CS100: Introduction to Computer Science

Lecture 21: Database & Data mining

Statistics of the Second Exam
- 100 and above: 8
- 90 – 99: 7
- 80 – 89: 7
- 70 – 79: 5
- 60 – 69: 3

Review: Course Objectives
- Fundamental understanding of the field,
- Experience with programming, and
- Research topics and applications
  - Database & Data Mining
  - Information Retrieval & Question Answering & Web Search

Review: Course Objectives
- Fundamental understanding of the field,
- Experience with programming, and
- Research topics and applications
  - Database & Data Mining
  - Information Retrieval & Question Answering & Web Search

Database Systems
- **Database**
  - A collection of related records
  - The conceptual view: table with rows and columns
- **Schema**
  - A structural description of the type of facts held in a database
  - E.g. (EmployeeID, Name, Address, Salary, JobNo)
  - Your own example?
- **Database models**: modeling database structure

Relational Database Models
- **Relational model**
  - Dominant in commercial data processing systems
- **Hierarchical model**
  - A tree structure, parent-child relationships, 1:m mapping
- **Network model**
  - More than 1 parent per child, m: m mapping
- **Object-oriented model**
  - Add database functionality to object programming language.
Relational Database Model

- **Relation**: A rectangular table
- **Attribute**: A column in the table
- **Tuple**: A row in the table

A relation containing employee information

<table>
<thead>
<tr>
<th>EmpId</th>
<th>Name</th>
<th>Address</th>
<th>SSN</th>
</tr>
</thead>
<tbody>
<tr>
<td>25X15</td>
<td>Joe E. Baker</td>
<td>33 Nowhere St.</td>
<td>11122333</td>
</tr>
<tr>
<td>23Y34</td>
<td>G. Jerry Smith</td>
<td>555 Circle Dr.</td>
<td>11100555</td>
</tr>
</tbody>
</table>

A relation containing redundancy

<table>
<thead>
<tr>
<th>EmpId</th>
<th>Name</th>
<th>Address</th>
<th>SSN</th>
</tr>
</thead>
<tbody>
<tr>
<td>34Y70</td>
<td>Cheryl H. Clark</td>
<td>560 Downtown Ave.</td>
<td>99999999</td>
</tr>
</tbody>
</table>

An employee database consisting of three relations

**EMPLOYEE relation**

<table>
<thead>
<tr>
<th>EmpId</th>
<th>Name</th>
<th>Address</th>
<th>SSN</th>
</tr>
</thead>
<tbody>
<tr>
<td>25X15</td>
<td>Joe E. Baker</td>
<td>33 Nowhere St.</td>
<td>11122333</td>
</tr>
<tr>
<td>23Y34</td>
<td>G. Jerry Smith</td>
<td>555 Circle Dr.</td>
<td>11100555</td>
</tr>
</tbody>
</table>

**JOB relation**

<table>
<thead>
<tr>
<th>EmpId</th>
<th>Job Id</th>
<th>Job Title</th>
<th>Skill Code</th>
<th>Dept</th>
</tr>
</thead>
<tbody>
<tr>
<td>25X15</td>
<td>2234</td>
<td>Secretary</td>
<td>F5</td>
<td></td>
</tr>
</tbody>
</table>

**ASSIGNMENT relation**

<table>
<thead>
<tr>
<th>EmpId</th>
<th>Job Id</th>
<th>Start Date</th>
<th>Term Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>25X15</td>
<td>2234</td>
<td>12/1/2001</td>
<td>12/1/2003</td>
</tr>
</tbody>
</table>

Relational Operations

- **Select**: Choose rows
- **Project**: Choose columns
- **Join**: Assemble information from two or more relations

The SELECT operation

```
NEW = SELECT from EMPLOYEE where EmpId = "34Y70"
```

NEW relation

<table>
<thead>
<tr>
<th>EmpId</th>
<th>Name</th>
<th>Address</th>
<th>SSN</th>
</tr>
</thead>
<tbody>
<tr>
<td>34Y70</td>
<td>Cheryl H. Clark</td>
<td>560 Downtown Ave.</td>
<td>99999999</td>
</tr>
</tbody>
</table>
The PROJECT operation

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe E. Baker</td>
<td>33 Nowhere St.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

MAIL = PROJECT Name, Address from EMPLOYEE

Another example of the JOIN operation

The JOIN operation

An application of the JOIN operation

Structured Query Language (SQL)
- Operations to manipulate tuples
  - insert
  - update
  - delete
  - select

SQL Examples

- select EmplId, Dept
  from ASSIGNMENT, JOB
  where ASSIGNMENT.JobId = JOB.JobId
  and ASSIGNMENT.TermData = "*"

- insert into EMPLOYEE
  values ('43212', 'Sue A. Burt',
          '33 Fair St.', '4446611111')
SQL Examples (continued)

- delete from EMPLOYEE
  where Name = 'G. Jerry Smith'

- update EMPLOYEE
  set Address = '1812 Napoleon Ave.'
  where Name = 'Joe E. Baker'

Set Operators (SQL)

- UNION ALL
  Combines the results of two SELECT statements into one result set.

- UNION
  Combines the results of two SELECT statements into one result set, and then eliminates any duplicate rows from that result set.

- MINUS
  Takes the result set of one SELECT statement, and removes those rows that are also returned by a second SELECT statement.

- INTERSECT
  Returns only those rows that are returned by each of two SELECT statements.

Database Systems

- A relational database management system
  - A software package used to create a database (Oracle, Microsoft SQL server, MySQL)

- Database applications
  - Human resource management system
  - Sales management system
  - Inventory management system
  - Decision support system

The conceptual layers of a database implementation

Data Mining

- What is data mining
- Basic data mining tasks

What is data mining?

- Data is growing at a phenomenal rate
- Users expect more sophisticated information
- How?

UNCOVER HIDDEN INFORMATION
DATA MINING
Basic Data Mining Tasks -- Classification

- Maps data into predefined groups or classes
  - **Example**
    - Credit card company must determine whether to authorize credit card purchases.
    - Four classes:
      - 1) Authorize,
      - 2) Ask for further identification before authorization
      - 3) do not authorize,
      - 4) do not authorize but contact police
    - How to classify a purchase?
    - Your own example?
- Techniques: Regression, Bayesian classification, Decision trees (20-questions game), Neural network

Basic Data Mining Tasks--clustering

- Groups similar data together into clusters. (The clusters are not predefined)
  - **Example**
    - A department store chain creates special catalogs targeted to various demographic groups based on attributes such as income, location, etc.

Basic Data Mining Tasks—Association Rules

- **Association Rules**
  - Identify items are frequently purchased together.
- **Example**
  - A grocery store retailer is trying to decide whether to put bread on sale. He finds that 60% of the time that bread is sold so are pretzels and 70% of the time jelly is also sold by using association rules.
  - Decisions?

Announcements:

- Homework 5: Due next Monday (May 7th).
  - Write a 2-page essay to discuss computer applications in your academic area.
  - For example: Economics, Biology, etc.
  - Cover current applications and discuss potential future applications
  - This written assignment should be typed.
- Lab 6 this week: (optional)
- Next lecture: Information Retrieval, Question Answering & Web search