Data Mining Research: Successes vs. Failures or Achievements vs. Lessons

Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign

August 20, 2007

18 Years of KDD Research

- 1989 Workshop on KDD: less than 30 participants
- Today, an army of researchers and industry practitioners
- Related conferences and journals
  - ACM SIGKDD
  - IEEE ICDM
  - SIAM Data Mining Conf.
  - PKDD
  - PAKDD
  - DB, ML, WWW (WSDM), PR, Stat conf's
  - DAMI (DMKD)
  - IEEE TKDE
  - ACM TKDD
  - Statistical Analysis and Data Mining
  - Other journals on ML, Web, Stat, DB, …

Major Achievements

- Scalable data mining methods
- Power demonstrated in a killer application: Web search and analysis
- Pattern mining penetrating into data mining tasks
- Data mining claims its success in many applications
- Invisible data mining: An integral part of some systems
- The growth of new workforce on data mining research and development
- Public interests and public knowledge
- Curriculum: Learning and education
- The disappearing boundary between data mining, machine learning, statistical analysis, Web search/mining, etc.

Scalable Mining Methods

- Scalable Pattern Mining Methods
  - Frequent pattern minning
    - Apriori, FPgrowth, EClat, closed itemset, FPClose, …
    - Mining long patterns by pattern fusion
  - Sequential pattern mining: GSP, PrefixSpan, Clospan, BIDE
  - Graph pattern mining: gSpan, CloseGraph, …
  - Constraint-based mining:
  - Patterns for clustering, feature extraction, classification
- Efficient Classification Methods: SVM, RainForest, BOAT, …
- Scalable Clustering Method: Clarans, BIRCH, Optics,…

Fast Expanding of Applications

- Web Search and Mining
  - PageRank and HITS: Google, Google News, …
  - Links contain rich information
  - Lots of in-depth mining done on web analysis
- Stream and sensor network mining
- Spatiotemporal mining, moving object data mining
- Social network analysis
- Data mining in bioinformatics and medical science
- Data mining for science and engineering (e.g., civil eng.)
- Data mining for software bug analysis and system improvements

Major Lessons

- Data mining: Poor public image or misunderstanding?
  - Violation of person's privacy and security?
  - Police in society?
- Large amount of real data is still not touched!
- Generating more rules than the original data?
- Yet another delta algorithm with tiny improvement?
- More success stories?!
- Where are many killer, flagship systems and applications?
- Where are killer, invisible data mining applications
- No industrial or well-accepted standards or benchmarks
- Data mining foundation: A science, technology, or an art?

Look into the Near Future

- Data mining, not a curse, not a mystery, not misunderstanding, but a pleasant, common practice
- Education: A basic skill for every undergraduate student in CS, business, statistics, and science & engineering
- Put DM in a user-friendly, multidimensional explorative (i.e., OLAP-like) environment
- Invisible data mining, penetrating many more daily applications