Big Data Integration & Processing: further approaches and applications

Vũ Tuyết Trinh

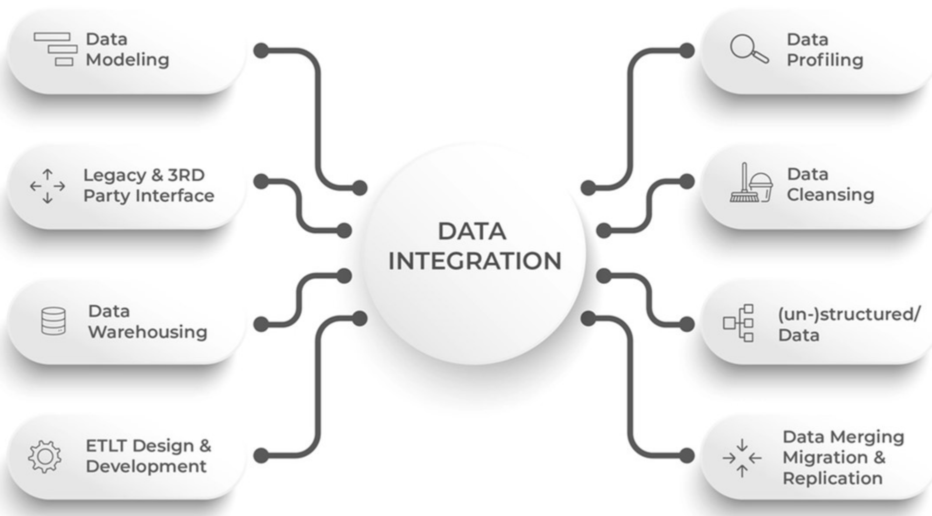
1

Adaptive

- Adaptable: easily adapted to accommodate a change.
  - Customized, personalized, configurable
- Adaptive: consistently able to change itself, to accommodate and maximize the benefits of change.
  - Flexible, scalable, intelligent, dynamic

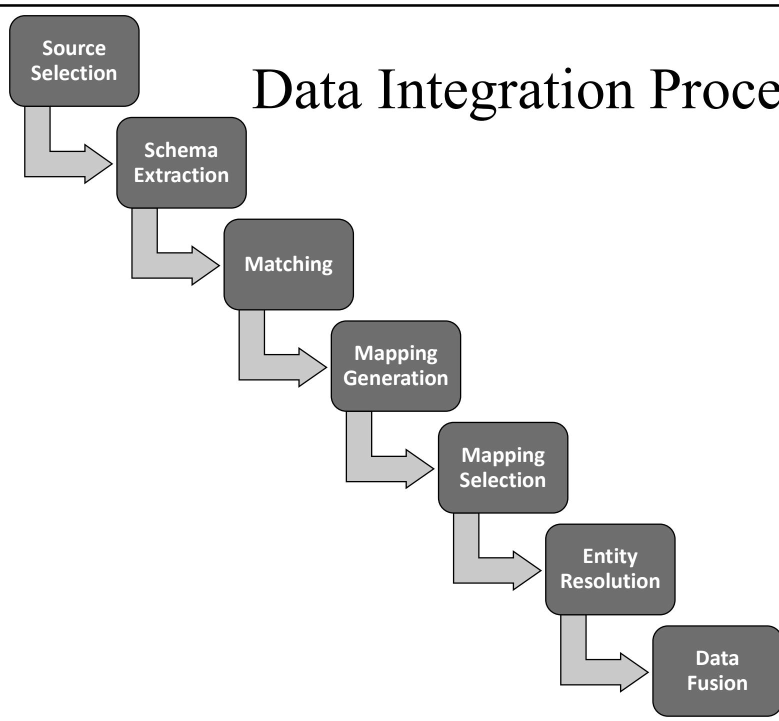
2

Data Integration



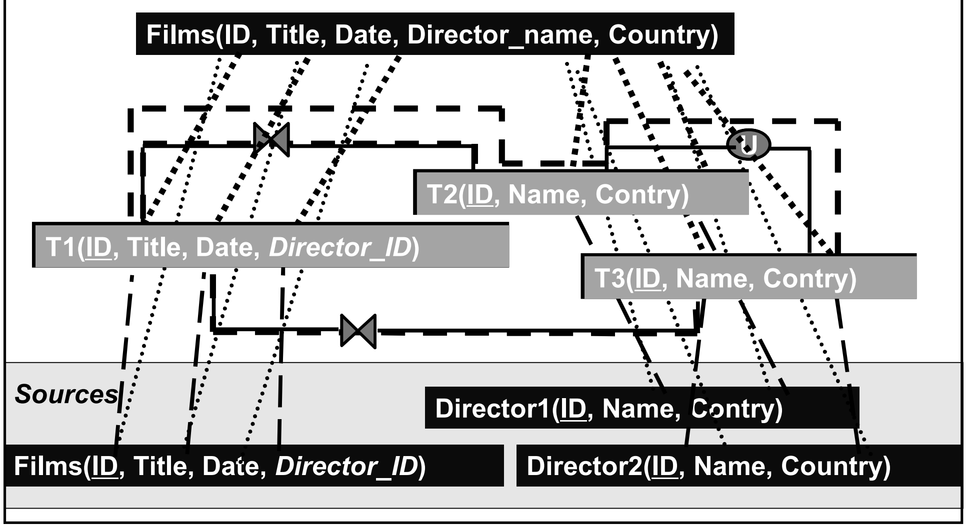
3

Data Integration Process



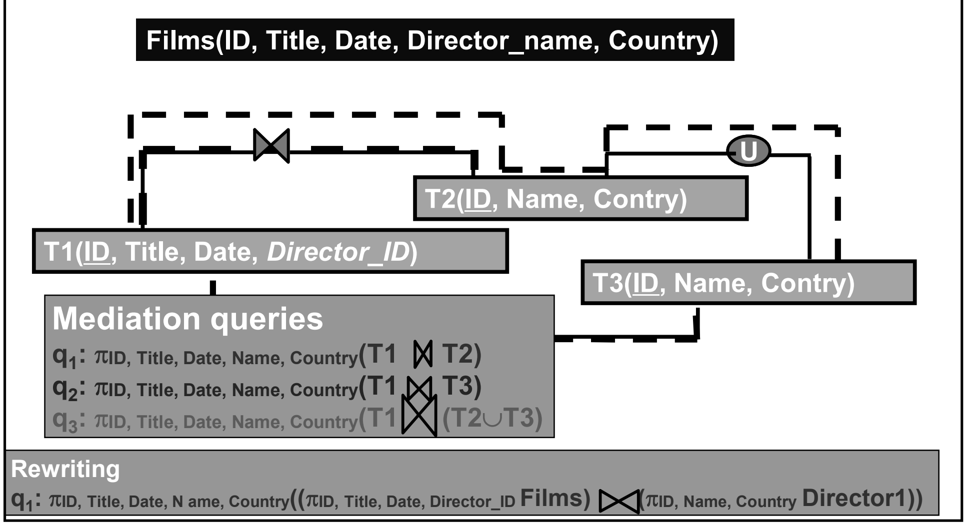
4

Generating mediation queries



5

Generating mediation queries (2)



6

Pay-as-you-go approach

- Accessing multiple data sources without full integration
- Starting with some mapping, improving/discovering more overtime

7

Using probabilistic model

PROBABILISTIC MEDIATED SCHEMA

$\{S_1, \ldots, S_n\}$ *be a set of schemas. A* probabilistic mediated schema (p-med-schema) for $\{S_1, \ldots, S_n\}$ *is a set*

$$\mathbf{M} = \{(M_1, Pr(M_1)), \ldots, (M_l, Pr(M_l))\}$$

where

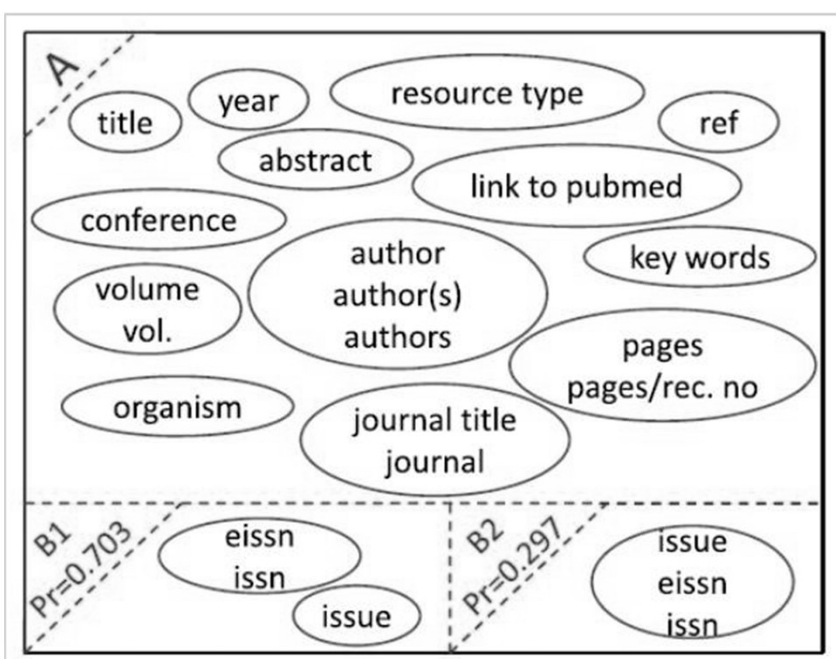
- *for each* $i \in [1, l]$, M_i *is a mediated schema for* $S_1, \ldots, S_n$, *and for each* $i, j \in [1, l], i \neq j$, M_i *and* M_j *correspond to different clusterings of the source attributes;*
- $Pr(M_i) \in (0, 1]$, *and* $\Sigma_{i=1}^{l} Pr(M_i) = 1$. □

8

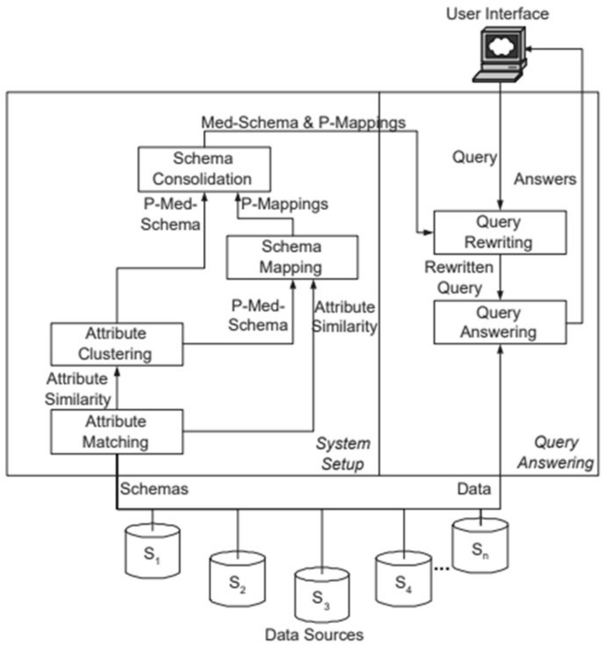
Using probabilistic model:
Mediated Schema Generation

- Remove infrequent attributes
  - Ensure mediated schema contain most relevant attribute
- Construct weighted graph
  - Nodes are remaining attributes
  - Edges are values of some similarity measure $s(a_i, a_j)$
    - Threshold τ
    - Error ε (uncertain)
- Cluster nodes
  - Cluster is a connected component of the graph

9



10



11

Using functional dependencies

$S_1(name, hPhone, hAddr, oPhone, oAddr)$
$S_2(name, phone, address)$

$F_1 = \{hPhone \to hAddr, oPhone \to oAddr\}$
$F_2 = \{phone \to address\}$

$M_1(\{name, name\}, \{phone, hP\}, \{oP\}, \{address, hA\}, \{oA\})$
$M_2(\{name, name\}, \{phone, oP\}, \{hP\}, \{address, oA\}, \{hA\})$

12

FD heuristics

Heuristic 1 *Let S_p and $S_q, p \neq q$, be two source schemas. Then,*

$$Match(a_{p,i}, a_{q,k}) \Rightarrow unmatch(a_{p,i}, a_{q,l}) \wedge unmatch(a_{q,k}, a_{p,j})$$

where $a_{p,i} \in att(S_p), a_{p,j} \in att(S_p) \setminus \{a_{p,i}\}, a_{q,k} \in att(S_q), a_{q,l} \in att(S_q) \setminus \{a_{q,k}\}$.

Heuristic 2 *Let $fd_p : a_{p,i} \rightarrow a_{p,j}$ and $fd_q : a_{q,k} \rightarrow a_{q,l}$ be two FDs, where $fd_p \in F_p, fd_q \in F_q, p \neq q$. Then, $similarity(a_{p,i}, a_{q,k}) > t_L \Rightarrow Match(a_{p,j}, a_{q,l})$ where t_L is a certain threshold and similarity is a given similarity function.*

Heuristic 3 *Let PK_p and $PK_q, p \neq q$, be the primary keys of S_p and S_q respectively. Then,*

$$(\exists a_{p,i} \in PK_p, a_{q,j} \in PK_q \mid (a_{p,i}, a_{q,j}) = \underset{a_p \in PK_p, a_q \in PK_q}{\arg\max} \; similarity(a_p, a_q)) \wedge$$
$$(similarity(a_{p,i}, a_{q,j}) > t_{PK}) \Rightarrow Match(a_{p,i}, a_{q,j})$$

13

FD heuristics (2)

Heuristic 4 *Let PK_p and $PK_q, p \neq q$, be the primary keys of S_p and S_q respectively. Then,*

$$(\exists a_{p,i} \in PK_p, a_{q,j} \in PK_q, fd_p \in F_p, fd_q \in F_q \mid$$
$$fd_p : a_{p,i} \rightarrow R_p, fd_q : a_{q,j} \rightarrow R_q) \Rightarrow Match(a_{p,i}, a_{q,j}) \qquad (1)$$

and also

$$(RHS(1) \wedge R_p = \{a_{p,r}\} \wedge R_q = \{a_{q,s}\}) \Rightarrow Match(a_{p,r}, a_{q,s}) \qquad (2)$$

Heuristic 5 *Let PK_p and $PK_q, p \neq q$, be the primary keys of S_p and S_q respectively. Then,*

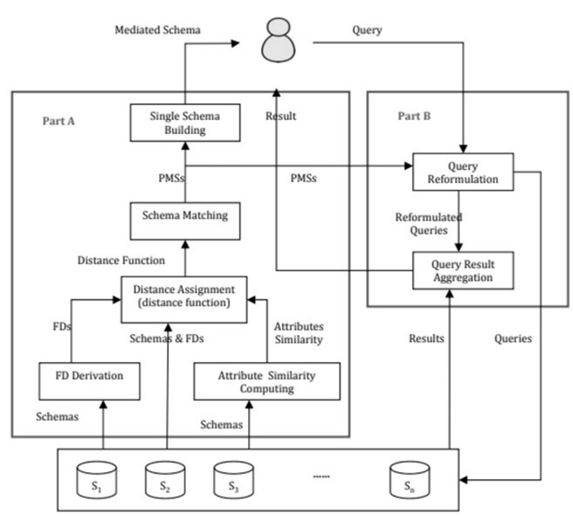
$$(\forall a_{p,r} \in PK_p \setminus \{a_{p,i}\}, \exists a_{q,s} \in PK_q \setminus \{a_{q,j}\} \mid Match(a_{p,r}, a_{q,s})) \wedge$$
$$(|PK_p| = |PK_q|) \Rightarrow Math(a_{p,i}, a_{q,j})$$

14

Distance Assignment

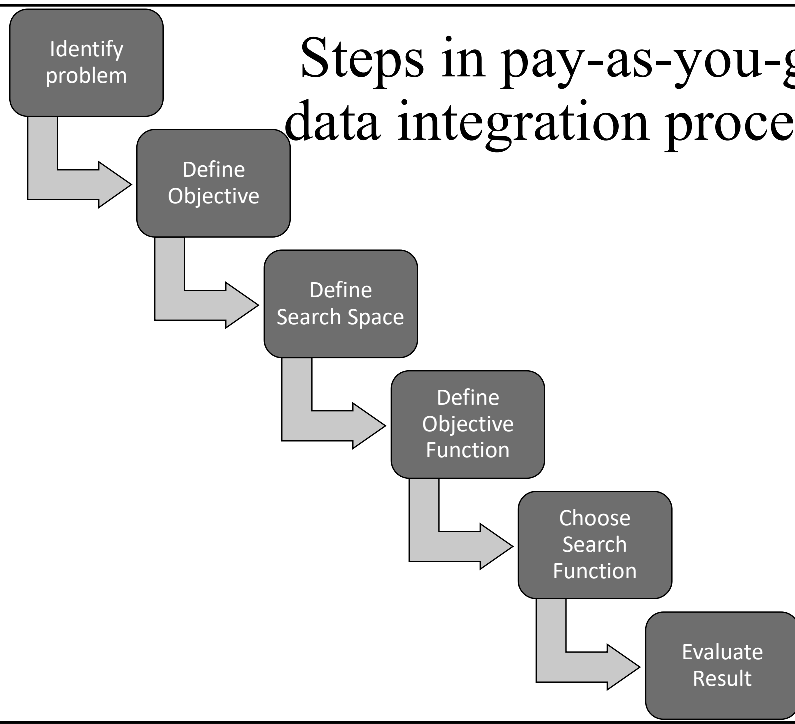
- find the attribute pairs (ap, aq) whose similarity is maximum
  - Probabilistic model, threshold
  - Match distance, unmatch distance
- Find FD pairs from different sources which their left sides match together and then try to match attribute pairs on the right sides of these FDs
- remove the matched attributes from the list of unmatched attributes, and repeat the matching process if there are still some attributes remaining for matching

15



16

Steps in pay-as-you-go data integration process



17

Mashup-based Linked Data Integration

- See 10.Mushup.pdf

18

Applications

- Job portal (see 11_BKWork.pdf)
- Tourist (example https://www.visitacity.com/)
- A Scientific Data and Workflow Sharing System (see 12_scientificFlow_nus.pdf)

19

References

- Bootstrapping Pay-As-You-Go Data Integration Systems by Anish D. Sarma, Xin Dong, Alon Halevy, Proceedings of SIGMOD'08, Vancouver, British Columbia, Canada, June 2008
- Daisy Zhe Wang, Xin Luna Dong, Anish Das Sarma, Michael J. Franklin, Alon Y. Halevy: Functional Dependency Generation and Applications in Pay-As-You-Go Data Integration Systems. WebDB 2009
- Naser Ayat, Hamideh Afsarmanesh, Reza Akbarinia, Patrick Valduriez: Pay-As-You-Go Data Integration Using Functional Dependencies. CD-ARES 2012: 375-389
- William Kokou Dedzoe, Philippe Lamarre, Reza Akbarinia, Patrick Valduriez: As-Soon-As-Possible Top-k Query Processing in P2P Systems. Trans. Large Scale Data Knowl. Centered Syst. 9: 1-27 (2013)
- Ruiming Tang, Dongxu Shao, Stéphane Bressan, Patrick Valduriez: What You Pay for Is What You Get. DEXA (2) 2013: 395-409
- Fábio Porto, Amir Khatibi, João N. Rittmeyer, Eduardo S. Ogasawara, Patrick Valduriez, Dennis E. Shasha: Constellation Queries over Big Data. SBBD 2018: 85-96
- Sakina Mahboubi, Reza Akbarinia, Patrick Valduriez: Privacy-Preserving Top-k Query Processing in Distributed Systems. Euro-Par 2018: 281-292

20