Have You Ever …

- Wondered how products are placed in supermarket aisles?
- Had your application for a no-interest-for-6-months Titanium credit card rejected?
- Puzzled over the two-hour phone call to Belize on your phone bill?
- Gazed at the sky and wondered if that bright star is a white dwarf?
- Data mining has the answers!!!

What is Data Mining?

- Finding “interesting” patterns in large amounts of data.
- Data mining encompasses several areas:
  - Machine learning (AI)
  - Statistics
  - Databases

Data Mining Needs Databases

- Machine learning and statistics often make the following assumptions:
  - small amount of data (or sample)
  - data fits in main memory
  - CPU time is crucial
- The reality:
  - huge amounts data
  - data on secondary storage
  - data management (disk I/O) is crucial

Data Mining Techniques

- Classification (supervised learning)
  - Build and train classifiers (decision trees, neural nets, etc.)
- Clustering (unsupervised learning)
  - Partition the data into groups with similar characteristics.
- Sequence and stream analysis
- Association rule-mining

Association-Rule Mining

- Flagship of data mining with database flavor.
- Find correlations among data without building a complete predictive or descriptive model.
- Data-centric approach.
Market Basket Data

- Consider supermarket customers.
- At the checkout each customer has a basket of items.
- Find correlation among the contents of baskets.
- The model works for many domains:
  - Online/offline shopping
  - Web surfing
  - Text analysis

Association Rules

- Find rules of the form:
  - People who buy X tend to buy Y.

Mythical Association Rule

- Suppose we know that people buy bread and milk frequently. So what?
  - Stock them together.
  - Stock them apart.
  - Run sales on one and up the price of the other.
- Amazon’s recommendations are based on association rules.
  - Order size went up 20% in the first week after recommendations were introduced.

Schema of Market Basket Data

- Several models possible depending on the application.
- Simplest, most general schema:
  \[ \text{Baskets}(\text{basketID}, \text{item}) \]
- Applicable to many different scenarios, online and offline.

Market Basket Example

<table>
<thead>
<tr>
<th>basketID</th>
<th>item</th>
</tr>
</thead>
<tbody>
<tr>
<td>11111</td>
<td>beer</td>
</tr>
<tr>
<td>11111</td>
<td>chips</td>
</tr>
<tr>
<td>11111</td>
<td>salsa</td>
</tr>
<tr>
<td>22222</td>
<td>vodka</td>
</tr>
<tr>
<td>22222</td>
<td>caviar</td>
</tr>
</tbody>
</table>
Support and Confidence

- Formally, we associate two numbers with every rule:
  - support
  - confidence
- Example: Diapers → Beers
  - Support is the fraction of all baskets that contain both beer and diapers.
  - Confidence is the fraction of baskets which contain diapers that also contain beers.

Thresholds

- Find association rules with high support and high confidence.
- Typically, high support means > 0.1% and high confidence means > 50%.
- Thresholds depend on the application.

Main Challenge

- Too many item combinations:
  - 100s of thousands of items
  - millions of transactions
- Direct approach too slow:
  - 100 million baskets, 20 items/basket
  - 19 billion pairs, 100+ billion triples,…

Two-Phase Approach

- Phase 1: Find all itemsets with high support.
  - These itemsets are called frequent.
- Phase 2: Construct rules with high confidence.
  - The computational cost of phase 1 dominates the total cost.
  - Focus on finding frequent itemsets.

Find All Frequent Pairs

- Write query in SQL:

The A-Priori Technique

- Key observation: a pair of items is frequent only if each item is frequent.
  - If \{bread, cheese\} is frequent then \{bread\} and \{cheese\} must be frequent.
- Levelwise pruning:
  - Consider \{bread, milk, cheese\} only if \{bread, milk\}, \{bread, cheese\}, \{milk, cheese\} are frequent
A-Priori in SQL

```
INSERT INTO Baskets1(bid, item)
SELECT * FROM Baskets
WHERE item IN
(SELECT item
FROM Baskets
GROUP BY item
HAVING COUNT(*) >= s);
```

• Rewrite join using Basket1 instead of Basket.

Extending Association Rules

- Causality vs. association
- much trickier
- hidden variables outside the domain
- More detailed associations:
  - Find items that are bought together frequently, in a particular region, in a particular month.
  - Additional information is already available at the data warehouse.

Example Data Warehouse

Need for Data Warehousing

- Integrated, company-wide view of high-quality information.
- Separation of operational and analytical systems and data.

Operational vs. Analytical Data

Data Differences

Typical Time-Horizon: Days/Months
Detailed
Current

Typical Time-Horizon: Years
Summarized (and/or Detailed)
Values over time (Snapshots)

Technical Differences

Can be Updated
Read (and Append) Only
Control of Update: Major Issue
Control of Update: No Issue
Small Amounts used in a Process
Large Amounts used in a Process
Non-Redundant
Redundancy not an Issue
High frequency of Access
Low/Moderate frequency of Access

Purpose Differences

For "Clerical Community"
Supports Day-to-Day Operations
Application Oriented

For "Managerial Community"
Supports Managerial Needs
Subject Oriented

Application vs. Subject Oriented

<table>
<thead>
<tr>
<th>Application: Health Club Members-Visit Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEALTHCLUBMEMBERS</td>
</tr>
<tr>
<td>111 Joe A</td>
</tr>
<tr>
<td>222 Sue B</td>
</tr>
<tr>
<td>333 Pat A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DAILYVISITSFROMNONMEMBERS</th>
<th>VisitType</th>
<th>VisitDate</th>
</tr>
</thead>
<tbody>
<tr>
<td>11xx22 YP</td>
<td>01/01/2000</td>
<td></td>
</tr>
<tr>
<td>11xx23 NP</td>
<td>02/01/2000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MEMBERSHIPLEVELS</th>
<th>ID</th>
<th>Type</th>
<th>Fee</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gold</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Basic</td>
<td>$50</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VISITLEVELS</th>
<th>ID</th>
<th>Type</th>
<th>Fee</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>With Pool Usage</td>
<td>$15</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Without Pool Usage</td>
<td>$10</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subject: Health Club Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>REVENUE</td>
</tr>
<tr>
<td>2235</td>
</tr>
<tr>
<td>2236</td>
</tr>
<tr>
<td>2237</td>
</tr>
<tr>
<td>2238</td>
</tr>
<tr>
<td>2239</td>
</tr>
<tr>
<td>2240</td>
</tr>
</tbody>
</table>

```
Application vs. Subject Oriented

**Application:**
Health Club Members-Visit Database

<table>
<thead>
<tr>
<th>HEALTHCLUBMEMBERS</th>
<th>MembId</th>
<th>Name</th>
<th>MembLevel</th>
<th>DatePayed</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>Joe</td>
<td>A</td>
<td>01/01/2000</td>
<td></td>
</tr>
<tr>
<td>222</td>
<td>Sue</td>
<td>B</td>
<td>01/01/2000</td>
<td></td>
</tr>
<tr>
<td>333</td>
<td>Pat</td>
<td>A</td>
<td>01/01/2000</td>
<td></td>
</tr>
</tbody>
</table>

**DAILYVISITSFROMNONMEMBERS**

<table>
<thead>
<tr>
<th>Trid</th>
<th>VisitType</th>
<th>VisitDate</th>
</tr>
</thead>
<tbody>
<tr>
<td>11122</td>
<td>YP</td>
<td>01/01/2000</td>
</tr>
<tr>
<td>11123</td>
<td>NP</td>
<td>02/01/2000</td>
</tr>
<tr>
<td>11124</td>
<td>YP</td>
<td>02/01/2000</td>
</tr>
</tbody>
</table>

**MEMBERSHIPS LEVELS**

<table>
<thead>
<tr>
<th>ID</th>
<th>Type</th>
<th>Fee</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Gold</td>
<td>$100</td>
</tr>
<tr>
<td>B</td>
<td>Basic</td>
<td>$50</td>
</tr>
</tbody>
</table>

**VISIT LEVELS**

<table>
<thead>
<tr>
<th>ID</th>
<th>Type</th>
<th>Fee</th>
</tr>
</thead>
<tbody>
<tr>
<td>YP</td>
<td>With Pool Usage</td>
<td>$15</td>
</tr>
<tr>
<td>NP</td>
<td>Without Pool Usage</td>
<td>$10</td>
</tr>
</tbody>
</table>

**Subject:**
Health Club Revenue

<table>
<thead>
<tr>
<th>REVENUE</th>
<th>Rid</th>
<th>Date</th>
<th>GeneratedBy</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>7235</td>
<td>01/01/2000</td>
<td>NonMember</td>
<td>$15</td>
<td></td>
</tr>
<tr>
<td>7236</td>
<td>01/01/2000</td>
<td>Member</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>7237</td>
<td>01/01/2000</td>
<td>Member</td>
<td>$50</td>
<td></td>
</tr>
<tr>
<td>7238</td>
<td>01/01/2000</td>
<td>Member</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>7239</td>
<td>02/01/2000</td>
<td>NonMember</td>
<td>$10</td>
<td></td>
</tr>
<tr>
<td>7240</td>
<td>02/01/2000</td>
<td>NonMember</td>
<td>$15</td>
<td></td>
</tr>
</tbody>
</table>

Standard ARM Question:
What products are frequently bought together?

Analyst may want to know:
What products are frequently bought together in a particular region and in a particular month?

New Challenges
- Interactive mining
- Collaborative/distributed mining
- Peer to peer systems
- Beyond relational data:
  - Text
  - XML
  - Audio
  - Video