DSE 203

DAY 1: DATA INTEGRATION TASKS
(A PREVIEW OF THE REMAING COURSE)

Source Description

- Catalog of sources and their capabilities
  - Schema if available
    - Export schema, views, …
    - Data description
      - Data types
      - Complex structures
    - Constraints that apply
  - Access methods
    - Can be modeled as schemas
    - Invocation specification
    - Calling restrictions, update rates
  - Auxiliary Information
    - Rough estimate of data size
    - Location, availability parameters, communication delays, reliability

Matching

- What needs to match?
  - Elements of the target schema with elements of source schemata
    - How to construct a “value” in the target element as a function of corresponding values in the source elements
- Schema matching
    - Defining the exact correspondence between elements of the source schemata and the value of a target schema element
- Schema mapping
    - Rules for construction of the target tuples
- Data Matching
  - Value matching
  - “Record” matching

Data Preparation

- For any given source can the data be used “as is” for integration?
- Does it need reformatting?
- Does it need restructuring?
- Does it need discarding “noise data”?
- Do we need all the data that is provided by the source?
- Do we need to transform/wrap/translate the data from its current state to an integration-ready state?
- Are the transformation rules completely known?
- Is human validation needed for the data or the data transformation?

Data Transformation

- Stems from the semantic incongruity problem
- Structural types vs. semantic types
  - Integer (with non-negative constraint) vs. age
- Transformations
  - Functions that convert structures (and values) but maintain semantic type
  - Unit conversions
  - Numerical manipulations
- Implementation through UDFs (User Defined Functions) in an SQL setting

Model Manipulation

- The data sources to be integrated can come from different data models
- Often different data models are transformed into a common model
- Are there a set of operations for model transformation?
- Are there ways to perform these model transformations automatically?

Query Planning and Evaluation

- The user asks a query to the integrated schema
- Which sources should be used for this query?
  - The source selection problem
- How does the query get reformulated to smaller, component queries that would be evaluated by each source?
  - The query reformulation problem
- Where does the query execute?
  - Different answers depending on the integration architecture
- What happens when data sources are to be accessed over the Internet?

Query Reformulation

- The problem
  - Given
    - A query against a target schema
    - Source descriptions
    - Mapping information
  - Rewrite new queries against the source

Target Relation: Products(ProductID, ProductName, ProductDescription) ←

Source Relations
- PolicyTypes('P-'+PolicyTypeKey, Name, Description) ∨
- AccountType('B-'+TypeID, Name, Description)

We will sometimes use a logical language to describe queries

Data Aggregation

- Special case of data integration – vertical integration
- The schema of each source is identical
  - Virtual bookstore integrating other stores for regular books, textbooks and antique books
- The sources may have overlapping content
- What happens when there are too many sources?

Data Fusion

- Information sources like the internet often have small fragments of information about an entity
- These information fragments are heterogeneous, almost never complete and sometimes inaccurate
- This gives rise to potential conflicts when all data are assimilated into the integration system
- Data Fusion
  - Combining multiple records representing the same real-world object into a single, consistent, and clean representation