Have You Ever ...

- Wondered how products are placed in supermarket isles?
- 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!!!
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: $\text{Baskets}(\text{basketID}, \text{item})$
- 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: Beer → Diapers
  - Support is the fraction of all baskets that contain both beer and diapers.
  - Confidence is the fraction of baskets which contain beer that also contain diapers.

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 computing 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, at a particular city, at a particular day of the week.

New Challenges

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