Motivation

- Companies and institutes usually have amassed a huge amount of sales data
  - Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases
  - these data may contain valuable business information
- Various applications have been ported to the Web at rapid pace
An Overview of Data Mining

- We are buried in data, but looking for knowledge
- Data mining: Knowledge discovery in databases
  - extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases
Reference Books

• Reference book
  – Class notes will be made available to you before each class
  – Data mining: Concepts and Techniques, by Jiawei Han and Micheline Kamber

Course Requirement

• Had better have backgrounds on
  – Database
  – Algorithm
  – Fundamental Web technology
  – Programming in C/C++
  – Knowledge in Java
Course Materials

- Data mining and data warehousing
- Association rules and generalized association rules
- Classification rules
- Data generalization process

Course Materials (cont’d)

- Mining sequential patterns
- Use of tools for data mining
- Mining Web applications
- Data clustering techniques
Class Announcement

• A class Web page is ready
  – Access from my Web page
  – Students are required to have email account and to check class Web page for announcement
  – Will try to make class materials ready there

Class Announcement

• One grader: Chin-Huang Yun at 306
• Grading (tentative):
  – HW 30%
    • at least 6
  – Final 35%
  – Project 35%, either of the following is allowed
    • with implementation: group of up to 3
    • in depth research paper: individual
    • abstract is due 4/29/01
Correspondence

- Email: mschen@cc.ee.ntu.edu.tw (preferred)
  http://www.ee.ntu.edu.tw/~mschen
- Tel: (02) 23635251 ext 523
  Fax: (02) 23638247

Mining Capabilities

- Association
- Classification and Clustering
- Traversal patterns
- Sequential patterns
- Data generalization, summarization
Notes

- Data mining is very application dependent
  - Small team with good skill and domain knowledge
- Lots of work has been done in other areas
- Emerging issues:
  - New journals, KDD, SIGMOD, ICDE, VLDB, CIKM, etc.
  - Special issues

Where to Find References?

- Data mining and KDD
  - Conference proceedings: SIG-KDD, and others, such as ICDM, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery
- Database field
  - Conference proceedings: ACM-SIGMOD, ACM-PODS, VLDB, ICDE, EDBT, DASFAA
  - Journals: ACM-TODS, J. ACM, IEEE-TKDE, etc.
- AI and Machine Learning:
  - Conference proceedings: Machine learning, AAAI, IJCAI, ICTAIetc.
  - Journals: Machine Learning, Artificial Intelligence, etc.
Data Warehousing

- Data warehousing an architectural foundation for decision support system, consisting of
  - integrated data, detailed and summarized data, historical data, and metadata
- Set up stages for effective data mining

OLAP

- On-Line Analytical Processing: simple data mining facility
  - responds to queries quickly
- A multidimensional, logical view of the data.
- Interactive analysis of the data: drill down and roll-up, etc.
- Operations for
  - analytical modeling: deriving ratios, variance, etc.
  - summarization and aggregations at every dimension intersection.
Challenges for Data Mining

- Handling of different types of data
- Efficiency and scalability of mining algorithms
- Usefulness and certainty of mining results
- Expression of various kinds of mining results

Challenges (cont’d)

- Interactive mining at multiple abstraction levels
- Mining information from different sources of data
- Protection of privacy and data security
Classifying Mining Techniques

- What kinds of databases
- What kinds of knowledge to be mined
- What kinds of techniques to be used

Mining Association Rules

- Transaction data analysis: Mining association rules
  - Given: (1) a database of transactions
    (2) each tx has a list of items purchased
- Find all asso. rules: the presence of one set of items implies the presence of another set of items
  - people who purchased hammers also purchased nails
Two Parameters

- Confidence (how true)
  - the rule X&Y => Z has 90% conf. means 90% of customers who bought X and Y also bought Z
- Support (how useful is the rule)
  - useful rules should have some minimum tx support

Mining Strong Association Rules in Transaction DBs

- Measurement of rule strength in a transaction DB.
  \[ A \rightarrow B \ [\text{support, confidence}] \]
  \[
  \text{support} = \text{Prob}(A \cup B) = \frac{\# \text{ of trans containing all the items in } A \cup B}{\text{total } \# \text{ of trans}}
  \]
  \[
  \text{confidence} = \text{Prob}(B|A) = \frac{\# \text{ of trans that contain both } A \text{ and } B}{\# \text{ of trans containing } A}
  \]
- We are often interested in only strong associations, i.e.
  \[ \text{support} \geq \text{min}_\text{sup} \text{ and } \text{confidence} \geq \text{min}_\text{conf} \]
- Examples.
  - milk \rightarrow bread [5%, 60%].
  - tire \& auto_accessories \rightarrow auto_services [2%, 80%].
Two Steps for Mining Asso.

- Determining “large itemsets”
  - the main factor for overall performance
- Generating rules
Two Steps for Mining Asso.
(cont’d)

for each large itemset m do
  for each subset p of m do
    if (sup(m)/sup(m-p) >= minconf) then
      output the rule (m-p) => p
      with conf = sup(m)/sup(m-p) and
      support = sup(m)

m = {a,c,d,e,f,g} 2000 tx’s
p = {a,d} 5000 tx’s
{a,d} => {c,e,f,g} conference: 40%, support: 2000
  tx’s

Redundant Rules

For the same support and confidence, if we have a rule {a,d} => {c,e,f,g}, do we have
[agga98a]
- {a,d} => {c,e,f}
- {a} => {c,e,f,g}
- {a,d,c} => {e,f,g}
- {a} => {d,c,e,f,g}
Methods for Mining Association Rules

• Apriori (Agrawal & Srikant’94).
  – derivation of large 1-itemsets L1: At the first iteration, scan all the transactions and count the number of occurrences for each item.
  – level-wise derivation: At the k-th iteration, the candidate set $C_k$ are those whose every $(k - 1)$-item subset is in $L_{k-1}$. Scan DB and count the # of occurrences for each candidate itemset.
  – the cardinality of C2 is huge
  – the exe time for the first 2 iterations is the dominating factor to overall performance

M.-S. Chen NTU Winter 2001

Methods for Association Rules (cont’d)

• DHP (direct hashing with pruning)
  – Apriori + hashing (Park, Chen and Yu’95)
  – use hash based method to reduce the size of C2
  – allow effective reduction on tx database size (tx number and each tx size)

• Variations of association rules have been explored, e.g., [agga98d]
HW#1

- Suppose we additionally have
  - 500 ABCE
  - 600 CD
  - support=3 tx’s (50%)
  - confidence = 66%
- Repeat the large itemset generation
- Derive up to 4 rules
- Due 2/28/00

Reading Assignments

Reading Assignments (cont’d)


