multiple, diverse sources?

• How to perform clustering using partial, prior knowledge?
Some Projects in My Group

- Intelligent Help Desk
- Database Navigation
- Music Information Retrieval
- Computing System Management
- Business Intelligence
  - Leading KPI Discovery
- Bioinformatics
- Social Network Analysis
- Matrix-Based Algorithms for Data Mining
Business Continuity Information Network (BCIN)
BCIN

• Central question
  – How to help businesses recover and resume operations quicker after a major hurricane (or other) disaster??
**Situation Analysis**

<table>
<thead>
<tr>
<th>John Smith</th>
<th>Asst. Manager</th>
<th>Store# 2345</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Status</strong></td>
<td><strong>Actions</strong></td>
<td><strong>Email</strong></td>
</tr>
<tr>
<td><strong>Recommendation Summaries</strong></td>
<td><strong>SA Overall</strong></td>
<td></td>
</tr>
<tr>
<td>Staff: Inadequate</td>
<td>10/1/07 12:32pm</td>
<td>J. Smith</td>
</tr>
<tr>
<td>Facility: Non-Functional</td>
<td>10/1/07 10:01am</td>
<td>R. Carr</td>
</tr>
<tr>
<td>Goods: Inadequate</td>
<td>10/1/07 10:58am</td>
<td>B. Falls</td>
</tr>
<tr>
<td>Trans: Functional</td>
<td>10/1/07 11:45am</td>
<td>L. Staff</td>
</tr>
</tbody>
</table>

**FACILITY: Building & Grounds**

- **Roof:** Damaged 10/1/07 7:15am R. Garcia
- **Windows:** Damaged 10/1/07 7:58am R. Garcia
- **Door:** Damaged 10/1/07 6:50am R. Garcia
- **Security:** N/A
- **Water:** N/A
- **Parking:** Damaged 10/1/07 6:40 am R. Garcia

Rcmd: Non-Functional 10/1/07 10:01am R. Carr
Comment: Parking lot may drain in 24hrs.
Acknowledgements

• Funding
  – NSF Career Award
  – NIH
  – Faculty Research Awards from IBM, Xerox and NEC
  – Equipment Donation from IBM

• Collaborators
  – Dr. Chris Ding (University of Texas at Arlington)
  – Dr. Michael I. Jordan (UC Berkeley)
  – Dr. Charles Perng (IBM Research)
  – Dr. Tong Sun (Xerox Research)
  – Dr. Shenghuo Zhu (NEC Research)

• Ph.D. Students
  – Dr. David Kaiser
  – Dr. Wei Peng
  – Bo Shao
  – Dingding Wang
  – Haifeng Wang
  – Yi Zhang
Question?

• Email: taoli@cs.fiu.edu
• http://www.cs.fiu.edu/~taoli

Thank You!