Information Integration Research

A Brief Outlook

Joachim Hammer

Lots of Progress

- Languages for mediation
- Query processing techniques
- Construction toolkits for wrappers and mediators

But …

- Much remains to be done
  - Managing semantic heterogeneity
  - Use of domain knowledge
  - Move from query-only tools to more active components

Driving Forces

- Continuous increase in the number of sources that must be integrated
  - Integrate sources at scale (100s to 1,000s)
  - Tools for automated discovery of new schemas
- Desire to share information via Web and Web services
- New applications
  - E.g., bioinformatics, autonomous agents, …
- Integrate a wide variety of sources: structured, semistructured, text, multimedia, streams, …
  - Adapt in the presence of unreliable data
- Protect data privacy, support secure data access

Let’s look at some specific research challenges

Reconciling Heterogeneous Schema and Ontologies

- Need semantic mappings between respective representation
- Schema matching or ontology alignment is currently human centered task
- Need tools to aid human designer and improve their productivity
  - Tools to improve over time, scale up

On-the-fly Integration

- Currently mediators (integration systems) rely on relatively static configuration with a set of long-lived data sources
- Need to integrate data from source “immediately”’ after discovering it
  - Source may only be used a few times

“Deep-Web” Integration

- Significant portion of Web data hidden behind query interfaces of searchable databases
- Need tools to
  - Discover these sources
  - Integrate them
  - Support efficient query processing

Management of Changes in Data Integration

- Sources change
  - Genomic repositories grow by 100s of GB every year
- Need to handle updates
- Want to be notified if data item is available in a source without repeatedly sending query
- Active mediators and wrappers?
  - Similar to publish/subscribe systems but need to deal with heterogeneity of data sources and large-scale distribution

Combining Structured and Unstructured Data

- E.g., databases and Web pages
- Not clear how to query such an integrated repository
- Need languages appropriate for such queries and efficient methods for processing them

Managing Uncertainty

- Most integrated data is inconsistent, incomplete, or uncertain
- Tell users about quality of data
  - Guiding them through process of reconciling the affected data items

Use of Domain Knowledge

- Domain knowledge can help
  - Find sources
  - Understand source data
  - Formulate queries
  - Explain results
  - Express semantic mappings
- How to extract useful domain knowledge
  - How to keep it updated?

Protect Privacy

- People don't share unless appropriate security polices in place
  - Hospitals, e-commerce
- E.g., need secure wrappers that can export patient data in such a way that
  - Aggregate information is visible but no correlation to individual can be made, or
  - Certain fields are hidden
  - Critical data is perturbed and cannot be guessed
  - …
- Very related to field of statistical databases

Into the Future…

- Need theory, shareable toolkits, benchmarks
- Progress in this field will be result of collaborations among experts from Database Systems, Artificial Intelligence, and Information Retrieval
- Lots of opportunities for research