From Complete to Incomplete Information and Back

L. Antova et al.
Presented by: Nedyalko Borisov

Instructor: Jun Yang
Duke University
Computer Science
Jan 29, 2008

Motivation

- Incomplete information
  - Database with missing information
    - web info extraction
    - manage information systems
    - no solution for now
  - Answers set programing
    - used in AI (knowledge representation)
    - not scalable

Example

Buy company, preserve skill web, one employee could leave

\[
U \leftarrow \text{select choice of Company.Emp} \mid \text{CID};
\]
\[
V \leftarrow \text{select R1.CID, R1.EID} \mid \text{from Company.Emp R1, (select \* from U choice of EID) R2}
\quad \text{where} \quad R1.CID = R2.CID \text{ and } R1.EID = R2.EID;
\]
\[
W \leftarrow \text{select certain CID, Skill} \mid \text{from V, Emp.Skill EID = Emp.Skill,EID (select CID from V)};
\]
\[
\text{select possible CID} \mid \text{from W where Skill = 'Web'};
\]

I-SQL vs SQL

Buy company, preserve skill web, one employee could leave

\[
U \leftarrow \text{select \* from Company.Emp} \mid \text{choice of CID};
\]
\[
V \leftarrow \text{select R1.CID, R1.EID} \mid \text{from Company.Emp R1, (select \* from U choice of EID) R2}
\quad \text{where} \quad R1.CID = R2.CID \text{ and } R1.EID = R2.EID;
\]
\[
W \leftarrow \text{select certain CID, Skill} \mid \text{from V, Emp.Skill EID = Emp.Skill,EID (select CID from V)};
\]
\[
\text{select possible CID} \mid \text{from W where Skill = 'Web'};
\]

World-set SQL

- I-SQL
  - analog of SQL
  - language of search engine
  - based on world-set algebra
  - generic - preserve independece of the data
  - expressive - common queries + incomplete info
  - conservative - analog to relational algebra
  - efficient evaluation

World-set Algebra (WSA)

- World-set algebra
  - world-set algebra to ISQL <=> relational algebra to SQL
  - set semantics
  - no bag semantics and aggregations
  - conservative over relational algebra => each 1W/1W query <=> SQL
World Manipulation 1

- Data manipulation
  - I-SQL uses the standard SQL commands - insert, update and delete

- Merging worlds
  - possible and certain - compute tuples that appear in some worlds and adds them to each of the input worlds
  - group-worlds by - group worlds by certain criteria

World Manipulation 2

- Splitting worlds
  - choice-of - freeze the values for a given set of attributes and analyze each combination in separate world
  - repair-by-key - constructs possible repairs of a relation that violates a uniqueness constraint

Syntax and Semantics 1

\[
\begin{align*}
\prod^{i}_{j=1} \phi \cdot \sigma\alpha (\mathbf{R}) & = \{ \mathbf{R} \mid \exists \mathbf{A} (\mathbf{R} \subseteq \mathbf{A}) \land \exists \mathbf{B} (\mathbf{A} \subseteq \mathbf{B}) \\
\prod^{i}_{j=1} \phi \cdot \sigma\alpha (\mathbf{R}) & = \{ \mathbf{R} \mid \exists \mathbf{A} (\mathbf{R} \subseteq \mathbf{A}) \land \exists \mathbf{B} (\mathbf{A} \subseteq \mathbf{B}) \land \exists \mathbf{C} (\mathbf{B} \subseteq \mathbf{C}) \\
\prod^{i}_{j=1} \phi \cdot \sigma\alpha (\mathbf{R}) & = \{ \mathbf{R} \mid \exists \mathbf{A} (\mathbf{R} \subseteq \mathbf{A}) \land \exists \mathbf{B} (\mathbf{A} \subseteq \mathbf{B}) \land \exists \mathbf{C} (\mathbf{B} \subseteq \mathbf{C}) \land \exists \mathbf{D} (\mathbf{C} \subseteq \mathbf{D}) \}
\end{align*}
\]

Syntax and Semantics 2

- Operator typing
  - relational algebra operators and groups operators are: 1 \rightarrow 1 or m \rightarrow m
  - possible and certain operators are: m \rightarrow 1
  - choice-of operator is: 1 \rightarrow m or m \rightarrow m

- Extending Word-set algebra
  - repair-by-key operator
  - NP hard problem
  - query evaluation in Ptime in respect to data complexity

WSA to Relational Algebra

- Condition: complete representation of the input world set
- Inlined representation of World-sets
- World-set translation (next slide)
WSA Translation

\[
\begin{align*}
\tau_f(R_1, \ldots, R_n, W) &= \varpi_f(R_1, \ldots, R_n, W) \\
\varpi_f(R_1, \ldots, R_n, W) &= \pi_f(T) \\
\pi_f(T) &= \pi_f(T) \\
\end{align*}
\]

Progress

- Currently under development system called MayBMS (Cornell database group)
- Built on top of PostgreSQL
- Only small amount of experiments completed

Experiments

- Selection/projections queries in practice have a constant overhead of a factor of 3 to 5 over PostgreSQL running the same queries on single world
- Joins shows exponential behaviour
- Universal operations (difference, certain aggregations) need further work

Related work

- Trio (Stanford University)
  - stores information in weak representation systems
  - use constraints (called lineage) to represent dependencies
  - always applies constraints for answering queries
  - no experiments have been reported

Problems and Discussions 1

- Implementation of I-SQL in real systems
  - Performance
    - not shown how they perform the experiments
    - what metrics are used
  - Translation to SQL?
- From incomplete to complete
  - Not presented

Problems and Discussions 2

- Bag semantics and aggregations
- Do we need the world set semantics?
- Is it worth the efforts? (see the example)
- Probabilistics – claimed to be easy. They have demo on ICDE’07
Thank you!