CSE 494/598
Lecture-14: Information Integration

LYDIA MANIKONDA
HTTP://WWW.PUBLIC.ASU.EDU/~LMANIKON/

**With permission, content adapted from last year’s slides**
Announcements

• Final exam: **May 2nd 2016, 12:10 – 2:00 PM BYAC 240** (same as our regular class room)
• Class standings till now will be up by tonight.
Today’s lecture

- Information Integration
- Revision
Information Integration

• Combining information from multiple autonomous information sources
  • Answering queries using the combined information

• Applications

• WWW
  • Comparison shopping
  • Portals integrating data from multiple sources
  • Electronic marketplaces
  • Mashups

• Science Informatics
  • Integrating genomic data, geographic data, archaeological data, astro-physical data, etc.

• Enterprise data integration
  • An average company has 49 different databases and spends 35% of its IT dollars on integration efforts
Deployed Information Integration Systems

- Travel sites: Kayak, Expedia, etc.
- Google Base, Dbpedia
- Map mashups
- Citeseer, Google Scholar, etc.
Blind Men & the Elephant: Differing views on Information Integration

Database View
- Integration of autonomous structured data sources
- Challenges: Schema mapping, query reformulation, query processing

Web service view
- Combining/composing information provided by multiple web-sources
- Challenges: learning source descriptions; source mapping, record linkage etc.

IR/NLP view
- Computing textual entailment from the information in disparate web/text sources
- Challenges: Information Extraction
--Search Model--
--Materialize the pages
--crawl & index them
--include them in search results

--Mediator Model--
--Design a mediator
--Reformulate queries
--Access sources
--Collate results

--Warehouse Model--
--Get all the data and put into a local DB
--Support structured queries on the warehouse
Dimensions of Variation

Conceptualization of (and approaches to) information integration vary widely based on

- **Type of data sources**: being integrated (text; structured; images etc.)
- **Type of integration**: vertical *vs.* horizontal *vs.* both
- **Level of up-front work**: *Ad hoc* *vs.* pre-orchestrated
- **Control over sources**: Cooperative sources *vs.* Autonomous sources
- **Type of output**: Giving answers *vs.* Giving pointers
- **Generality of Solution**: Task-specific (Mashups) *vs.* Task-independent (Mediator architectures)
Dimensions: Type of Data Sources

Data sources can be

- **Structured** (e.g. relational data)
  - Schema mapping (GAV/LAV)
  - Precise integration (Could I have done better by going to the sources myself?)

- **Text oriented**
  - Information Extraction

- **Mixed**
Dimensions: Type of Output

The cost-effective approach may depend on the quality guarantees we would want to give.

At one extreme, it is possible to take a “web search” perspective—provide potential answer pointers to keyword queries

- Materialize the data records in the sources as HTML pages and add them to the index

At the other, it is possible to take a “database/knowledge base” perspective

- View the individual records in the data sources as assertions in a knowledge base and support inference over the entire knowledge.
  - Extraction, Alignment etc. needed
Dimensions: Vertical vs. Horizontal

**Vertical:** Sources being integrated are all exporting same type of information. The objective is to collate their results

- Eg. Meta-search engines, comparison shopping, bibliographic search *etc.*
- Challenges: Handling overlap, duplicate detection, source selection

**Horizontal:** Sources being integrated are exporting different types of information

- E.g. Composed services, Mashups,
- Challenges: Handling “joins”

Both..
Dimensions: Level of Up-front Work Ad Hoc vs. Pre-Orchestrated

FULLY QUERY-TIME

- Get a query from the user on the mediator schema
- Go “discover” relevant data sources
- Figure out their “schemas”
- Map the schemas on to the mediator schema
- Reformulate the user query into data source queries
- Optimize and execute the queries
- Return the answers

FULLY PRE-FIXED

- Decide on the only query you want to support
- Write a (java)script that supports the query by accessing specific (pre-determined) sources, piping results (through known APIs) to specific other sources
- Examples include Google Map Mashups
Dimensions: Control over Sources (Cooperative vs. Autonomous)

Cooperative sources can (depending on their level of kindness)
- Export meta-data (e.g. schema) information
- Provide mappings between their meta-data and other ontologies
  - Could be done with Semantic Web standards...
- Provide unrestricted access
  - Examples: Distributed databases; Sources following semantic web standards

...for uncooperative sources all this information has to be gathered by the mediator
- Examples: Most current integration scenarios on the web
Collated Challenges

Source Selection
- Which sources are likely to be most relevant to answer the query?
  - Source quality
  - Source overlap

Schema Mapping
- How to map schemas of the sources to the mediator?

Record Linkage
- How to couple data items that refer to the same entity

Information quality
- Data with null values
- Inconsistent data

Query reformulation
- To convert query from the mediator schema to the source schema
- To bring out relevant data despite incompleteness and inconsistency
Term Frequency & Inverse Document Frequency
Inverted Index
Distance Metrics
Scalar Clusters
HITS algorithm
PageRank Algorithm
Network Theory
Scale –free Networks
Clustering
Classification
Today

• Information Integration
• Revision