Web Algorithms and Applications

Moon Jung Chung
• Prerequisites: CPS830, CS835, or CSE860


• Reference: Data Mining: Concepts and Techniques by J. Han, M. Kamber, Morgan Kaufmann Publishers, 1st edition (August 2000)
Grading and Project

• Grading
  – Homework 20%
  – Midterm  30%
  – Project  30%
    ◆ Mid Progress Presentation
    ◆ Midterm Report
    ◆ Final Report
    ◆ Final Presentation
  – Class Participation 20%  -- Presentations and Discussions

• Project: A team of 2-4 students will select a topic and work on the project. Each student will present background papers and results of the project in the class.
Outline

• Web Mining: From a huge collection of data, how to reduce it to meaningful information?

• Searching: A query -> extract information from data base classification

• Better web searching
  provide intelligent services such as advice, inference, etc.
  Find travel agents who handle ski trip to Aspen with Ski-in ski-out accommodation, lift tickets, but no air

• Behaviors on a Web (Negotiations, Auctions, etc)

• Performance of Network Communications
Topics

Part I: Data Mining Basics
- Mining Association Rules
- Apriori Algorithm
- Classification and Prediction – Decision Tree Induction, Bayesian Classification
- Clustering
Part II: Graph models of the Web and the Internet

- Algorithms for analyzing the web graph
- Graph oriented statistical sampling of the web
- Practical discovery algorithmic techniques on the Web
- Mining web link structure – Hub, Authority
- Mining Web Content
- Mining Web Usage
- Dimension Reduction
- Application to web-searching, information retrieval on the web, data mining.
- Web mining and Internet security
PART III: Game theory and economics of the Internet

- Routing on the Internet
- Game theory
- Nash's equilibrium
- Network design with selfish agents
- Fixed point models for performance engineering in enterprise IP networks
- Performance and measurement issues in the Web
- Empirical studies and issues
- Congestion pricing and capacity expansion games
Nature of the Web

- No standards, unstructured and heterogeneous
- Growing and changing very rapidly
  - One new WWW server every 2 hours
- Size
  - >350 million pages (1999)
  - Grows at about 1 million pages a day
  - Google indexes 3 billion documents
- Indices get stale very quickly

Need for better resource discovery and knowledge extraction.
Web Algorithms and Other Disciplines

• DBMS
• Information Retrieval
• AI: Agents, Knowledge representation
• Visualization
• Mathematical modeling
• Game Theory
• *Parallel and Distributed Computing*
• Security
Web Mining

- **Huge:** The “abundance” problem:
  - too huge for effective data warehousing and mining
  - 99% of the Web information is useless to 99% of users.
- **Unstructured:** *Complexity of Web pages:* far greater than text document collection
- **Dynamic:** information constantly updated.
- **Limited coverage** of the Web (hidden Web sources)
- **Limited query interface:** keyword-oriented search
- **Limited customisation** to individual users
Web Mining: On What Kind of Data?

• The content of the Web -- **Content Mining**
  – Getting the essence from within web pages

• The structure of the Web -- **Structure Mining**
  – extract knowledge from the interconnections of hypertext documents in the Web
  – Who is the authority? -- College Football Ranking
  – Who is the hub? --

• The usage of the Web -- **Usage Mining**
  – Exploiting Web access logs
Data Mining Functionality

• **Characterization:** summarization of general features
  
  *Characterize Bush supporters*

• **Discrimination:** Comparison of general features of two objects
  
  *Compare supporters between Bush and Gore*

• **Association:** Studies the frequency of items occurring together
  
  *buy (lexus) -> support(bush)*
  
  *buy (x, lexus) -> buy (x, criminal lawyer)*