Best-Effort Data Integration

AnHai Doan
University of Wisconsin-Madison

Joint work with Fei Chen, Pedro DeRose, Robert McCann, Yoonkyong Lee, Mayssam Sayyadian, Warren Shen, Luis Gravano, Raghu Ramakrishnan
Data Integration: Current Status

- We have made tremendous progress in the last 30 years
  - develop foundations: mediator model, GAV, LAV
  - build on the foundation: query reformulation, provenance, uncertainty, schema matching and mapping, entity resolution, adaptive query processing, managing inconsistent data, P2P, etc.
  - branch into applications: bio-informatics, geo-spatial, Web, ...
  - join forces: databases, AI, Web

- But data integration remains hard
  - intractable, AI-complete, etc.

- Partly because we often want exact, precise integration
Precise Data Integration

Find houses with 4 bedrooms priced under 300K

- Original motivation: business applications
  - e.g., payroll, human resources, banking
  - here, anything less is NOT usable
However, the Application Landscape Has Changed in the Past Decade

- **Today**, precise integration continues to be critical
  - e.g., expedia.com

- **But for many emerging application domains**, best-effort data integration
  - often incurs far less cost
  - may already prove very useful

- **Examples**
  - citation tracking (e.g., Citeseer, Google Scholar)
  - personal information management
  - scientific, exploratory data analysis
  - intelligence analysis for homeland security
  - business intelligence
  - Web integration scenarios (e.g., Froogle)
Best-Effort Data Integration

- Remove, simplify, or make “less precise” certain components
- Employ automatic techniques
- To go “the last mile”: learn from human interaction
Example 1: Simplify Global Schema

Keyword Search over Multiple Databases

- Novel problem
- Very useful for urgent / one-time DI needs
  - also when users are SQL-illiterate
- Proposed solution in ICDE-07a
  - combines IR, schema matching, entity resolution, and AI planning

Can do joins across data sources
Example 2: Simplify Wrappers
Structured Queries over Text/Web Data

- Novel problem
- Proposed solution in ICDE-07b
Example 3: Best-Effort Data Integration for Web Communities

- Numerous data-rich communities
  - database researchers, movie fans, legal professionals, bioinformatics, enterprise intranets, etc.
- Each community = many disparate data sources + people
- Members often want to discovery, query, monitor information in the community
  - any interesting connection between researchers X and Y?
  - find all citations of this paper in the past one week on the Web
  - what is new in the past 24 hours in the database community?
  - what are current hot topics? who has moved where?
The Cimple Project @ Wisconsin/Yahoo!

Builds structured data portals using
extraction + integration + mass collaboration

Keywords:
- Personalize system, provide feedback
- Keyword search
- SQL querying
- Question answering
- Browse
- Mining
- Alert/Monitor
- News summary

Researchers:
- Jim Gray

Web pages
Text documents
Jim Gray
Jim Gray
SIGMOD-04
give-talk
SIGMOD-04
Prototype System: DBLife

Integrate data of the DB research community

1164 data sources

Crawled daily, 11000+ pages = 160+ MB / day
Data Integration

Selected Publications
- Raghu Ramakrishnan, J. Valduriez, J. M. Naughton, (EDM 90 Tutorial).

Raghu Ramakrishnan

co-authors = A. Doan, Divesh Srivastava, ...
Resulting ER Graph
Querying The ER Graph

Query: “David DeWitt Jennifer Widom”

1. Jennifer Widom - coauthor of David DeWitt

2. Jennifer Widom - PC-Chair of SIGMOD 2005
   - SIGMOD 2005 - PC-member of Jennifer Widom

   - Shivnath Babu - coauthor of Jennifer Widom

Database Bibliographies by Topic

- active databases, constraint management
- applications and middleware
- approximation and uncertainty
- architecture, engines, and internals
- change management, maintenance
- data cleaning, data translation, data exchange, schema matching, record linkage
- data integration, heterogeneous database systems, interoperability
- data mining, classification, clustering
- data models, query languages, design analysis
- data reduction, compression, sampling
- data replication
- data storage, indexing, and access methods
- data warehousing and olap, decision support
- deductive databases, datalog
- derived data and materialized views
- extensibility and database evolution
Mass Collaboration: A Simplified Example

Picture is removed if enough users vote “no”.

16
More on Mass Collaboration
DBLife: Key Lessons Learned

- Built relatively simple best-effort integration tools
- Combined them in a flexible, bottom-up fashion
- System appears already interesting/useful
  - see dblife.cs.wisc.edu (still very preliminary & slow)
- Hence possible strategy for best-effort integration:
  - build relatively simple integration tools
  - learn how to combine them effectively
- Relative simple integration tools = Lego blocks
  - easier to build, debug, work with, enable quick tech transfer?
- Building systems bring much benefits
  - suggests many interesting / unexpected research challenges
  - helps bridge the research/tech transfer gap
Research Challenges (1)

- Information extraction
- Data integration
- Mass collaboration
  - how to collectively edit extracted and integrated data?
Research Challenges (2)

- **Exploiting extracted data**
  - keyword search, structured querying, mining, monitoring
  - how to seamlessly transition among these?
- **Handling uncertainty / provenance / explanation**
- **Dealing with evolving data**

Personalize system, provide feedback

Keyword search
SQL querying
Question answering
Browse
Mining
Alert/Monitor
News summary
Research Challenges (3)

- New data model?
- Should we use / extend relational databases?
- How to build continuously running systems?

Personalize system, provide feedback
Summary

- Precise vs. best-effort data integration
- Sample research
  - keyword search over multiple databases
  - SQL queries over text
  - Cimple project @ Wisconsin/Yahoo! Research
- The topic is wide open
- Our community can contribute much
- Prototype system: DBlife
  - can serve as a data integration challenge / testbed / benchmark
  - potentially provides useful service to our community (as DBWorld+)
  - provides data for researchers (on a variety of topics)

More details: search “anhai cimple”
Mass Collaboration Meets Jeff Naughton

Jeffrey F. Naughton swears that this is David J. DeWitt