Data Mining
Information Retrieval
Web Search
Logistics

- Instructor: Byron Gao
  - http://cs.txstate.edu/~jg66/
  - Contact, office hours ...
  - Course page

- Database prerequisite?

- Textbook
  - Jiawei Han, Micheline Kamber and Jian Pei
  - Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze

- Workload

- Grading: curve
  - Weighting
  - Late policy, academic honesty
Course Project

- Type I: mobile + data mining/IR
  - Parking lot app
  - Craigslist faceted search app
- Type II:
  - Closed: implementation of selected data mining algorithm in the textbook; using Yahoo or Google API for web search
  - Open: stock prediction, KDD cup, practical data analytics such as in MIT Sloan Sports Analytics Conference (http://en.wikipedia.org/wiki/MIT_Sloan_Sports_Analytics_Conference)
  - Correctness: must; Utility: preferred; Novelty: a plus, but not necessary
- Examples (last year’s proposed projects)
Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary
Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, ...
    - Science: Remote sensing, bioinformatics, scientific simulation, ...
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets
Interest in Data Mining

- Sergey Brin
  - database lab at Stanford
  - http://infolab.stanford.edu/~sergey
  - “major research interest is data mining …”

- Jeff Ullman (~Larry Page) wikipedia entry
  - “Ullman's research interests include … data mining”

- Jon Kleinberg
  - active in KDD community

- Fast growing:
  - data-rich
  - knowledge-poor

- Job opportunities: Google, Yahoo, Microsoft …
  - http://labs.google.com/
Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. **Data mining** is a major new challenge!
Evolution of Database Technology

- **1960s:**
  - Data collection, database creation, IMS and network DBMS

- **1970s:**
  - Relational data model, relational DBMS implementation

- **1980s:**
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)

- **1990s:**
  - Data mining, data warehousing, multimedia databases, and Web databases

- **2000s**
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems
Data Mining Community

- At Texas state
  - Data mining lab
  - Other faculty members ...

- Worldwide
  - US: Stanford, UIUC, Wisconsin, Minnesota
  - Canada: SFU, UBC, U of A
  - Europe: LMU, U of Helsinki
  - Asia: NUS
  - ...

- KDD: knowledge discovery and data mining
What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?

- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

- Watch out: Is everything “data mining”?
  - Simple search and query processing
  - (Deductive) expert systems
Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities.
- Data mining plays an essential role in the knowledge discovery process.

Diagram:
- Data Cleaning
- Data Integration
- Databases
- Data Warehouse
- Task-relevant Data
- Selection
- Data Mining
- Pattern Evaluation
Data Mining in Business Intelligence

- Increasing potential to support business decisions

- Decision Making
  - Data Presentation
    - Visualization Techniques
  - Data Mining
    - Information Discovery
  - Data Exploration
    - Statistical Summary, Querying, and Reporting
  - Data Preprocessing/Integration, Data Warehouses
  - Data Sources
    - Paper, Files, Web documents, Scientific experiments, Database Systems

- End User
- Business Analyst
- Data Analyst
- DBA
Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks

- **Knowledge to be mined (or: Data mining functions)**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels

- **Techniques utilized**
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.

- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.
Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
  - Object-relational databases, Heterogeneous databases and legacy databases
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and information networks
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web
Data Mining Functions: Generalization

- Information integration and data warehouse construction
  - Data cleaning, transformation, integration, and multidimensional data model
- Data cube technology
  - Scalable methods for computing (i.e., materializing) multidimensional aggregates
  - OLAP (online analytical processing)
- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region
Data Mining Functions: Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer [0.5%, 75%] (support, confidence)
    - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?
Data Mining Functions: Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels

- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, ...

- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, ...
Data Mining Functions: Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications
Data Mining Functions: Outlier Analysis

- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Noise or exception? — One person’s garbage could be another person’s treasure
  - Methods: by product of clustering or regression analysis, ...
  - Useful in fraud detection, rare events analysis
Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- Sequence, trend and evolution analysis
  - Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., first buy digital camera, then buy large SD memory cards
- Periodicity analysis
- Motifs and biological sequence analysis
  - Approximate and consecutive motifs
  - Similarity-based analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams
Structure and Network Analysis

- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)

- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, ...
  - Links carry a lot of semantic information: Link mining

- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, ...
Data Mining: Confluence of Multiple Disciplines

- Machine Learning
- Pattern Recognition
- Statistics
- Applications
- Visualization
- Algorithm
- Database Technology
- High-Performance Computing
Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of science and information technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of data
- Data mining functionalities: characterization, discrimination, association, classification, clustering, trend and outlier analysis, etc.
- Data mining technologies and applications
- Major issues in data mining
Major Issues in Data Mining (1)

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: An interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided mining

- User Interaction
  - Interactive mining
  - Incorporation of background knowledge
  - Presentation and visualization of data mining results
Major Issues in Data Mining (2)

- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- Data mining and society
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining
10. Top-10 Most Popular DM Algorithms: 18 Identified Candidates (I)

Classification


Statistical Learning

- #8. FP-Tree: Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In SIGMOD '00.
The 18 Identified Candidates (II)

- Link Mining

- Clustering

- Bagging and Boosting
The 18 Identified Candidates (III)

- Sequential Patterns
  - #15. PrefixSpan: J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In ICDE '01.

- Integrated Mining
  - #16. CBA: Liu, B., Hsu, W. and Ma, Y. M. Integrating classification and association rule mining. KDD-98.

- Rough Sets

- Graph Mining
  - #18. gSpan: Yan, X. and Han, J. 2002. gSpan: Graph-Based Substructure Pattern Mining. In ICDM '02.
Top-10 Algorithm Finally Selected at ICDM’ 06

- #1: C4.5 (61 votes)
- #2: K-Means (60 votes)
- #3: SVM (58 votes)
- #4: Apriori (52 votes)
- #5: EM (48 votes)
- #6: PageRank (46 votes)
- #7: AdaBoost (45 votes)
- #7: kNN (45 votes)
- #7: Naive Bayes (45 votes)
- #10: CART (34 votes)
A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD’ 95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
- ACM Transactions on KDD (2007)
Conferences and Journals on Data Mining

- **KDD Conferences**
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (**KDD**)
  - SIAM Data Mining Conf. (**SDM**)
  - (IEEE) Int. Conf. on Data Mining (**ICDM**)
  - European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (**ECML-PKDD**)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (**PAKDD**)
  - Int. Conf. on Web Search and Data Mining (**WSDM**)

- **Other related conferences**
  - DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, ...
  - Web and IR conferences: WWW, SIGIR, WSDM
  - ML conferences: ICML, NIPS
  - PR conferences: CVPR,

- **Journals**
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD
Where to Find References? DBLP, CiteSeer, Google

- **Data mining and KDD (SIGKDD: CDROM)**
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD

- **Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)**
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA

- **AI & Machine Learning**
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.

- **Web and IR**
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: WWW: Internet and Web Information Systems,

- **Statistics**
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.

- **Visualization**
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.
Recommended Reference Books

- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011
- B. Liu, Web Data Mining, Springer 2006
- Y. Sun and J. Han, Mining Heterogeneous Information Networks, Morgan & Claypool, 2012
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005
- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998