Probabilistic Data Integration Systems

Dan Suciu

Joint with Nilesh Dalvi, Brian Harris, Brent Louie, Chris Re

Motivation

- Today's data integration techniques are well-suited for well-structured data integration tasks
  - Enterprise applications
- Do not cope well with imprecisions
  - Scientific data, heterogeneous data (text), even enterprise data
- This talk proposes: *Probabilistic Data Integration*
  - Imprecisions represented explicitly as probabilities at all levels of integration

Example of Imprecision: Object Reconciliation

- Same object has different representations in two different databases:
  - Same gene in Entrez and Swissprot
- *Record linkage*, *deduplication*, etc
- Today's techniques:
  - Similarity score
  - *Threshold* to classify into match/non-match

PDIS: keeps the scores natively, as probabilities

Information Extraction

- Increasing amount of data extracted from text
  - Email messages
  - User manuals
  - Blogs
- Example: the Avatar System at IBM
- Data is *inherently* imprecise:
  - Extracted data has limited use
- Data integration needs to account for imprecision

PDIS: keeps the scores natively, as probabilities

Scientific Data

- Different sources have various degrees of quality:
  - Curated/uncurated
  - Links computed using various methods

Uncertainty Metrics For Result Ranking in BioMediator

- Quality of data source
  - 1.0: SwissProt (Human curated)
  - 0.7: Trembl (Predictions)
- Quality of record in a source
  - 1.0 : Validated (RefSeq status code)
  - 0.7: Provisional (RefSeq status code)
- Quality of links between sources
  - 1.0: Foreign keys (RefSeq ids in records)
  - 0.7: Text search of record (OMIM)
- Quality of a specific link between records
  - E-values (BLAST, HMM's)
  - Foreign keys
- ***Overall Uncertainty (UII) Score***
  - Per entity, takes into account all sources of uncertainty
  - Belief in the quality of the "relevance" between query seed and a particular result
  - ***The results can then be ranked by UII score***

Courtesy of Brent Louie

PDIS

- Represent imprecisions at all level of data integration:
  - Wrappers
  - Data transformations
  - Mediated schema
  - Query language
  - Query answers

Two Challenges

- Efficient query processing
  - At UW we have made progress with the MystiQ system
- Constraints

Query Processing in MystiQ (Demo)

http://mystiq.cs.washington.edu/

Constraints

- “address” may be shippingAddress or billingAddress:

$\forall$ t $\in$ LocalData $\Rightarrow$ $\exists$ u $\in$ MediatedData. t.address=u.shippingAddress
confidence = 0.8

$\forall$ t $\in$ LocalData $\Rightarrow$ $\exists$ u $\in$ MediatedData. t.address=u.billingAddress
confidence = 0.2

Taxonomy of Constraints

	Deterministic data	Probabilistic data
Deterministic constraints	Constraint violations	Cleaning
Probabilistic constraints	Probabilistic data exchange…	… of probabilistic data

Conclusions

- Trend in Computer Science:
  - deterministic --> probabilistic
- Data integration under similar pressures

- What's in for us:
  - This stuff is hard, researchers needed !!!
  - Query processing: still ill understood
  - Constraints, mappings,XML etc: way in the future