D2K and Data Analytics

Infectious Disease Workshop

September 7, 2006

Loretta Auvil
Automated Learning Group
National Center for Supercomputing Applications
University of Illinois
217. 265.8021
lauvil@ncsa.uiuc.edu
Outline

- Motivation for D2K
- Analytics
- Some Current Projects Leveraging D2K
Motivation for D2K

Create a framework that could be used and extended for our needs in knowledge discovery (data mining).

D2K is a flexible data mining system that integrates effective analytical data mining methods for prediction, discovery, and anomaly detection with data management and information visualization.

Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

The understandable patterns are used to:
- Make predictions about or classifications of new data
- Explain existing data
- Summarize the contents of a large database to support decision making
- Create graphical data visualization to aid humans in discovering complex patterns
Knowledge Discovery Process

An Overview of the Steps That Compose the KDD Process

1. Data Selection & Cleaning
2. Data Prep
3. Transformation
4. Data Mining / Pattern Discovery
5. Interpretation / Knowledge
Required Effort for each KDD Step

Arrows indicate the direction we want the effort to go.
Three Primary Paradigms

- **Predictive Modeling** - supervised learning approach where classification or prediction of one of the attributes is desired.
  - Classification is the prediction of predefined classes
    - e.g. Naive Bayesian, Decision Trees, and Neural Networks
  - Regression is the prediction of continuous data
    - e.g. Neural Networks, and Decision (Regression) Trees

- **Discovery** - unsupervised learning approach for exploratory data analysis.
  - e.g. Association Rules, Link Analysis, Clustering, and Self Organizing Maps

- **Deviation Detection** - identifying outliers in the data.
  - e.g. Visualization
Advantages of a Framework for Analytics

- Provides scalable environment from the Desktop to Web Services
- Employs a visual programming system for data/work flow paradigm
- Provides capability to build custom applications
- Provides capability to access data management tools
- Contains data mining algorithms for prediction and discovery
- Provides data transformations for standard operations
- Integrated environment for models and visualization
- Supports an extensible interface for creating one’s own algorithms
- Provides access to distributed computing capabilities
- Employs multi-layered learning strategies
D2K and Its Many Components

- **D2K Infrastructure**
  D2K API, data flow environment, distributed computing framework and runtime system
- **D2K Modules**
  Computational units written in Java that follow the D2K API
- **D2K Itineraries**
  Modules that are connected to form an application
- **D2K Toolkit**
  User interface for specification of itineraries and execution that provides the rapid application development environment
- **D2K-Driven Applications**
  Applications that use D2K modules with a custom user interface
- **D2K Streamline (SL)**
  Task driven system that uses D2K modules
- **D2K Web Services**
  Enables web deployment
Stream Data Analytics/Information Visualization

Online Stream Query Engine
- Uses novel methods to do real-time stream data analysis.

Online Stream Classification
- Adaptable to the changes and evolution of data streams.

Online Frequent Pattern Mining
- Discovers association and correlation rules in data stream environment.

Online Clustering of Data Streams
- Detects outliers and finds evolution of clusters in data streams.
Rule Association: Overview

- Unsupervised learning problem.
- Find all rules that correlate the presence of one set of items X with another item Y.
  - Example: When a customer buys bread and butter, they buy milk 85% of the time.
- Support is the percentage of the records that contain both X and Y.
  - A rule must have some minimum user-specified support to show its impact.
- Confidence is the percentage of records that contain X and Y out of the number of records that contain X.
  - A rule must have some minimum user-specified confidence to show its value.
Rule Association: Strengths and Weaknesses

Strengths

• It produces easy to understand results.
• It supports undirected data mining.
• It works on variable length data.
• Rules are relatively easy to compute.

Weaknesses

• It produces many rules.
• For large numbers of attribute-value combinations, considerable cpu and memory resources are consumed.
Rule Association: Visualization

- Read rules down the column.
- Example - the first rule is
  - If petal-length Binned=[...:2.] and petal-width Binned=[...:0.7] then flower-type=Iris-setosa
    - Support = 25%
    - Confidence = 100%
- Use brushing to find out support and confidence.
- Click on the Confidence label to sort by confidence.
- Click on the Support label to sort by support.
- Additional functionality for searching/sorting is planned.
Exploratory Analysis for GeoSentinel

- Exploratory evaluation indicates that interesting patterns can be detected in the data using frequent pattern analysis to generate hypothesis.
Evolution Highway: Vis for Chromosome Comparison

- Uses the D2K web service to deliver the visualization application
- Provides a visual means for simultaneously comparing mammalian genomes of humans, horses, cats, dogs, pigs, cattle, rats, and mice
- Removes the burden of manually aligning these maps
- Allows cognitive skills to be used on something more valuable than preparation and transformation of data
- Adding analytical comparisons for pattern finding
- evolutionhighway.ncsa.uiuc.edu went live on July 22, 2005
Phylomat

- Uses the D2K web service to create and execute the job, and present the results.
- Finds every protein in each proteome that contains a specific consensus sequence (motif).
- Aligns motifs to derive clustering of motifs based on similarity.
- Evolutionary trends from protein groups can then be inferred from the data and insight into the evolutionary origins or the motif class can be hypothesized.
- alg.ncsa.uiuc.edu/pmat
NCASSR Streaming Text: Knowledge Extraction

- Leveraging some earlier work on information extraction from text streams

Information extraction
- process of using advanced automated machine learning approaches
- to identify entities in text documents
- extract this information along with the relationships these entities may have in the text documents

This project demonstrates information extraction of names, places and organizations from real-time news feeds. As news articles arrive, the information is extracted and displayed.
NCASSR Streaming Text: Semantic Text Processing Diagram

**Ingest Management**
- Document Streams
- Document Queue

**Text Analytics**
- Syntactic Processing
- Semantic Processing
- Logical Processing

**Data Store**
- RDF Semantic Graph Store

**Visual User Interface**

**Knowledge Discovery**
- Question/Answer
- Frequent Patterns
- Visual Analytics

**Filtering**
- By topic, date, location, etc.

**Semantic Graph Generation**
NCASSR Streaming Text: Semantic Text Itinerary and GUI
Grids are being built to work with distributed earth, space and environmental science data stores. A next step is to undertake distributed data analysis utilizing remote data.

- GeoLearn
  - Praveen Kumar, Hydrology
- EMO - Evolutionary-based Multiobjective Optimization for Hazard Management
  - Barbara Minsker, Civil and Environmental Engineering
- MAEViz - Multi-modal Data Integration and Information Visualization
  - Dan Abrams, Civil Engineering
Now that the human genome has been sequenced, attention is turning to the mining of proteomic and structural biological data and looking for patterns that arise when examining data from a wide variety of different “omic” data sets.

- **Evolution Highway**
  - Harris Lewin, Animal Science
- **Phylomat**
  - Rex Gaskins, Cell and Structural Biology
- **Disease Susceptibility**
  - Larry Schook, Animal Science
- **Constructing Biological Networks**
  - David Rivier, Cell and Structural Biology
Although science is leading the way, the exploring, analyzing, and mining of social science data stores is beginning to change these fields, too.

- Nora - Mining Literary Works
  - John Unsworth, GSLIS
- Distributed Innovation and Scalable Collaboration in Uncertain Settings (DISCUS)
  - David Goldberg, General Engineering
- Music Information Retrieval - MIR
  - Stephen Downey, Graduate School LIS
- Ticket To Work, Job Demands
  - Tayna Gallagher, College of Applied Life Studies
- Mining Bugzilla
  - Les Gasser, Graduate School LIS
- Multi-modal Global Economic Modeling
  - Gerald Nelson, AG and Consumer Economics
- Concept Modeling in War Periodical
  - Bruce Rosenstock, LAS-Religion
Mining homeland defense data is difficult because the data is massive, distributed, complex and heterogeneous.

- Mining Alarming Incidents in Data Streams  
  - Jiawei Han, Computer Science
- Distributed Innovation and Scalable Collaboration in Uncertain Settings  
  - David Goldberg, General Engineering
- NIBRS - Mining the National Incident Based Reporting System  
  - Tracy McGee, Illinois State Police
- Intelligence Gathering from API New Feeds
# The ALG Team

## Staff
- Bernie Acs
- Loretta Auvil
- David Clutter
- Vered Goren
- Eugene Grois
- Xavier Llora
- Luigi Marini
- Robert McGrath
- Greg Pape
- Barry Sanders
- Andrew Shirk
- David Tcheng
- Michael Welge

## Students
- Chen Chen
- Hong Cheng
- Yaniv Eytani
- Fang Guo
- Qian Yang
- Feida Zhu