Real World Performance Of Association Rule Algorithm

Zijian Zheng, Ron Kohavi, and Llew Mason
{zijian, ronnyk, lmason}@bluenmartini.com

Blue Martini Software
San Mateo, California

August 11, 2001
Very narrow min-sup range of interest

Super exponential growth in number of rules on real world data

Impossible

Super exponential growth in # Rules

>1,000,000,000 rules

Fast, but incorrect results
Performance improvements on artificial data did not generalize to real world data

Improvement over Apriori:

<table>
<thead>
<tr>
<th></th>
<th>Closet</th>
<th>Charm</th>
<th>FP-growth</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial</td>
<td>1.9 x</td>
<td>2.8 x</td>
<td>12.1 x</td>
<td>3.1 x</td>
</tr>
<tr>
<td>Real-POS</td>
<td>1.2 x</td>
<td>1.1 x</td>
<td>0.8 x</td>
<td>1.0 x</td>
</tr>
</tbody>
</table>

Are algorithms overfitting the artificial data?
Association rule discovery:

- Active research area
- Good application potential

Many promising new algorithms

Each new algorithm has significant performance improvements

- mainly based on results on IBM Almaden artificial datasets

How do these algorithms perform on real-world data?
## Datasets

<table>
<thead>
<tr>
<th></th>
<th>Transactions</th>
<th>Distinct Items</th>
<th>Maximum Transaction Size</th>
<th>Average Transaction Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM-Artificial</td>
<td>100,000</td>
<td>870</td>
<td>29</td>
<td>10.1</td>
</tr>
<tr>
<td>BMS-POS</td>
<td>515,597</td>
<td>1,657</td>
<td>164</td>
<td>6.5</td>
</tr>
<tr>
<td>BMS-WebView-1</td>
<td>59,602</td>
<td>497</td>
<td>267</td>
<td>2.5</td>
</tr>
<tr>
<td>BMS-WebView-2</td>
<td>77,512</td>
<td>3,340</td>
<td>161</td>
<td>5.0</td>
</tr>
</tbody>
</table>
IBM artificial dataset is very different from the real-world datasets
Datasets (Cont’d)

- IBM-Artificial
- BMS-POS
Ratio of #Freq itemsets over #Closed freq itemsets

BMS-POS

BMS-WebView-2
Experimental Setup

- Dual 550MHz Pentium III with 1 GB of memory
- Windows NT 4.0
- Measures: time (seconds) for generating frequent itemsets/association rules
- Minimum support: 1.00%, 0.80%, 0.60%, 0.40%, 0.20%, 0.10%, 0.08%, 0.06%, 0.04%, 0.02%, and 0.01%
- Minimum confidence: 0%
Comparison Algorithm

- **Apriori**: C. Borgelt’s implementation
- **Charm**: M. Zaki
- **FP-growth**: J. Han’s research group
- **Closet**: J. Han’s research group
- **MagnumOpus**: G. Webb
Experimental Results

Rankings (left is better) of the algorithms for generating frequent itemsets on the four datasets with high minimum supports and low minimum supports (Ap: Apriori, FP: FP-growth, Ch: Charm, Cl: Closet):

<table>
<thead>
<tr>
<th>Ranking</th>
<th>High Min-Support</th>
<th>Low Min-Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM-Artificial</td>
<td>Ap &gt; FP &gt; Ch &gt; Cl</td>
<td>FP &gt; Ch &gt; Cl &gt; Ap</td>
</tr>
<tr>
<td>BMS-POS</td>
<td>Ap &gt; Cl &gt; FP &gt; Ch</td>
<td>Ch &gt; FP &gt; Ap &gt; Cl</td>
</tr>
<tr>
<td>BMS-WebView-1</td>
<td>Ap &gt; FP &gt; Cl &gt; Ch</td>
<td>Ch &gt; FP &gt; Ap &gt; Cl</td>
</tr>
<tr>
<td>BMS-WebView-2</td>
<td>Ap &gt; FP &gt; Ch &gt; Cl</td>
<td>Ch &gt; FP &gt; Ap &gt; Cl</td>
</tr>
</tbody>
</table>
IBM-Artificial (frequent itemsets)
IBM-Artificial (frequent itemsets)

TIME (Seconds)

Minimum Support (%)

- Closet
- Charm
- FP-growth
- Apriori
BMS-POS (frequent itemsets)

TIME (Seconds)

Minimum Support (%)

Charm
FP-growth
Apriori

© Copyright 2000-2001, Blue Martini Software, San Mateo California, USA
BMS-POS (frequent itemsets)
BMS-WebView-1 (frequent itemsets)

TIME (Seconds)

0.00 0.02 0.04 0.06 0.08 0.10

Minimum Support (%)
BMS-WebView-1 (frequent itemsets)

TIME (Seconds)

Minimum Support (%)

- Closet
- Charm
- FP-growth
- Apriori
BMS-WebView-2 (frequent itemsets)

- Charm
- FP-growth
- Apriori
- Closet

Minimum Support (%) vs. Time (Seconds) graph.
BMS-WebView-2 (frequent itemsets)

TIME (Seconds)

Minimum Support (%)

- Closet
- Charm
- FP-growth
- Apriori
On some real-world datasets, when the minimum support is small, the number of frequent itemsets increases super-exponentially, thus no algorithm can handle it.

E.g. BMS-WebView-1:

<table>
<thead>
<tr>
<th>Minimum support(%)</th>
<th>Frequent itemsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.06</td>
<td>461,521</td>
</tr>
<tr>
<td>0.04</td>
<td>* 6.82 x 10^{10}</td>
</tr>
<tr>
<td>0.02</td>
<td>* 1.08 x 10^{26}</td>
</tr>
<tr>
<td>0.01</td>
<td>* 1.78 x 10^{45}</td>
</tr>
</tbody>
</table>

*: estimated
IBM-Artificial (association rules)
IBM-Artificial (association rules)

The graph shows the relationship between **Minimum Support (%)** and **TIME (Seconds)** for different algorithms:

- **Apriori**
- **MO**
- **MO-1000**

As the **Minimum Support (%)** increases, the **TIME (Seconds)** decreases for all algorithms, indicating improved performance with higher support levels.
BMS-POS (association rules)
BMS-WebView-1 (association rules)

![Graph showing time in seconds vs minimum support percentage for different algorithms: Apriori, MO, and MO-1000.](image)

- **TIME (Seconds)**
- **Minimum Support (%)**

- **Curves**:
  - **Apriori**
  - **MO**
  - **MO-1000**
BMS-WebView-2 (association rules)

![Graph showing the relationship between minimum support and time for different methods: Apriori, MO, and MO-1000. The x-axis represents minimum support (%), and the y-axis represents time (seconds). The graph illustrates that as minimum support increases, the time decreases for all methods, with MO-1000 generally taking the least time.]
MO could be a solution when
- the number of association rules is very large
- only the top-N rules are needed (based on some criteria such as lift or confidence)
Contributions

- First objective evaluation and comparison of association rule algorithms on real-world e-commerce and retail datasets.
- Donated one e-commerce datasets for use in the research community.
Artificial datasets have very different characteristics from the real-world datasets. On real world data:

- Very narrow min-sup range of interest
- Super exponential growth in number of rules

Performance improvements on artificial data did not generalize to real world data.

Optimizing algorithms for these artificial datasets can mislead research effort.