A General Model for OLAP of Complex Data

Jian Pei
State University of New York at Buffalo, USA
http://www.cse.buffalo.edu/faculty/jianpei/
Outline

• Motivation
• GOLAP – a general OLAP model
• Applying GOLAP on complex data
• Conclusions
### OLAP on Relational Data

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Store</td>
<td>Sales</td>
</tr>
<tr>
<td>Product</td>
<td></td>
</tr>
<tr>
<td>Season</td>
<td></td>
</tr>
<tr>
<td>Store S1</td>
<td>Product P1</td>
</tr>
<tr>
<td>Store S1</td>
<td>Product P2</td>
</tr>
<tr>
<td>Store S2</td>
<td>Product P1</td>
</tr>
</tbody>
</table>

**Operations:**
- Roll-up
- Drill-down
- Slice, dice, pivot (rotate)
Why OLAP is Desirable?

• Multi-level, multi-dimensional summarization
  – Identify multi-level, multi-dimensional trends, changes and exceptions

• Can we conduct OLAP on complex data?
  – Data types: strings, time series, sequences, XML documents, …
  – “What are the major patterns among the gene expressions that are similar to the given new sample?”
Gene Expression Matrix
Can We OLAP Gene Expression Data?

• Gene expression data – matrices
  – Oh, it can be treated as a relational table! 😊

• Syntax problem: what should be the measure?
  – SUM, MAX, MIN, AVG? They do not make sense! 😞
  – The patterns are wanted

• Semantic problem: what should be the OLAP operations? 😞😞😞
  – What is the meaning by generalizing (roll up) a sample/gene?
Good News, We Are Not Far Away

• Two major issues in defining an OLAP model
  – How to partition the data into summarization units at various levels?
  – How to summarize the data?
• The summarization units for OLAP should yield to some nice hierarchical structure
  – What about a lattice? – It’s nice
GOLAP – A General OLAP Model

• Base database – a set of objects
• Grouping function
  – Map a set of query objects in the base database to the smallest summarization unit covering the query set
  – Containment: a summarization unit is still in the base database
  – Monotonicity: $Q_1 \subseteq Q_2 \Rightarrow g(Q_1) \subseteq g(Q_2)$
  – Closure: a summarization unit is self-closed
Grouping Function and Class

- **Class**: a subset of objects $S$ s.t. $g(S) = S$

The whole base database itself is a class.
Grouping Function – Lattice

• The classes generated by a grouping function form a lattice

• **Good news:** containment, monotonicity and closure are sufficient to get a nice hierarchical structure!

• Member function: from class to the set of members
Summarization Function

- A mapping from a set of objects to a summary
  - A set of sequences → the sequential patterns
  - A set of time series → the dominant pattern
  - A set of XML trees → the frequent subtrees
OLAP Operations

• Given
  – A grouping function
  – A summarization function

• OLAP operations
  – Summarize: return the summary of the smallest class covering the query set
  – Roll up: return the summary of the smallest class covering the query set and the current class
  – Drill down: return the summary of the smallest class covering the current class except for the query set
GOLAP Model and Data Warehouse

• GOLAP model \((g, f)\)
  – \(g\) – grouping function
  – \(f\) – summarization function

• G-warehouse \(\{(c, f(c))\}\)
  – \(c\) is a class

• \((g_1, f_1)\) and \((g_2, f_2)\) are two GOLAP models. Then, \(((g_1, g_2), (f_1, f_2))\) is also a GOLAP model

• GOLAP on relational data is consistent with the traditional OLAP model
Applying GOLAP on Complex Data

• How to find a meaningful grouping function?
  – Use clusters from hierarchical clustering

• What kind of hierarchical clustering can lead to a grouping function in GOLAP?
  – Each cluster contains a subset of objects
  – The hierarchy covers every object
  – The whole set of objects is the root cluster
  – Ancestor/descendant relation based on containment
  – For any two clusters \( c_1 \) and \( c_2 \), \( c_1 \cap c_2 \) is a cluster if it is not empty
Fixing the Clustering Methods

• Many hierarchical clustering methods, but not all, satisfy the requirements
  – The requirement “c₁ ∩ c₂ is a cluster” may be violated by some methods

• Fix: make the non-empty intersections of clusters as “intermediate clusters”
GeneXplorerer: A GOLAP System

• OLAP gene expression time series data
• Use a hierarchical clustering
  – Based on attraction tree – the index structure of G-data warehouse
• Coherent patterns as summarization
• Basic operations
  – Roll up
  – Drill down
  – Slice
Towards Interactive Exploration of Gene Expression Patterns

- Mine hierarchical clusters of co-expressed genes and coherent patterns
Indexing Clusters
Interactive Exploration on Iyer’s Data Set
## Comparison with Other Methods

<table>
<thead>
<tr>
<th>Pattern</th>
<th>GeneXplorer(9)</th>
<th>Adapt(7)</th>
<th>CLICK(7)</th>
<th>CAST(9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.993</td>
<td>0.956</td>
<td>0.884</td>
<td>0.955</td>
</tr>
<tr>
<td>2</td>
<td>0.957</td>
<td>0.911</td>
<td>0.991</td>
<td>0.887</td>
</tr>
<tr>
<td>3</td>
<td>0.984</td>
<td>0.993</td>
<td>0.994</td>
<td>0.997</td>
</tr>
<tr>
<td>4</td>
<td>0.980</td>
<td>0.984</td>
<td>0.883</td>
<td>0.968</td>
</tr>
<tr>
<td>5</td>
<td>0.958</td>
<td>0.855</td>
<td>0.868</td>
<td>0.855</td>
</tr>
<tr>
<td>6</td>
<td>0.952</td>
<td>0.989</td>
<td>0.970</td>
<td>0.984</td>
</tr>
<tr>
<td>7</td>
<td>0.967</td>
<td>0.976</td>
<td>0.990</td>
<td>0.719</td>
</tr>
<tr>
<td>8</td>
<td>0.991</td>
<td>0.997</td>
<td>0.914</td>
<td>0.999</td>
</tr>
<tr>
<td>9</td>
<td>0.702</td>
<td>0.824</td>
<td>0.844</td>
<td>0.800</td>
</tr>
<tr>
<td>10</td>
<td>0.974</td>
<td>0.981</td>
<td>0.976</td>
<td>0.996</td>
</tr>
</tbody>
</table>

Each cell represents the similarity between the pattern reported by different approaches and the corresponding pattern in the ground truth.
Other Features of GeneXplorer

• Model adjustment – GOLAP models as plug-ins
  – User can change the grouping function and summarization function

• Gene annotation panel
  – Link patterns to ground truth from public annotations
  – Pattern and object visualization
Conclusions

• Problem: how to construct a general model for OLAP on complex data?

• Solution: GOLAP – a general model
  – Consistent with traditional OLAP on relational data
  – Can handle complex data

• A case study: GeneXplorer
Future Work

• Is it necessary to introduce new OLAP operations for complex data?
  – Data/application oriented or general?
• Efficient implementation of G-warehouse
• Data integration based on general OLAP on complex data
Thank You!

http://www.cse.buffalo.edu/faculty/jianpei/