DATABASE INTEGRATION FOR DATA MINING

Databases and DBMS

Data Warehousing
Jennifer Widom, “Research Problems in Data Warehousing”, Int’l Conf. on Information and Knowledge Management ’95.

Data Mediation
INTEGRATION OF HETEROGENEOUS DATA SOURCES


Databases:

- DBMS: Data + Management system

- (Some) Data models:
  - relational
  - network
  - hierarchical
  - entity–relationship (ER)
  - object–oriented
  - object–relational
  - logic–based

- Management Systems:
  - query processing
  - transaction management
  - metadata management
  - storage management
  - maintaining integrity and security
  - fault tolerance
Relational Data Model [Codd70]

- A database is a collection of relations:

**EMP**

<table>
<thead>
<tr>
<th>SS#</th>
<th>Ename</th>
<th>Salary</th>
<th>D#</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>John</td>
<td>20K</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>Paul</td>
<td>30K</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>Mary</td>
<td>40K</td>
<td>20</td>
</tr>
</tbody>
</table>

**DEPT**

<table>
<thead>
<tr>
<th>D#</th>
<th>Dname</th>
<th>Mgr</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Math</td>
<td>Smith</td>
</tr>
<tr>
<td>20</td>
<td>Physics</td>
<td>Jones</td>
</tr>
</tbody>
</table>
Relational Data Model (cont.)

- Query language: SQL (Structured Query Language)

The query:

```sql
SELECT E.Ename, D.Dname
FROM EMP E, DEPT D
WHERE E.Salary < 35K and E.D# = D.D#
```

returns:

```
<table>
<thead>
<tr>
<th>Ename</th>
<th>Dname</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>Math</td>
</tr>
<tr>
<td>Paul</td>
<td>Physics</td>
</tr>
</tbody>
</table>
```
Integrating Databases with Data Mining

**Approach 1:** Loose Integration between DBMS and data mining tool

```
Data Mining Tool
| V
| DBMS
| V
| Database
```

**Approach 2:** Tight Integration between DBMS and data mining tool

```
Data Mining Tool + DBMS
| V
| Database
```
Approaches to Integration

Data Warehouses vs. Mediators

● Data Warehouse
  Data is normalized and copied into a central, single database

  + Static normalization of the data
  - Not up-to-date data

● Mediators
  Virtual integration of the data ("integration-on-demand")

  + Up-to-date data
  - Dynamic normalization of data
Data Warehousing

Data Warehouse: (central) database that brings together selected data from multiple databases

• replicates information: alternatives:
  – no pre-processing
  – eliminating inconsistencies/redundancies
  – selecting a subset of the data
  – pre-analyzing data for predictable uses

• persistent
• in-advance integration
• information might be out-of-date
• application-oriented
• historic data
• aggregated and summarized information
• data sources are usually operated independently of the D.W.
• facilitates mining (but is not essential)
Data Warehouse – Architecture
Data Warehouse – Architecture

Wrapper/Monitor:

• translating information from source format into warehouse format
• automatically detecting changes of interest in the source data
  – cooperative sources
  – logged sources
  – queryable sources
  – snapshot sources
• reporting changes in the source data to the integrator
Data Warehouse – Architecture

Integrator:

• installing the reported information in the warehouse
  – filtering information
  – summarizing information
  – merging information
Mediators

- Proposed by Wiederhold [1992,1993]
- Current mediator/data integration projects:
  - TSIMMIS: Stanford Univ.
  - HERMES: Univ. of Maryland
  - INFORMATION MANIFOLD: AT&T Bell Labs.
  - SIMS: Information Sciences Institute (ISI/USC)
  - SOFTBOT: Univ. of Washington
  - DIGITAL LIBRARY PROJECT: Univ. Michigan
  - NOMENCLATOR: AT&T Bell Laboratories.
  - KQML: Univ. of Maryland, Baltimore.
  - WEBWATCHER: Carnegie Mellon Univ.
  - OCCAM: Univ. of Washington
  - KIF and Interoperable agents project, Stanford Univ.
Mediator Architecture

(Taken from [Ull97])
Queries and Query Languages

- **Languages**
  - **Mediators:**
    All mediators speak the same common query language.
  - **Wrappers:**
    Each wrapper translates between the mediators’ language and the data source’s language.

- **Queries and Query Reformulation**
  - **Users:** generate queries $Q$ in the common language
  - **Mediators:** reformulate $Q$ in terms of queries $Q_1, \ldots, Q_n$ (in the common language)
  - **Wrappers:** translate each $Q_i$ into the source’s language
  - **Data Sources:** answer queries $Q_1, \ldots, Q_n$
Mediator Components

• Mediator

-----------------------------------------------
<table>
<thead>
<tr>
<th>Program</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source Description</td>
</tr>
</tbody>
</table>

• Source Description

World-View Relations

\[
\text{\( \sim \)}
\]
| Mapping
| V

Data Source Relations
Approaches to Mediation

• Source-driven:
  All mediators use the same program ("facilitator")

  + Easy to add new data sources
  - Potential conflict between answers from different mediators

• Specification-driven:
  Each mediator supports a different, fixed set of queries.

  + No conflicts between answers from different mediators
  - Laborious to add new data sources - may require to reprogram some mediators.
Query Reformulation Problem

Given:
a query $Q$ in terms of world-view relations

Find:
a query $Q'$ in terms of source relations only
s.t. the answers to $Q'$ provide all possible answers to $Q$. 