Collection of data is outpacing analytical power

- Discover hidden (and useful) information to help us make decisions
Association Rule Mining
(Agrawal et al. 1993, 1994)

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diaper, Beer, Egg</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diaper, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diaper, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diaper, Coke</td>
</tr>
</tbody>
</table>

Association Rule: \{Milk, Diaper\} \rightarrow \{Beer\}

- **Support**: Probability(Milk, Diaper and Beer in T) = 2/5
- **Confidence**: Probability(Beer in T | Milk and Diaper in T) = 2/3
Apriori Principle

- **Apriori**: support has the anti-monotone property
Research Motivation

A gap between
- Association Pattern Mining
- Statistical Correlation Analysis

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
<th>Pattern</th>
<th>Support</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Milk, Ring, Necklace</td>
<td>{Milk} → {Coffee}</td>
<td>0.60</td>
<td>-0.25</td>
</tr>
<tr>
<td>2</td>
<td>Milk, Tea</td>
<td>{Ring} → {Necklace}</td>
<td>0.10</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Coffee, Tea</td>
<td>{Tea} → {Coffee}</td>
<td>0.67</td>
<td>-0.22</td>
</tr>
<tr>
<td>4</td>
<td>Milk, Coffee</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Milk, Coffee</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Milk, Coffee</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Milk, Coffee</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Milk, Coffee</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Coffee, Tea</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Coffee</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Application Scenario

- Amazon.com product suggestions

- A computation challenge!
  - For a database of $10^6$ items, it raises to the scale of $10^{12}$ pairs
  - Several million transactions make things even worse
  - Recommendations should be updated dynamically
Naïve Approach: r-TAPER

- For each update, re-compute all correlations
- Use an efficient static algorithm to boost each query
  - TAPER for each one-time query (Xiong et al. 2006)
  - r-TAPER: repeated TAPER for dynamic data

![Diagram showing data over time with ASP updates]
Naïve Approach: SAVE-ALL

\[
\phi_{\{a,b\}} = \frac{NN_{ab} - NaNb}{\sqrt{Na(N - Na)Nb(N - Nb)}}
\]

- SAVE-ALL: to save all pair-wise frequencies
  - Need \(O(n^2)\) space
  - What if \(n=10^6\)? Too large to fit in the memory.

- Can we save selected pairs to meet our needs?
The Checkpoint Principle

- Checkpoints can help us derive a computation buffer: a collection of candidate item pairs.
CHECK-POINT Algorithm

- At a checkpoint: establish the buffer
  - For each possible item pair \( \{a,b\} \)
    - Compute \( \text{upper}(\phi_{ab}) \) based on next checkpoint
    - If above \( \theta \), append \( \{a,b\} \) to the candidate list

- Each update: search within buffer
  - For each candidate item pair \( \{a, b\} \)
    - Update \( Nab \) and calculate \( \phi_{ab} \)
    - If above \( \theta \), output \( \{a,b\} \) as a strong pair

- Establishing a checkpoints for multiple updates will save computation
A Loose Upper Bound

□ Let \( v(x) = \sqrt{x(N - x)} \) and \( W_{ab} = NN_{ab} - N_a N_b \)

\[
\phi_{ab} = \frac{NN_{ab} - N_a N_b}{\sqrt{N_a(N - N_a)N_b(N - N_b)}} = \frac{W_{ab}}{v(N_a)v(N_b)}
\]

□ A loose upper bound:

\[
\phi_{ab} \leq \frac{\max W_{ab}}{\min v(N_a) \min v(N_b)}
\]

where \( \min v(N_x) = \min\{v(n_x), v(n_x + \Delta)\}, \forall x \in \{a, b\} \)

\[
\max W_{ab} = \begin{cases}
  w(0), & \text{if } c_0 \leq 0; \\
  w(c_0), & \text{if } 0 < c_0 \leq \Delta; \\
  w(\Delta), & \text{if } c_0 > \Delta;
  \end{cases}
\]

\[
  c_0 = \frac{N - n_a - n_b}{2}
\]

\[
  w(t) = N(n_{ab} + t) - (n_a + t)(n_b + t)
\]

The Tight Upper Bound

- **Case I**: If $n_{ab} = n_a = n_b$, then
  \[
  \text{upper}(\phi_{ab}) = 1
  \]

- **Case II**: If $n_{ab} < n_a = n_b$, then
  \[
  \text{upper}(\phi_{ab}) = \begin{cases} 
  f(0), & \text{if } n_a + n_b \leq n; \\
  f(\Delta), & \text{otherwise}.
  \end{cases}
  \]

- **Case III**: If $n_a \neq n_b$, then
  \[
  \text{upper}(\phi_{ab}) = \begin{cases} 
  f(\Delta), & \text{if } c_0 > \Delta; \\
  f(0), & \text{if } c_0 < 0; \\
  f(\lfloor c_0 \rfloor), & \text{otherwise}.
  \end{cases}
  \]

Local Correlation Patterns
(Zhou et al., ICDM 2009)

- Association patterns may behave differently at the local level from the global level

<table>
<thead>
<tr>
<th>Global Observation</th>
<th>Local Observation</th>
<th>Pitfalls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significant</td>
<td>Insignificant</td>
<td>False Positive</td>
</tr>
<tr>
<td>Insignificant</td>
<td>Significant</td>
<td>False Negative</td>
</tr>
</tbody>
</table>

- Simpson’s Paradox
  - The (global) pattern differs from each local segment
  - Direction of the correlation might be reversed
## Types of Confounders

<table>
<thead>
<tr>
<th>Type</th>
<th>Correlation</th>
<th>Partial Correlation</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutralizing Confounder</td>
<td>$</td>
<td>\phi_{AB}</td>
<td>\geq \theta$</td>
</tr>
<tr>
<td>Reversing Confounder</td>
<td>Positively Correlated ($\phi_{AB} \geq \theta$)</td>
<td>Negatively Correlated ($\phi_{AB</td>
<td>C} \leq -\theta$)</td>
</tr>
<tr>
<td></td>
<td>Negatively Correlated ($\phi_{AB} \leq -\theta$)</td>
<td>Positively Correlated ($\phi_{AB</td>
<td>C} \geq \theta$)</td>
</tr>
</tbody>
</table>

Problem: find all item triplets in the form of \{A,B|C\}, such that item C is a reversing confounder of item pair \{A,B\}.
Property of Confounders

- Given threshold $\theta (0 < \theta < 1)$, $C$ may be a reversing confounder of $\{A,B\}$ only if

$$\begin{align*}
\phi_{AC} &\geq \phi_{AB} \\
\phi_{BC} &\geq \phi_{AB}
\end{align*}$$

or

$$\begin{align*}
\phi_{AC} &\leq -\phi_{AB} \\
\phi_{BC} &\leq -\phi_{AB}
\end{align*}$$

- $C$ may be a reversing confounder of $\{A,B\}$ only if $A$, $B$, and $C$ are pairwise correlated (either positively or negatively).

- For any three items $A$, $B$, and $C$ that are pairwise correlated, none of them will be a reversing confounder if $\phi_{AB}\phi_{AC}\phi_{BC} \leq 0$. 

Selected Publications

- **Journal Papers**

- **Conference Papers**
Thank You!

Please send your offline questions to wzhou4@utk.edu