Interactive Data Exploration

Drawbacks found

Ali El Adi | Bruno Rubio | Deepak Barua | Hussain Syed

Databases and Information Retrieval Integration Project
Recap – Smart-Drill

- Goal: To discover and summarize interesting group of tuples using smart drill down operator. (keeping in mind the fundamental concepts of drill down).
- Group of tuples can be described as a rule.
  - e.g (a, b, *, 1000)
- A simple explanation to the above mention rule will be → thousand tuples with the a in first col, b in second col, anything in third, total count 1000 for tuples.
# Recap – Smart-Drill

## TABLE I: Initial Summary

<table>
<thead>
<tr>
<th>Store</th>
<th>Product</th>
<th>Region</th>
<th>Count</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>*</td>
<td>*</td>
<td>6000</td>
<td>0</td>
</tr>
</tbody>
</table>

## TABLE II: Result After First Smart Drill Down

<table>
<thead>
<tr>
<th>Store</th>
<th>Product</th>
<th>Region</th>
<th>Count</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>*</td>
<td>*</td>
<td>6000</td>
<td>0</td>
</tr>
<tr>
<td>▶ Target</td>
<td>bicycles</td>
<td>*</td>
<td>200</td>
<td>2</td>
</tr>
<tr>
<td>▶ *</td>
<td>comforters</td>
<td>MA-3</td>
<td>600</td>
<td>2</td>
</tr>
<tr>
<td>▶ Walmart</td>
<td>*</td>
<td>*</td>
<td>1000</td>
<td>1</td>
</tr>
</tbody>
</table>
Problems with Smart-Drill

- It would not be fair to call these drawbacks, instead we will refer to as problems or things that we didn’t like this much with smart-drill.
- Finding the best marginal rule can be difficult in cases of big data sets.
Problems with Smart-Drill

- A brute way to do this can be to enumerate all possible rules and to find marginal value for each rule in single pass of data.
- To avoid this we have to rely on other techniques such as a-priori algorithm for frequent item set mining.
Problems with Smart-Drill

• So it looks like finding interesting rules is also largely dependent on size of the dataset and in case of very large datasets, sample data is used to get the idea of whole data which may lead to less accuracy than a user desired for initially.
Problems with Smart-Drill

- A system component called sample handler is responsible for creating multiple samples of different parts of the table in memory.
- MinSS (minimum Sample Size)
- A higher value of minSS increasing accuracy but also increases computation cost.
Problems with Smart-Drill

![Graph showing time in milliseconds to expand empty rule against minSS Parameter value for different weighting schemes.]

- Marketing Size weighting
- Marketing Bits weighting
- Census Bits weighting
- Census Bits weighting
### Further Practical Problems

#### Row Oriented (RDBMS Model)

<table>
<thead>
<tr>
<th>id</th>
<th>Name</th>
<th>Age</th>
<th>Interests</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ricky</td>
<td></td>
<td>Soccer, Movies, Baseball</td>
</tr>
<tr>
<td>2</td>
<td>Ankur</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Sam</td>
<td>25</td>
<td>Music</td>
</tr>
</tbody>
</table>

#### Column Oriented (Multi-value sorted map)

<table>
<thead>
<tr>
<th>id</th>
<th>Name</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ricky</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Ankur</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>Sam</td>
<td>25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>id</th>
<th>Interests</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Soccer</td>
</tr>
<tr>
<td>1</td>
<td>Movies</td>
</tr>
<tr>
<td>1</td>
<td>Baseball</td>
</tr>
<tr>
<td>3</td>
<td>Music</td>
</tr>
</tbody>
</table>

Smart Drill Down might pose a stress on traditional relational database access might be better to use a columnar database like Hbase.
High Memory usage to store sample datasets during drill down
By its very definition, smart drill down is a NP-Hard Problem which means at any point in time we do not know how much resources to allocate to the processing of a smart drill down (it is variable).
More than a billion tuples leads to frequent hard disk access which slows down the whole process.
Recap - Query Steering

• Process of assisting a user to navigate through a complex data space.
  ➔ Assistance to the user (recommendations)
  ➔ Faster process of the queries.

• Query session (i.e. sequence of queries) generated by:
  - by the user
  - by the system (learning)
Recap - Query Steering

Steering modes:

1. Manual Steering
   User manually specifies the queries in the session one by one.

2. Power Steering
   User can specify either an arbitrarily long prioritized query sequence at once or the steering goals.

3. Auto Steering
   System automatically recommends queries through the user profile.
Query Steering drawbacks

- There is only a theoretical approach on how to develop the steps for an efficient query steering.

- Studies merely suggest that this technology can be one of the solutions to deal with complex databases rather than assuming itself as an end point.
Query Steering challenges

Profile-driven prefetching and caching

Query checkpointing to facilitate reusable process

Efficient query learning

Each presented steering optimization has a clear challenge that needs to be addressed.
Query Steering challenges

1. Process the query
2. Classify returned objects to relevant or irrelevant
3. Adjust the query for the next iteration

Likely next? Uncertainty

Profile-driven prefetching and caching

User System
Query Steering challenges

Given the probabilistic branching, there will be often multiple likely next queries.

→ The challenge: how to prefetch under such uncertainty

**Profile-driven prefetching and caching**

*Prefetching*: transfer (data) from main memory to temporary storage in readiness for later use.
Query Steering challenges

1. Process the query
2. Classify returned objects to relevant or irrelevant
3. Adjust the query for the next iteration

How to decompose?

Query Checkpointing

Query checkpointing to facilitate reusable process

Prefetching Reusability
The challenge: Design these checkpoints to maximize reusability in consideration of the predicted future queries.
Query Steering challenges

1. Process the query
2. Classify returned objects to relevant or irrelevant
3. Adjust the query for the next iteration

Classification of a sample of objects by the user

Learning by the system

Sample selection
Query Steering challenges

- Users characterize data samples (objects, features, etc.) as relevant or irrelevant to their interests.
  → Sample selection in learning
- Minimize the number of iterations required to converge to a result of an acceptable quality