NAS Workshop Panel

Major Challenges of Data Mining and Search

Chid Apte

Manager, Data Abstraction Research
T.J. Watson Research Center
IBM Research Division
http://www.research.ibm.com/dar

April 29, 2000

Most challenging and essential research problems in data mining

- It would cause a revolution in data mining if we could solve the problem of ....

Data Mining - A Business Intelligence Perspective

- Extracting actionable insights from data
  - frequent patterns, predictive models, clusters
- Incorporate in decision support solutions
  - risk management
  - targeted marketing
  - web personalization
- Current state of art
  - ► data must be mapped into a standard problem such as classification, regression, clustering
  - ► many strong methods can then be applied (to mostly uniform measurements)
- Revolutionary
  - ► Automation of all the steps; here's the raw data, here are the actionable results

Data

- Verifying data integrity
  - Is it just noise ?
    - methods for detecting presence of signal
  - Is it backfitted (past data updated based on future) ?
    - methods for detecting unusual correlations

Representation

- What is the problem to be solved ?
  - Posing the right problem is more important than the subsequent methods we use
- What are the set of defining features ?
  - feature transformation, creation, selection
    - methods for selecting the best transforms

What will make data mining ubiquitous?

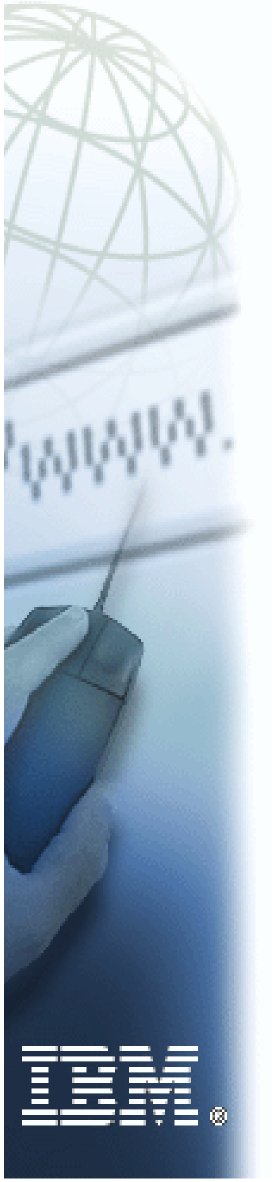
- Automation & quality are the keys
  - How automatic can the modeling be ?
    - searching for alternate models within modeling vocabulary space to find the optimum
      - methods for heuristic search
  - How accurate can the models be ?
    - robust techniques that can scale and handle heterogeneous data
      - methods for integrating learning techniques and data management technology
  - How understandable can the models be ?
    - does human comprehension necessitate sacrifices in model accuracy ?

Other Open Areas

- Clustering -- metric based, goal directed
  - data segmentation needs better evaluation metrics
- Privacy preserving data mining
  - how to take out all specifics and yet retain the statistical properties to enable pattern detection and model formulation
- Unstructured (text) data mining
  - combining information retrieval, natural language processing, and data mining methods
- Sequence / ordering based pattern detection
  - representational issues for temporal data streams
  - handling order-based constraints in patterns
- Online Learning
- Non-stationary data over time
  - when to discard old data

Where will the breakthroughs come from?

- Interdisciplinary
  - Statistics, Machine Learning, Data Management



NAS Workshop Panel

Major Challenges of Data Mining and Search

Chid Apte
Manager, Data Abstraction Research
T.J. Watson Research Center
IBM Research Division
http://www.research.ibm.com/dar

April 29, 2000