CS 245: Database System Principles

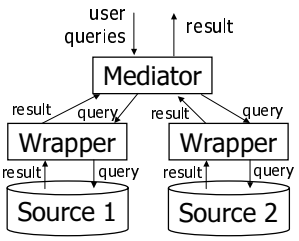
Notes 14: Coping with Limited Capabilities of Sources

Hector Garcia-Molina
(Some modifications by Chris Olston)

Recall …

- Three approaches to information integration:
  - Federated databases ← did teaser
  - Data warehousing ← did overview
  - Mediation ← next

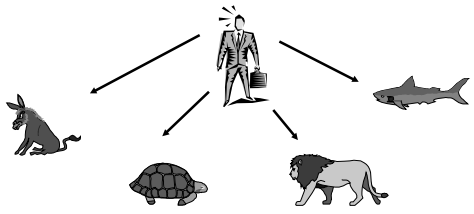
Third Approach to Information Integration: Mediation



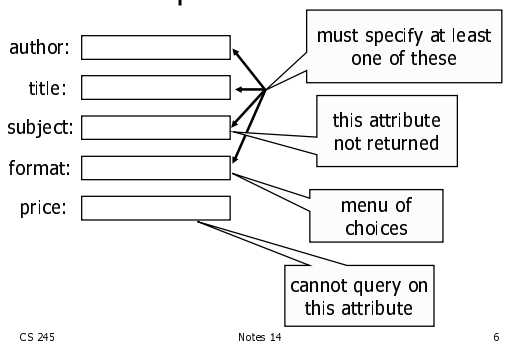
Recall Advantages of Mediation

- No need to copy data
  - less storage
  - no need to purchase data
- More up-to-date data
- Query needs can be unknown
- Only query interface needed at sources
  - Sources may not support "upload all data"

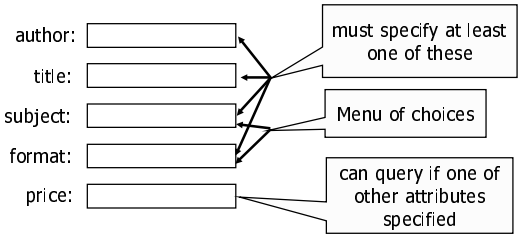
- Worse, sources may provide very limited and differing query capabilities
- One major challenge for mediation: coping with limited source capabilities



Example: Amazon.com



Example: BarnesAndNoble.com



Why Limited Capabilities?

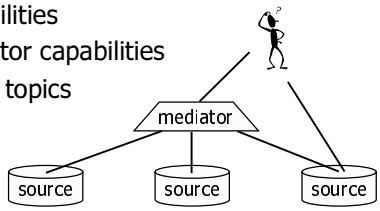
- Search forms
- Security
- Indexes
- Legacy

Capability vs. Content

- Capability description
  - Can only search for subject = "art," "history," "science"
- Content description
  - Source only contains subject = "art," "history," "science"

Outline

- Describing source capabilities
- Extending source capabilities
- How mediators cope with limited capabilities
- Mediator capabilities
- Other topics



Describing Query Capabilities

R(X, Y, … Z)

Adornments:
- **f**: may or may not specify
- **u**: cannot be specified
- **b**: must be specified
- **c[S]**: specified from list S
- **o[S]**: optional, chose from S

Describing Query Capabilities

R(X, Y, … Z)

Adornments:
- **f**: may or may not specify
- **u**: cannot be specified
- **b**: must be specified
- **c[S]**: specified from list S
- **o[S]**: optional, chose from S

With output restriction
- **f'**
- **u'**
- **b'**
- **c'[S]**
- **o'[S]**

Example

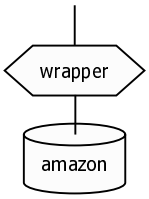
- Relation R(X, Y, Z)
- Description Templates: bu'f, uf'c[z_1, z_2]
- Answerable queries: R(x_1, Y, Z), R(X, Y, z_1)
- Unanswerable queries:
  R(X, y_1, Z), R(X, Y, z_3)

Many Other Description Mechanisms Exist …

- Tsimmis
- Information Manifold
- Disco
- Garlic
- Contex-free grammars

Extending Source Capabilities

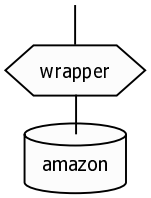
Query: author="Freud" AND price > 10



Source: R(author, price, ...)
Template: b, u, ...

Extending Source Capabilities

Query: author="Freud" AND price > 10



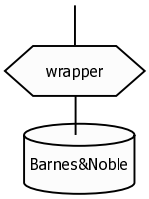
Wrapper Filter: price > 10

Source Query: author="Freud"

Source: R(author, price, ...)
Template: b, u, ...

Another Example

Query: (author = "Freud" OR author = "Jung") AND price < 10

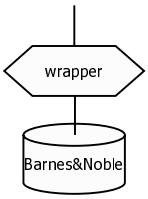


R(author, price, ...)
No disjunctive conditions;
Price can only be specified with author

Another Example

Query: (author = "Freud" OR author = "Jung") AND price < 10

Union Operation



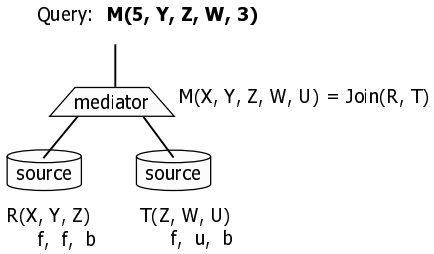
Q1: author = "Freud" AND price < 10
Q2: author = "Jung" AND price < 10

R(author, price, ...)
No disjunctive conditions;
Price can only be specified with author

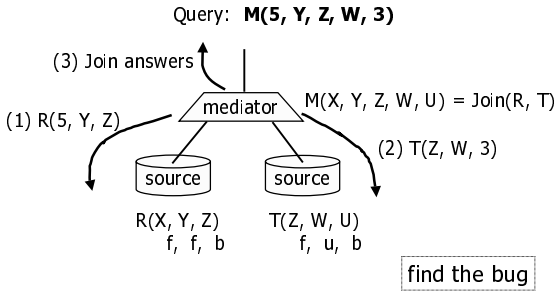
Extending Source Capabilities

- General scheme:
  - try many query rewritings
  - check if query fragments supported by source
  - check if wrapper can combine answer fragments
  - do all this very efficiently!! [See ICDE99 paper]
- Tsimmis, Info Manifold: no disjunctive queries
- DISCO: no query splitting
- Garlic: only CNF queries

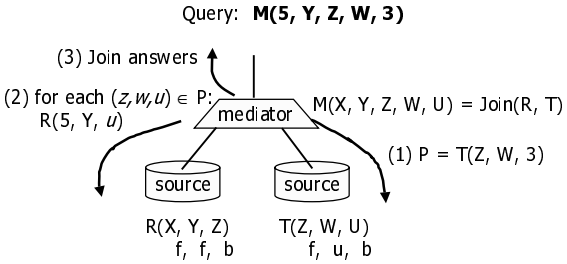
Mediator Processing



Plan 1



Plan 2



Mediator Plan Generation

- Need feasible and efficient plan
- Search space is huge
- Tsimmis, Info Manifold, Garlic:
  - exponential algorithms
- Polynomial algorithms:
  - often find optimal or near-optimal plan
  - bounded performance
  - [See ICDT99 Paper]

Conclusion

- Not all sources are created equal!
- Need to
  - describe what sources can do
  - efficiently process queries even though sources have limited capabilities
  - describe what mediators can do
  - exploit content information
  - deal with unavailable sources