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

A Lesson in Marketing

- 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:
  - Baskets(basketID, 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 $\rightarrow$ 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

```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

![Diagram of a data warehouse with tables and relationships]

Need for Data Warehousing

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

Operational vs. Analytical Data

<table>
<thead>
<tr>
<th>Data Differences</th>
<th>Technical Differences</th>
<th>Purpose Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical Time-Horizon: Days/Months</td>
<td>Typical Time-Horizon: Years</td>
<td>For &quot;Clerical Community&quot;</td>
</tr>
<tr>
<td>Detailed</td>
<td>Summarized (and/or Detailed)</td>
<td>Supports Day-to-Day Operations</td>
</tr>
<tr>
<td>Current</td>
<td>Values over time (Snapshots)</td>
<td>Application Oriented</td>
</tr>
<tr>
<td>Can be Updated</td>
<td>Read (and Append) Only</td>
<td></td>
</tr>
<tr>
<td>Control of Update: Major Issue</td>
<td>Control of Update: No Issue</td>
<td>For &quot;Managerial Community&quot;</td>
</tr>
<tr>
<td>Small Amounts used in a Process</td>
<td>Large Amounts used in a Process</td>
<td>Supports Managerial Needs</td>
</tr>
<tr>
<td>Non-Redundant</td>
<td>Redundancy not an Issue</td>
<td>Subject Oriented</td>
</tr>
<tr>
<td>High frequency of Access</td>
<td>Low/Most frequency of Access</td>
<td></td>
</tr>
</tbody>
</table>

Application vs. Subject Oriented

**Application:**

- Health Club Members-Visit Database
  - HEALTHCLUBMEMBERS
    - MembId: Name: MembLevel: DatePayed
      - 111: Joe: A: 01/01/2000
      - 222: Sue: B: 01/01/2000
      - 333: Pat: A: 01/01/2000
  - DAILYVISITSFROMNONMEMBERS
    - Trid: VisitType: VisitDate
      - 11xx22: YP: 01/01/2000
      - 11xx23: NP: 02/01/2000

**Subject:**

- Health Club Revenue
  - REVENUE: GenerateBy: Amount
    - 7235: 01/01/2000: NonMember: $15
    - 7236: 01/01/2000: Member: $100
    - 7237: 02/01/2000: Member: $50
    - 7238: 02/01/2000: NonMember: $10
Application vs. Subject Oriented

**Application:**
- **Health Club Members-Visit Database**
  - **HEALTHCLUBMEMBERS**
    - MembId | Name | MembLevel | DatePayed
    - 111 | Joe | A | 01/01/2000
    - 222 | Sue | B | 01/01/2000
    - 333 | Pat | A | 01/01/2000
    ...
  - **DAILYVISITSFROMNONMEMBERS**
    - Trid | VisitType | VisitDate
    - 11xx22 | YP | 01/01/2000
    - 11xx23 | NP | 02/01/2000
    - 11xx24 | YP | 02/01/2000
    ...
  - **MEMBRSHPLEVELS**
    - ID | Type | Fee
    - A | Gold | $100
    - B | Basic | $50
  - **VISITLEVELS**
    - ID | Type | Fee
    - YP | With Pool Usage | $15
    - NP | Without Pool Usage | $10
- **Subject:**
  - **Health Club Revenue**
    - Rid | Date | GeneratedBy | Amount
    - 7235 | 01/01/2000 | NonMember | $15
    - 7236 | 01/01/2000 | Member | $100
    - 7237 | 01/01/2000 | Member | $50
    - 7238 | 01/01/2000 | Member | $100
    - 7239 | 02/01/2000 | NonMember | $10
    - 7240 | 02/01/2000 | NonMember | $15

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