Beyond Relational Queries

CS 2541: Database Systems
Information Retrieval
Data Mining

Data Analytics anyone?
- What is data analytics (also analytics)?
- Examples:
  - Sports analytics
    - If pitcher A is pitching then send hitter B to hit
    - If X is playing point guard, then send Y to guard X
  - Customer analytics — (amazon) if customer buys X then suggest buying Y
  - Financial analysis
    - If warm winter then what happens to oil prices

DM vs ML?
- Conclusion vs Prediction
  - Data mining vs machine learning
  - This is an extremely naïve (and not totally correct) explanation
- DM is an info source that ML can use
- DM pulls existing patterns in data
  - ML could (?) predict future patterns
- ML can make automated co-relations and learn from them
  - The algo changes as it learns from data

Next....Analyzing data in DBMS
- Relational DBMS (RDBMS)
  - Good for unstructured data
- NoSQL DBMS
  - Scales well on multiprocessors
- Big Data?
Predictive Modelling

- A “black box” that makes predictions about the future based on information from the past and present

<table>
<thead>
<tr>
<th>Model</th>
<th>How much will customer spend on next catalog order?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>balance</td>
</tr>
<tr>
<td>income</td>
<td></td>
</tr>
</tbody>
</table>

Preprocessing and Mining

- Original Data
- Target Data: Data Warehouse
- Preprocessed Data
- Patterns
- Knowledge
- Interpreation

OLAP: OnLine Analytical Processing -- Early Analytics?

- Way of presenting data to facilitate understanding any patterns inside it
  - Provide for drilling down into the data starting from summarized views
- OLAP lets you look at data and manipulate interactively
- OLAP allows users to “slice and dice” data
  - Allows user to drill-down into detail data
- Anyone heard of Tableau?
  - Dashboard analysis?

Important: Legal and Ethical Issues

- Privacy concerns…hello Facebook!
  - Becoming more important
  - Will impact the way that data can be used and analyzed
  - Ownership issues
- Often data included in the data warehouse cannot legally be used in decision making process
  - Race, Gender, Age
  - Bias in machine learning…..
  - Google translate example
- Data contamination will become critical
Relational vs Multidimensional

<table>
<thead>
<tr>
<th>Product</th>
<th>Region</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuts</td>
<td>East</td>
<td>50</td>
</tr>
<tr>
<td>Nuts</td>
<td>West</td>
<td>60</td>
</tr>
<tr>
<td>Nuts</td>
<td>Central</td>
<td>100</td>
</tr>
<tr>
<td>Screws</td>
<td>East</td>
<td>40</td>
</tr>
<tr>
<td>Screws</td>
<td>West</td>
<td>70</td>
</tr>
<tr>
<td>Screws</td>
<td>Central</td>
<td>80</td>
</tr>
<tr>
<td>Bolts</td>
<td>East</td>
<td>90</td>
</tr>
<tr>
<td>Bolts</td>
<td>West</td>
<td>120</td>
</tr>
<tr>
<td>Bolts</td>
<td>Central</td>
<td>140</td>
</tr>
<tr>
<td>Washers</td>
<td>East</td>
<td>20</td>
</tr>
<tr>
<td>Washers</td>
<td>West</td>
<td>15</td>
</tr>
<tr>
<td>Washers</td>
<td>Central</td>
<td>30</td>
</tr>
</tbody>
</table>

Consolidated Data

<table>
<thead>
<tr>
<th>Region</th>
<th>East</th>
<th>West</th>
<th>Central</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuts</td>
<td>50</td>
<td>90</td>
<td>100</td>
<td>240</td>
</tr>
<tr>
<td>Screws</td>
<td>40</td>
<td>70</td>
<td>100</td>
<td>210</td>
</tr>
<tr>
<td>Bolts</td>
<td>90</td>
<td>120</td>
<td>140</td>
<td>350</td>
</tr>
<tr>
<td>Washers</td>
<td>20</td>
<td>10</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>240</td>
<td>310</td>
<td>750</td>
</tr>
</tbody>
</table>

Can be generated by groupby Queries
Two dimensional view of data

OLAP Views

OLAP – how is it useful in decision making

Manager: “total sales of 810 are not meeting target. Which region is underperforming?”

<table>
<thead>
<tr>
<th>Region</th>
<th>East</th>
<th>West</th>
<th>Central</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuts</td>
<td>50</td>
<td>90</td>
<td>100</td>
<td>240</td>
</tr>
<tr>
<td>Screws</td>
<td>40</td>
<td>70</td>
<td>100</td>
<td>210</td>
</tr>
<tr>
<td>Bolts</td>
<td>90</td>
<td>120</td>
<td>140</td>
<td>350</td>
</tr>
<tr>
<td>Washers</td>
<td>20</td>
<td>10</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>240</td>
<td>310</td>
<td>750</td>
</tr>
</tbody>
</table>

OLAP – how is it useful in decision making

Manager: “total sales of 810 are not meeting target. What are the problem areas/why?”

<table>
<thead>
<tr>
<th>Region</th>
<th>East</th>
<th>West</th>
<th>Central</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuts</td>
<td>50</td>
<td>90</td>
<td>100</td>
<td>240</td>
</tr>
<tr>
<td>Screws</td>
<td>40</td>
<td>70</td>
<td>100</td>
<td>210</td>
</tr>
<tr>
<td>Bolts</td>
<td>90</td>
<td>120</td>
<td>140</td>
<td>350</td>
</tr>
<tr>
<td>Washers</td>
<td>20</td>
<td>10</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>240</td>
<td>310</td>
<td>750</td>
</tr>
</tbody>
</table>

Answer: “Drill” down by region and/or category
East region sales are low
OLAP – how is it useful in decision making

Manager: “total sales of 810 are not meeting target. Which product is underperforming?”

<table>
<thead>
<tr>
<th></th>
<th>East</th>
<th>West</th>
<th>Central</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nails</td>
<td>50</td>
<td>100</td>
<td>100</td>
<td>250</td>
</tr>
<tr>
<td>Screws</td>
<td>45</td>
<td>70</td>
<td>50</td>
<td>165</td>
</tr>
<tr>
<td>Bottle</td>
<td>60</td>
<td>120</td>
<td>140</td>
<td>320</td>
</tr>
<tr>
<td>Washers</td>
<td>20</td>
<td>30</td>
<td>30</td>
<td>80</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>280</td>
<td>260</td>
<td>740</td>
</tr>
</tbody>
</table>

Answer: very few washers are being sold

DM techniques: Prediction and Classification

- Prediction and Classification
- Directed
- Algorithms: Decision trees, Neural networks, memory based reasoning, etc.
- Examples:
  - Forecast: How many units will be sold on a given day?
  - What will be the stock price on a given day?
  - Will a customer buy the product or not?
  - Detect sequences:
    - balance increase -> missed payment -> default

DM techniques: Affinity Grouping

- Affinity grouping
  - Undirected
  - Which products go together naturally?
  - Market basket analysis
- Examples:
  - Which products peak in demand simultaneously?

DM techniques: Clustering

- Clustering task
  - Undirected
  - Segmenting into similar clusters
- Different from classification
- Examples
  - Customers with similar buying profiles
  - Products with similar demand patterns
  - Classification and segmentation: yes/no
Data Mining Algorithms

- Four algorithms commonly cited
  - Association Rule (used in over 90% of the cases!)
  - Nearest Neighbor
    - quick and easy but models get large
  - Decision Tree
  - Neural Network
    - difficult to interpret and large time

Decision Trees

- Series of if/then rules
  - easy to understand, complexity in implementation

Cluster Detection

- Undirected data mining
- Finds records that are similar to each other (clusters)
- Clusters are found using geometric methods, statistical methods, and neural networks
- Good way to start any analysis

Market Basket Analysis: Association Rule Mining

- Form of clustering used for finding items that occur together (in a transaction or market basket)
- Likelihood of different products being purchased together as rules
- Planning store layouts, limiting specials to one of the products in a set,...
**Association Rules: Support and Confidence Parameters**

- looking for a rule that says: If A then B
- **Support** is defined as the ratio of number of transactions that include both A and B to total number of transactions
- How useful are A and B?
- **Confidence** is defined by the ratio of the number of transactions that include both A and B to the number of transactions that include A.
- If A then B with probability \( p \)
- How do you specify 'significant' support and confidence?

---

**A very simple (and bad performance) algorithm**

- Build a co-occurrence matrix from the data
  - Transactions in which two items occurred together
  - From this table, filter out entries that meet the support and confidence ratios
- If we have N transactions,
  - To meet support of \( x \), the corresponding entry in the matrix must be \( \geq x^N \)
  - To meet the confidence cutoff,
    - confidence = (Entry for A and B) / (Total number of A)
- In reality, this is a very inefficient algorithm
- In practice: use concepts of “frequent itemsets” to build an association rule mining algorithm

---

**Transaction data**

<table>
<thead>
<tr>
<th>StudentID</th>
<th>Courses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CS2461, CS2113</td>
</tr>
<tr>
<td>2</td>
<td>CS2461, CS2312, PHIL51,</td>
</tr>
<tr>
<td></td>
<td>Math33</td>
</tr>
<tr>
<td>3</td>
<td>CS2461, Math33</td>
</tr>
<tr>
<td>4</td>
<td>CS2312, PHIL51, CS2113</td>
</tr>
<tr>
<td>5</td>
<td>CS2312, CS2113</td>
</tr>
</tbody>
</table>

**Association Rule example**

- Using the data to create a co-occurrence matrix \( C \)
  - Initially \( C_{ij} = 0 \) for all \( ij \)
  - For each transaction \( T_k \) (the transaction consists of a list of courses/items that appear in this transaction)
    - If course (item) \( i \) appears in the transaction, increment entry \( C[i,j] \) (the diagonal entries)
    - If any pair of courses (items) \( i \) and \( j \) both appear in the transaction then increment \( C[i,j] \) (and \( C[j,i] \))
  - \( N \) is the total number of transactions
  - If there are \( M \) different items then matrix \( C \) has \( M \) by \( M \) entries and each entry is no greater than \( N \)
**Example**

<table>
<thead>
<tr>
<th>StudentID</th>
<th>Courses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CS2461, CS2113</td>
</tr>
<tr>
<td>2</td>
<td>CS2461, CS2312, Phil51,</td>
</tr>
<tr>
<td></td>
<td>Math33</td>
</tr>
<tr>
<td>3</td>
<td>CS2461, Math33</td>
</tr>
<tr>
<td>4</td>
<td>CS2312, Phil51, CS2113</td>
</tr>
<tr>
<td>5</td>
<td>CS2312, CS2113</td>
</tr>
</tbody>
</table>

Total number of transactions = \( N = 5 \)

Since both CS2461 and CS2113 appear in \( T_1 \), the entry for \( C[cs2461, cs2113] \) is incremented by 1.

Index the courses as:
1 = cs2461, 2 = cs2113, 3 = cs2312, 4 = phil51, 5 = math33

\( C[2, 3] \) is number of times cs2113 and cs2312 appear together in a transaction.

---

**Co-occurrence Matrix for Example**

<table>
<thead>
<tr>
<th></th>
<th>cs2461</th>
<th>cs2113</th>
<th>cs2312</th>
<th>Phil51</th>
<th>Math33</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs2461</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>cs2113</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>cs2312</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Phil51</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Math33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

---

**Association Rule example**

- Let relevant Support = 25% and Confidence = 50%:
  - CS2312 and CS2113 appear in 2/5 = 40%
  - CS2312 appears in 3 students
  - If CS2312 then CS2113 has confidence of 2/3 = 67%
  - Thus, "If student registers for 2312 then student registers for 2113" satisfies 25% support & 50% confidence
- Conclusion drawn by mining system:
  - Students who register for 2312 also register for 2113

---

**Example – finding all rules/associations**

- Since support \( x = 25\% \) and number of trans \( N = 5 \) we only consider pairs \( (i, j) \) where
  \[
  C[i, j] \geq x \times N = 0.25 \times 5 = 1.25
  \]
- Next, out of the pairs that satisfy the support above, since confidence \( y = 50\% \), we only consider pairs \( C[i, j] \) where
  \[
  C[i, j] / C[i, i] \geq y = 0.50
  \]
- The above conditions will lead to
  - \((cs2312, cs2113)\)
  - \((cs2312, phil51)\)
In-Class example

Below is transaction data from a supermarket transaction ID, <list of items purchased>
We set support=25% and confidence = 50%
What are the rules identified?

Data

- Support=25%  Confidence=50%
- (100, Milk, Soda)
- (200, Milk, Beer, Diapers)
- (300, Milk, Cleaner)
- (400, Beer, Diapers, Soda)
- (500, Beer, Soda)

Association Rule example

- Using the sample data create a co-occurrence table
- Let relevant Support = 25% and Confidence= 50%:
  - Beer and Diapers appear in 2/5= 40%
  - If beer then diapers has confidence of 2/3=67%
  - Thus, “if customer buys beer then customer buys diapers” satisfies 25% support & 50% confidence
- Conclusion drawn by mining system:
  - Customers who buy beer also buy diapers

Applying the Results

- Is the relationship useful?
  - Beer and Diapers may not be of use
  - Amazon’s clustering, Sports analytics, Finance, …
- Who defines “usefullness”
  - only as good as rules specified by humans/marketing workforce
**Data Mining: Summary**

- “Using the new media of the one-to-one future, you will be able to communicate directly with customers individually...” - Don Peppers & Martha Rogers, 2001 (?) (One-to-One Future)
- “What are you afraid of?.....Even if you’re not afraid of these things, the beauty is, with proper marketing, we can make you afraid”-- Michael Saylor, CEO Microstrategy, 2000.