

Functional Dependency Generation and Applications in Pay-As-You-Go Data Integration Systems

Daisy Zhe Wang, Luna Dong, Anish Das Sarma,
Michael J. Franklin, and Alon Halevy


UC Berkeley, AT&T Research,
Stanford University, and Google Inc.

Web-scale Structured Data


HTML Tables extracted from the Web

EnchantedLearning.com

The Presidents of the United States of America



President's Day
[Activities](#)



Abraham
Lincoln

In the order in which they served

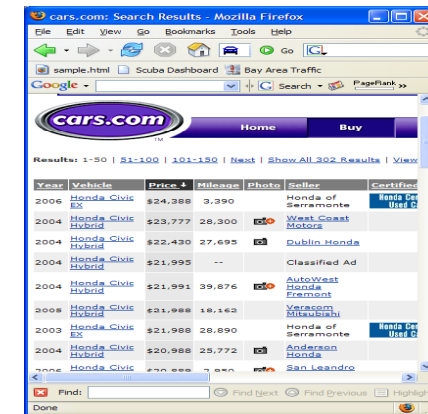
Alphabetical order

Short table of Data

The President and Vice-President are elected every four years. They must be at least 35 years of age, they must be native-born citizens of the United States, and they must have been residents of the U.S. for at least 14 years. (Also, a person cannot be elected to a third term as President.)

President	Party	Term as President	Vice-President
1. George Washington (1732-1799)	None, Federalist	1789-1797	John Adams
2. John Adams (1735-1826)	Federalist	1797-1801	Thomas Jefferson
3. Thomas Jefferson (1743-1826)	Democratic-Republican	1801-1809	Aaron Burr, George Clinton
4. James Madison (1751-1836)	Democratic-Republican	1809-1817	George Clinton, Elbridge Gerry
5. James Monroe (1758-1831)	Democratic-Republican	1817-1825	Daniel Tompkins

Database Views in the Deep Web accessed through HTML Forms on the Web



cars.com Search Results - Mozilla Firefox

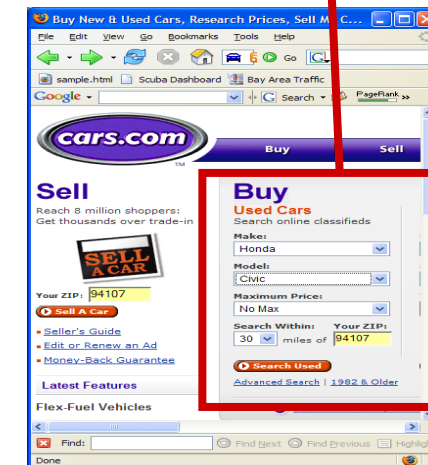
Year	Vehicle	Price	Mileage	Photos	Seller	Certified
2006	Honda Civic EX	\$24,288	3,390		Honda of Serramonte	Honda Certified
2004	Honda Civic Hybrid	\$23,777	28,300		West Coast Motors	
2004	Honda Civic Hybrid	\$22,430	27,695		Dublin Honda	
2004	Honda Civic Hybrid	\$21,995	--		Classified Ad	
2004	Honda Civic Hybrid	\$21,991	39,876		AutoWest Honda Fremont	
2008	Honda Civic Hybrid	\$31,988	18,163		Veracon Mitsubishi	
2003	Honda Civic EX	\$21,988	28,890		Honda of Serramonte	Honda Certified
2004	Honda Civic Hybrid	\$20,988	25,772		Anderson Honda	
2006	Honda Civic	\$20,988	2,850		San Leandro	

For years, [Microsoft Corporation CEO Bill Gates](#) was against open source. But today he appears to have changed his mind. "We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft VP](#). "That's a super-important shift for us in terms of code access."

[Richard Stallman](#), founder of the [Free Software Foundation](#), countered saying...

Relations generated by information extraction from web pages

Name	Title	Organization
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	Founder	Free Soft..



Buy New & Used Cars, Research Prices, Sell My Car

Reach 8 million shoppers: Get thousands over trade-in

Make:

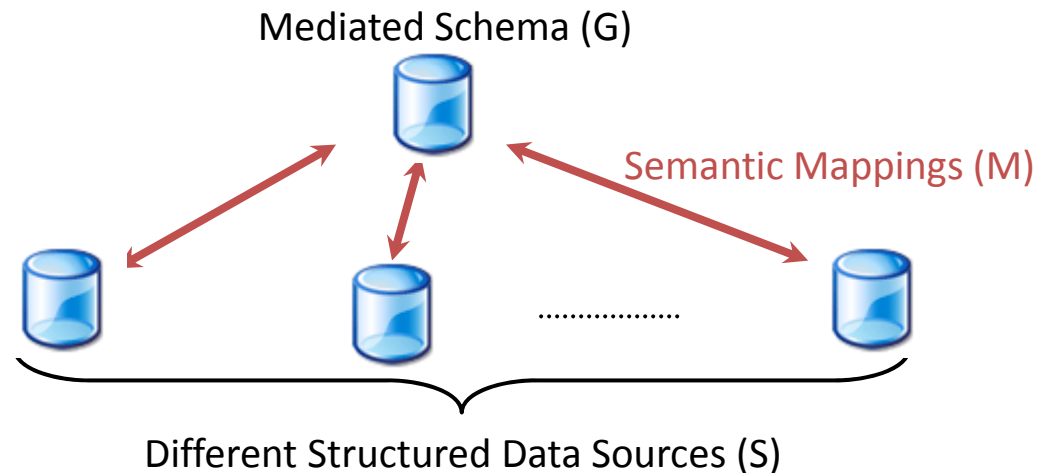
Model:

Maximum Price:

Search Within: miles of

Advanced Search | 1982 & Older

A Typical Data Integration System



- Data Sources (S): a set of data sources of a specific domain
- Mediated Schema (G): a set of relations and attributes that we wish to expose to users
- Schema Mapping (M): a set of mappings from the attributes in S to the attributes in G
- Query processing
 - A user query over G is reformulated into multiple queries over S using M
 - Results are retrieved from multiple data sources and combined

Data Integration at Web-scale

- A typical data integration solution is **impractical** for web-scale data
 - Too many domains of interest (Web Data is about everything)
 - Huge number of sources for each domain
 - Designing mediated schema is infeasible
 - Data sources are dirty, incomplete and lack of meta-data
- A web-scale data integration system
 - can only afford pay-as-you-go [Franklin et. al 2005]
 - Support automated schema design and mapping
 - Provide best-Effort services

Functional Dependency (FD)

- **Can we use FD theory in some way to automate the massive data integration problem?**
- FDs are specified top-down in the database design process as statements of truth on how attributes relates to each other
- FD $X \rightarrow Y$ holds if and only if each X value is associated with precisely one Y value
- One of Armstrong's Axioms for Normalization
Transitivity: if $X \rightarrow Y$, $Y \rightarrow Z$, then $X \rightarrow Z$

Probabilistic Functional Dependencies (pFDs)

- Why probabilistic FDs?
- Definition of a probabilistic FD (pFD)
 $X \rightarrow^p A$, p is the likelihood of FD holds in general
- Related work
 - TANE [Huhtala et al. 1999]
 - CORDS [Ilyas et al. 2004]
- The new challenge: from single large table to many, potentially incomplete and dirty tables

Generating pFDs

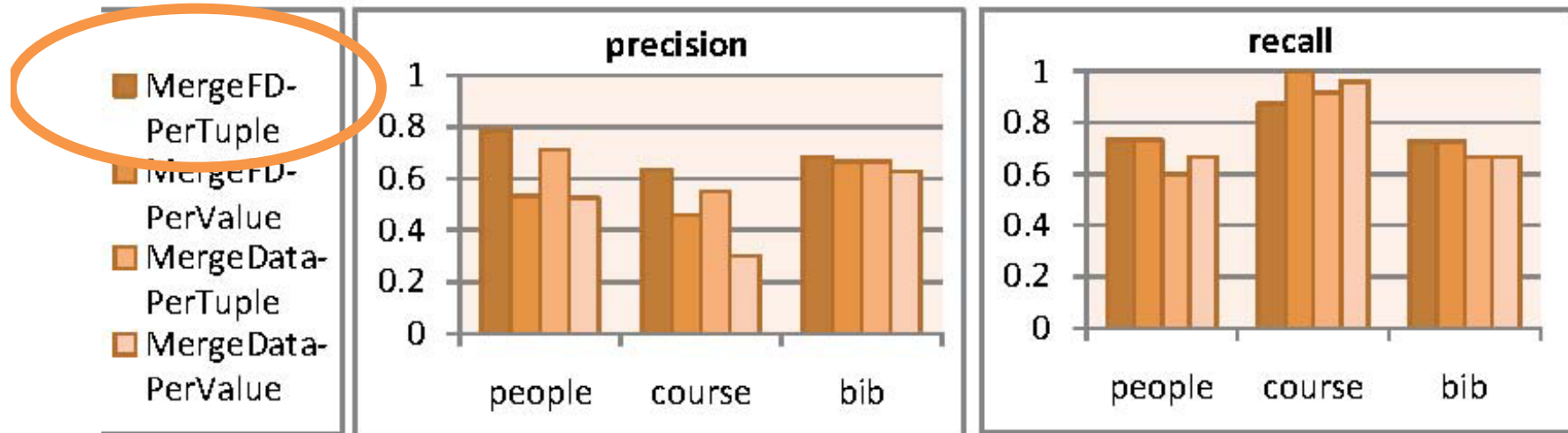
- Probability of pFD over single data source R

- Per-Tuple counting: $Pr(\bar{X} \rightarrow A, R)_{\text{PERTUPLE}} = \frac{\sum_{V_X \in \mathcal{D}_{\bar{X}}} |V_A, V_X|}{\sum_{V_X \in \mathcal{D}_{\bar{X}}} |V_X|}$
- Per-Value counting: $Pr(\bar{X} \rightarrow A, R)_{\text{PERVALUE}} = \frac{\sum_{V_X \in \mathcal{D}_{\bar{X}}} \frac{|V_A, V_X|}{|V_X|}}{|\mathcal{D}_{\bar{X}}|}$

- Probability of pFD over multiple data sources

- Merge pFDs: $Pr(\bar{X} \rightarrow A, \bar{S})_{\text{MERGEFD}} = \sum_{R \in \bar{S}} Pr(\bar{X} \rightarrow A, R)$
- Merge Data

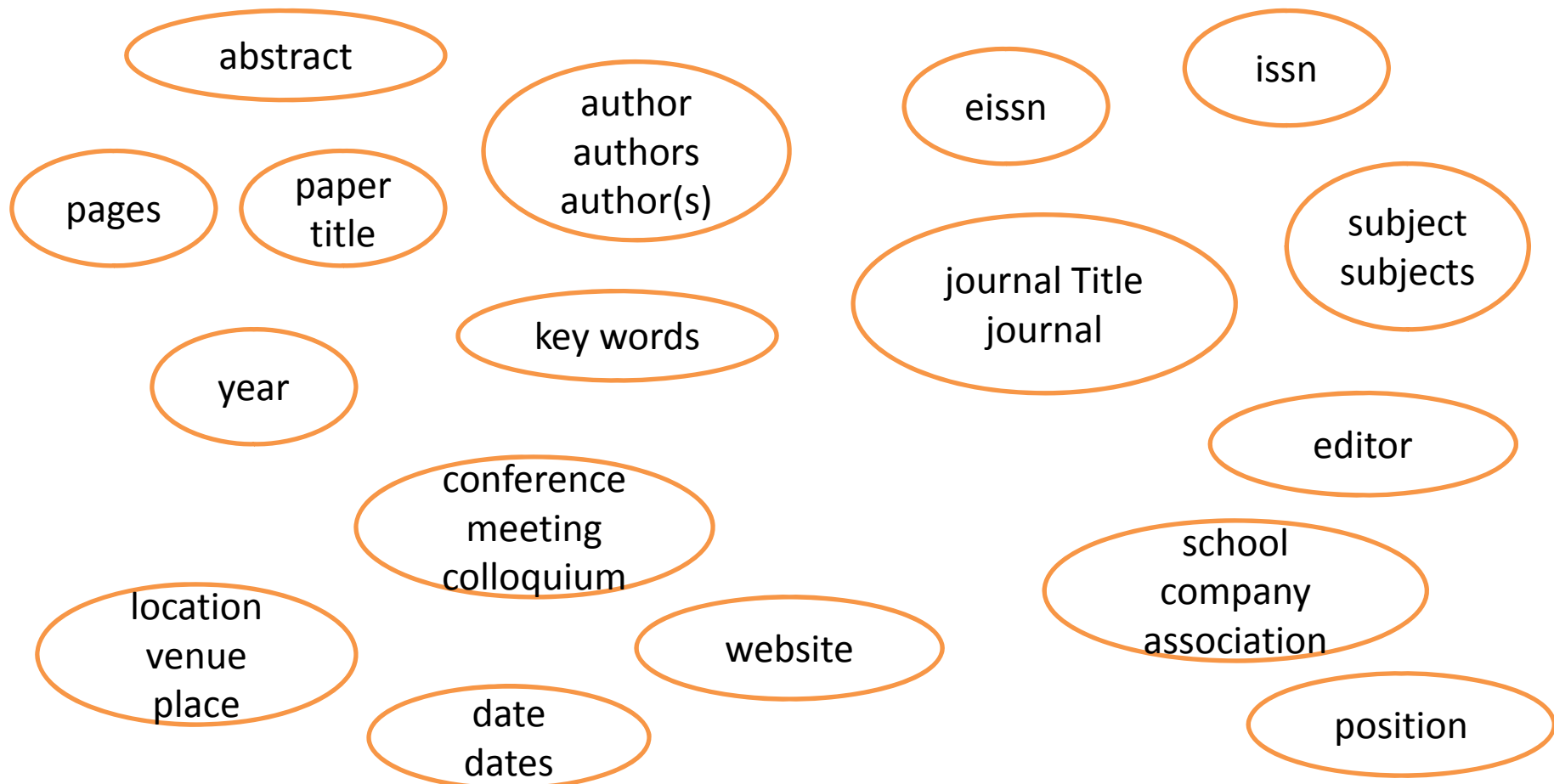
Results for pFDs Generation Algorithms



Number of data sources: 50 -- 600

App1: Normalize Mediated Schema Example (I)

Attributes in the mediated schema of the Bibliography Domain



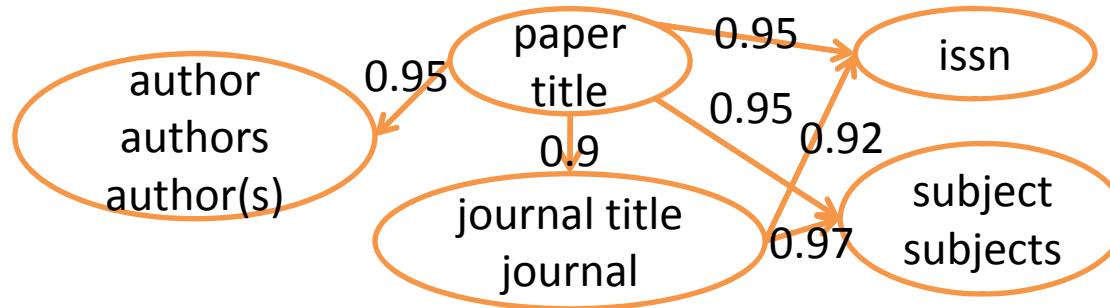
App1: Normalize Mediate Schema

Example (II)

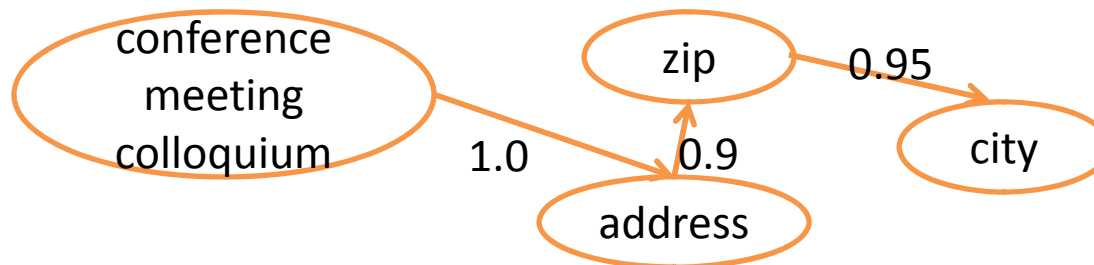


Normalizing Mediated Schema

- Prune pFD set
 - Prune low-probability pFDs
 - Prune pFDs that can be generated by transitivity



- Avoid over-splitting



Results for Schema Normalization

Domain	Mediated Schema	Normalized Schemas
People	(organization, name, country, work phone, zip, city address, fax, title, email, state)	Organization(<u>organization</u> , country, zip, city, address, fax) People(<u>name</u> , work phone, title, organization, email, state)
Course	(catalog number, class number, section, units, location, subject, fee, time, title, instructor days, institution, catalog number, term)	Course(<u>catalog number</u> , location, subject, units, class number, fee, institution) Class(<u>class number</u> , section, time, title, instructor, days, catalog number, term)
Bib	(journal title, title, id, subject, source, eissn authors, volumne, years, issue, issn)	Journal(<u>journal title</u> , issn, subject, source, eissn) Article(<u>title</u> , id, authors, volumne, years, issue)

Table 4: Results for mediated schema normalization application using pFDs.

App2: Identify Dirty Data Sources

- Structured data sources from the Web can be dirty

Dummy Values

name	company	email
Alice	IBM	email
Bob	Google	email
Cathy	Yahoo	email
David	MSR	email

Entity Ambiguity

name	country	city
Alice	USA	Boston
Bob	US	Boston
Cathy	u.s.a	Boston
David	United States	Boston

Nested Columns

name	city	country
Alice	Boston	02101,USA
Bob	Seattle	98101,USA
Cathy	Chicago	60601,USA
David	New York	12201,USA

- We report data sources that violate pFDs with high probabilities:

Results

People: 3 out of 3 reported are dirty

Course: 31 out of 80 reported are dirty (estimate total 66)

Bib: 3 out of 7 reported are dirty

Conclusion: FD in Pay-as-you-go

- Web-scale data integration can only afford to pay-as-you-go
- Automation is the key
 - Automatically setting up mediated schema, mapping
[Das Sarma et. al. 2008]
 - Automatically measuring and improve the quality of data integration
 - Measuring quality of data sources
 - Measuring and Improving quality of mediated schema, schema mapping, etc.
- FD-based Quality Measuring and Improvement
 - Identify dirty data sources
 - Improving mediated schemas

Future Work

- Automatic mediated schema design for millions of HTML tables
 - Domain cluster (clustering over source schema)
 - Entity/Relationship cluster (clustering + pFD normalization)
 - Attribute cluster (synonyms + string similarity)
- Related Issues
 - Scalability
 - Visualization