DATA FUSION – RESOLVING DATA CONFLICTS IN INTEGRATION

Tutorial at VLDB 2009

Xin Luna Dong – AT&T Labs-Research
Felix Naumann – Hasso Plattner Institute (HPI)
Origins of Data Conflicts

ACM Computing Survey [BN08]
Origins of Data Conflicts

Integrated data

Schering CRM

Bayer CRM

Data Fusion | VLDB 2009 Tutorial | Luna Dong & Felix Naumann
Search Results for 'dessloch'

- Stefan Deßloch
- Stefan Dessloch

DBLP: [Home | Search: Author, Title | Conferences | Journals]
Michael Levy (ley@uni-trier.de) Thu Jan 31 10:44:06 2008
| 488941 britney spears | 29 britent spears | 9 brittany spears | 5 brney spears | 3 britiy spears |
| 40134 brittany spears | 29 brittnany spears | 9 brittany spears | 5 britney spears | 3 britney spears |
| 36315 brittney spears | 29 brittnany spears | 9 brittany spears | 5 britney spears | 3 britney spears |
| 24342 brittney spears | 29 britney spears | 9 britney spears | 5 britney spears | 3 britney spears |
| 7331 brittny spears | 26 birttny spears | 9 brittny spears | 5 btrtney spears | 3 britteny spears |
| 6633 brittney spears | 26 britney spears | 9 britney spears | 5 britney spears | 3 britney spears |
| 2696 brittney spears | 26 britney spears | 9 britney spears | 5 britney spears | 3 britney spears |
| 1807 briney spears | 26 britteny spears | 9 brittney spears | 5 brittney spears | 3 brittney spears |
| 1635 brittny spears | 26 brittny spears | 9 brittney spears | 5 brittney spears | 3 brittney spears |
| 1479 britteny spears | 26 brittany spears | 9 brittney spears | 5 brittney spears | 3 brittney spears |
| 1479 brittany spears | 26 brittany spears | 9 brittney spears | 5 brittney spears | 3 brittney spears |
| 1338 britnny spears | 26 brittny spears | 9 brittney spears | 5 brittney spears | 3 brittney spears |
| 1211 britneet spears | 24 birtney spears | 8 brittney spears | 4 britteyn spears | 3 brittney spears |
| 1096 britney spears | 24 birtney spears | 8 britney spears | 4 britteny spears | 3 britteny spears |
| 991 britney spears | 24 britteny spears | 8 brittney spears | 4 britteny spears | 3 britteny spears |
| 811 brittney spears | 24 brittny spears | 8 brittney spears | 4 britteny spears | 3 britteny spears |
| 811 britney spears | 24 brittny spears | 8 brittney spears | 4 britteny spears | 3 britteny spears |
| 664 britney spears | 24 brittny spears | 8 brittney spears | 4 britteny spears | 3 britteny spears |
| 664 brittny spears | 24 brittney spears | 8 brittney spears | 4 britteny spears | 3 britteny spears |
| 664 britney spears | 24 brittney spears | 8 brittney spears | 4 britteny spears | 3 britteny spears |
| 601 britney spears | 24 brittny spears | 8 brittney spears | 4 britteny spears | 3 britteny spears |
| 601 britny spears | 24 brittny spears | 8 brittney spears | 4 britteny spears | 3 britteny spears |
| 544 brittney spears | 21 biritney spears | 7 brittney spears | 4 britney spears | 2 britney spears |
| 544 brittney spears | 21 biritney spears | 7 brittney spears | 4 britney spears | 2 britney spears |
| 364 britney spears | 21 britney spears | 7 britney spears | 4 britney spears | 2 britney spears |
| 364 brittiny spears | 21 brittny spears | 7 brittny spears | 4 brittny spears | 2 brittny spears |
| 329 britney spears | 21 britnie spears | 7 britney spears | 4 britney spears | 2 britney spears |
| 269 britney spears | 21 britney spears | 7 britney spears | 4 britney spears | 2 britney spears |
| 269 britneys spears | 21 brittay spears | 7 britney spears | 4 britney spears | 2 britney spears |
| 244 britne spears | 21 brittinay spears | 7 britney spears | 4 britney spears | 2 britney spears |
| 244 brytney spears | 21 britney spears | 7 britney spears | 4 britney spears | 2 britney spears |
| 220 breetney spears | 21 britney spears | 7 britney spears | 4 britney spears | 2 britney spears |
| 220 brittanny spears | 19 birney spears | 7 brittney spears | 4 brittney spears | 2 brittney spears |
| 195 britnny spears | 19 britney spears | 7 brittney spears | 4 brittney spears | 2 brittney spears |
| 163 britnry spears | 19 britney spears | 7 brittney spears | 4 brittney spears | 2 brittney spears |
| 159 brnney spears | 19 britney spears | 7 brittney spears | 4 brittney spears | 2 brittney spears |
| 157 britny spears | 19 britney spears | 7 brittney spears | 4 brittney spears | 2 brittney spears |
| 153 britnry spears | 19 britney spears | 7 brittney spears | 4 brittney spears | 2 brittney spears |
| 140 britney spears | 19 britney spears | 7 brittney spears | 4 brittney spears | 2 brittney spears |
| 140 brittny spears | 19 brittny spears | 7 brittny spears | 4 brittny spears | 2 brittny spears |
| 128 britney spears | 19 britney spears | 7 brittney spears | 4 brittney spears | 2 brittney spears |
| 128 brittny spears | 19 brittny spears | 7 brittny spears | 4 brittny spears | 2 brittny spears |
Origins of Intra-Source Conflicts

- No integrity or consistency checks
- Redundant schemata
- Typos, transmission errors, incorrect calculations
- Variants
  - Kantstr. / Kantstrasse / Kant Str. / Kant Strasse
  - Kolmogorov / Kolmogoroff / Kolmogorow
- Typical confusion (OCR)
  - U<->V, 0<->o, 1<->l, etc.
- Obsolete values
  - Different update frequencies, forgotten update
Origins of Inter-Source Conflicts

- Locally consistent but globally inconsistent
- Different data types
- Local spelling variations and conventions
  - Addresses
    - St → Street, Ave → Avenue, etc.
    - R.-Breitscheid-Str. 72 a → Rudolf-Breitscheid.-Str. 72A
    - 128 spellings for Frankfurt am Main
  - Names
    - Dr. Ing. h.c. F. Porsche AG
    - Hewlett-Packard Development Company, L.P.
  - Numerical data
    - 10.000 € = 10T EURO = 10k EUR = 10.000,00€ = 10,000.- €
  - Phone numbers, birth dates, etc.
“... focus is on fusing data management and collaboration: merging multiple data sources, discussion of the data, querying, visualization, and Web publishing.”

“The power of data is truly harnessed when you combine data from multiple sources. Fusion Tables enables you to fuse multiple sets of data when they are about the same entities. In database speak, we call this a join on a primary key but the data originates from multiple independent sources.”
Web Integration—Google Fusion Tables

- Allows discussion of values between users
- Error correction
  - Reference tables
    - Cities, countries, products ...
    - Similarity measures
  - Standardization and transformation
  - Domain-knowledge (meta data)
    - Conventions (country/region-specific spelling)
    - Ontologies
    - Thesauri, dictionaries for homonyms, synonyms, ...
  - Outlier detection and elimination
- And data fusion...
Overview

- Data fusion in the integration process
- Foundations of data fusion
  - Conflict resolution strategies and functions
  - Conflict resolution operators
- Advanced truth-discovery techniques
- Existing data fusion systems
- Open problems
Information Integration

Source A

<pub>
  <Titel> Federated Database Systems </Titel>
  <Autoren>
    <Autor> Amit Sheth </Autor>
    <Autor> James Larson </Autor>
  </Autoren>
</pub>

Source B

<publication>
  <title> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </title>
  <author> Sheth & Larson </author>
  <year> 1990 </year>
</publication>

- Schema Mapping
- Data Transformation
- Duplicate Detection
- Data Fusion

Data Fusion | VLDB 2009 Tutorial | Luna Dong & Felix Naumann
Information Integration

Source A

Source B

Schema Mapping

Data Transformation

Duplicate Detection

Data Fusion

Schema Integration

Federated Database Systems

Amit Sheth

James Larson

1990

Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases

Amit Sheth & Larson

1990

Duplicate Detection
Information Integration

Source A

<pub>
<Titel> Federated Database Systems </Titel>
<Autoren>
<Autor> Amit Sheth </Autor>
<Autor> James Larson </Autor>
</Autoren>
</pub>

Source B

<publication>
<title> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </title>
<Autoren>
<author> Scheth & Larson </author>
</Autoren>
<year> 1990 </year>
</publication>

Transformation queries or views

<pub>
<title> Federated Database Systems </title>
<Autoren>
<Autor> Amit Sheth </Autor>
<Autor> James Larson </Autor>
</Autoren>
</pub>

<pub>
<title> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </title>
<Autoren>
<Autor> Scheth & Larson </Autor>
</Autoren>
<year> 1990 </year>
</pub>
Information Integration

Source A

Source B

Schema Mapping

Data Transformation

Duplicate Detection

Data Fusion
Information Integration

Source A

Source B

Schema Mapping

Data Transformation

Duplicate Detection

Data Fusion

<publication>
  <title> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </title>
  <author> Scheth & Larson </author>
  <year> 1990 </year>
</publication>

<pub>
  <title> Federated Database Systems </title>
  <author> Amit Sheth & Larson </author>
  <year> 1990 </year>
</pub>

<pub>
  <title> Federated Database Systems </title>
  <author> Amit Sheth & Larson </author>
  <year> 1990 </year>
</pub>

<pub>
  <title> Federated Database Systems </title>
  <author> Amit Sheth & Larson </author>
  <year> 1990 </year>
</pub>
Information Integration

Source A

Federated Database Systems

<pub>
<title>Federated Database Systems</title>
<Autoren>
<author>Amit Sheth</author>
<author>James Larson</author>
</Autoren>
</pub>

Source B

Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases

<pub>
<title>Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases</title>
<Autoren>
<author>Amit Sheth</author>
<author>James Larson</author>
</Autoren>
<year>1990</year>
</pub>

Preserve lineage

Schema Mapping
Data Transformation
Duplicate Detection
Data Fusion
Completeness, Conciseness, and Correctness

Schema Matching:
Same attribute semantics
Duplicate detection:
Same real-world entities
Completeness, Conciseness, and Correctness

Data Fusion: Resolve uncertainties and contradictions

Intensional completeness

Extensional completeness

Extensional conciseness

Intensional conciseness
Problem
- Given two schemata, find all correspondences between their attributes

Difficulties
- Schematic heterogeneity (synonyms & homonyms)
- Data heterogeneity
- n:m mappings
- Transformation functions
- User interaction

Then: Derive a schema mapping
Duplicate Detection

- **Problem**
  - Given one or more data sets, find all sets of objects that represent the same real-world entity.

- **Difficulties**
  - Duplicates are not identical
    - Similarity measures – Levenshtein, Soundex, Jaccard, etc.
  - Large volume, cannot compare all pairs
    - Partitioning strategies – Sorted neighborhood, Blocking, etc.
Ironically, “Duplicate Detection” has many Duplicates

- Household matching
- Mixed and split citation problem
- Match
- Deduplication
- Entity resolution
- Identity uncertainty
- Hardening soft databases
- Doubles
- Duplicate detection
- Record linkage
- Object identification
- Object consolidation
- Entity clustering
- Approximate match
- Reference reconciliation
- Merge/purge
- Householding
- Reference matching
Problem

- Given a duplicate, create a single object representation while resolving conflicting data values.

Difficulties

- Null values: Subsumption and complementation
- Contradictions in data values
- Uncertainty & truth: Discover the true value and model uncertainty in this process
- Metadata: Preferences, recency, correctness
- Lineage: Keep original values and their origin
- Implementation in DBMS: SQL, extended SQL, UDFs, etc.
The Field of Data Fusion

Data Fusion

Conflict types
- Uncertainty
- Contradiction
- Ignorance
- Avoidance
- Resolution

Resolution strategies
- Join-based
- Union-based
- Possible worlds
- Consistent answers

Operators
- Instance-based
- Metadata-based

Resolution functions
- Subsumption
- Complementation
- Aggregation
- Advanced functions
Overview

- Data fusion in the integration process
- Foundations of data fusion
  - Conflict resolution strategies and functions
  - Conflict resolution operators
- Advanced truth-discovery techniques
- Existing data fusion systems
- Open problems
Uncertainty and Contradiction

- **Uncertainty**
  - NULL value vs. non-NUL value
  - “Easy” case

- **Contradiction**
  - Non-NUL value vs. (different) non-NUL value
Semantics of NULL

- “unknown”
  - There is a value, but I do not know it.
  - E.g.: Unknown date-of-birth

- “not applicable”
  - There is no meaningful value.
  - E.g.: Spouse for singles

- “withheld”
  - There is a value, but we are not authorized to see it.
  - E.g.: Private phone line
Classification of Strategies

conflict resolution strategies

- conflict resolution
  - conflict ignorance
    - instance based: PASS IT ON
  - conflict avoidance
    - instance based: TAKE THE INFORMATION NO GOSSIPING
  - conflict resolution
    - instance based: CRY WITH THE WOLVES
    - mediating: MEET IN THE MIDDLE
    - metadata based: NOTHING IS OLDER THAN THE NEWS FROM YESTERDAY
    - deciding: ROLL THE DICE
    - mediating: TRUST YOUR FRIENDS
## Conflict Resolution Functions

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min, Max, Sum, Count, Avg</td>
<td>Standard aggregation</td>
<td>NumChildren, Salary, Height</td>
</tr>
<tr>
<td>Random</td>
<td>Random choice</td>
<td>Shoe size</td>
</tr>
<tr>
<td>Longest, Shortest</td>
<td>Longest/shortest value</td>
<td>First_name</td>
</tr>
<tr>
<td>Choose(source)</td>
<td>Value from a particular source</td>
<td>DoB (DMV), CEO (SEC)</td>
</tr>
<tr>
<td>ChooseDepending(val, col)</td>
<td>Value depends on value chosen in other column</td>
<td>city &amp; zip, e-mail &amp; employer</td>
</tr>
<tr>
<td>Vote</td>
<td>Majority decision</td>
<td>Rating</td>
</tr>
<tr>
<td>Coalesce</td>
<td>First non-null value</td>
<td>First_name</td>
</tr>
<tr>
<td>Group, Concat</td>
<td>Group or concatenate all values</td>
<td>Book_reviews</td>
</tr>
<tr>
<td>MostRecent</td>
<td>Most recent (up-to-date) value</td>
<td>Address</td>
</tr>
<tr>
<td>MostAbstract, MostSpecific, CommonAncestor</td>
<td>Use a taxonomy / ontology</td>
<td>Location</td>
</tr>
<tr>
<td>Escalate</td>
<td>Export conflicting values</td>
<td>gender</td>
</tr>
</tbody>
</table>

...
Classification of Functions

Conflict resolution strategies

- Conflict ignorance
  - Escalate

- Conflict avoidance

- Conflict resolution
  - Instance based
    - Coalesce
    - Choose Depending
    - Concat
  - Metadata based
    - Choose
      - Deciding
        - MIN, MAX
        - Random
        - Vote
      - Mediating
        - AVG, SUM
  - Instance based
    - Deciding
      - MostRecent
      - MostAbstract
      - MostSpecific
    - Mediating
      - CommonAncestor
Mehrfach vorhandener Kontakt

Der Name oder die E-Mail-Adresse dieses Kontakts ist bereits im Ordner 'Kontakte' vorhanden. Möchten Sie:

- Neuen Kontakt hinzufügen
- Informationen des ausgewählten Kontakts aktualisieren. Eine Sicherungskopie wird im Ordner 'Gelöschte Objekte' gespeichert.

Name  Position  Firma  E-Mail
Marina Mustermann  CEO  Acme Corp.  marina@acme.com

Vorschau der aktualisierten Visitenkarte:

Mustermann, Marina
Acme Corporation
CEO
+1493315509280 Geschäftlich
mmustermann@yahoo.com
marina@acme.com

Änderungen am ausgewählten Kontakt:

Name: Mustermann, Marina
Position: CEO
Firma: Acme Corporation
E-Mail: mmustermann@yahoo.com
E-Mail: marina@acme.com
Telefon geschäftlich: +1493315509280 +1 (49) 331 5509 287
Fax geschäftl.: +1 (49) 331 5509 287
Kontaktbild: Keine Änderung
Notizen: Keine Änderung
Overview

- Data fusion in the integration process
- Foundations of data fusion
  - Conflict resolution strategies and functions
  - Conflict resolution operators
- Advanced truth-discovery techniques
- Existing data fusion systems
- Open problems
Data Fusion Goals

Identical tuples

Source 1(A,B,C) →

\[ a, b, c \]

\[ a, b, - \]

Source 2(A,B,D) →

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

Identical tuples

Subsumed tuples

\[ a, b, c \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]

\[ a, b, - \]
Relational Operators – Overview

- Identical tuples
  - UNION, OUTER UNION

- Subsumed tuples (uncertainty)
  - MINIMUM UNION

- Complementing tuples (uncertainty)
  - COMPLEMENT UNION, MERGE

- Conflicting tuples (contradiction)
  - Relational approaches: Match, Group, Fuse, ...

- Other approaches
  - Possible worlds, probabilistic answers, consistent answers
Minimum Union

- **Union**: Elimination of exact duplicates

- **Minimum Union**: Elimination of subsumed tuples
  - Outer union
  - Subsumption

- **Rewriting in SQL using DWH extensions** (Windows) and assuming existence of favorable ordering [RPZ04]

### A B C | A B D
---|---
a b c | a b ⊥
e f g | e f h
m n o | m p ⊥

A tuple $t_1$ subsumes a tuple $t_2$, if it has the same schema, has less NULL-values, and coincides in all non-NULL-values.

### A B C D
---
a b c ⊥
e f g ⊥
e f h
m n o ⊥
m p ⊥ ⊥
Full Disjunction

- Represents all possible combinations of source tuples
  - Full outer join on all common attributes
    - All combinations for more than two sources
  - Minimum union over results
- Combines complementing tuples (only inter-source)
- Algorithms: [GL94, RU96, CS05]
Elimination of complementing tuples

- Outer union
- Complementation

No known SQL rewriting

A tuple $t_1$ complements a tuple $t_2$, if it has the same schema and coincides in all non-NULL-values.

Includes duplicate removal and subsumption.
Mixes Join and Union to a new operator [GPZ01]

- Idea: Build two versions for each common attribute, one “favoring” S1, the other “favoring” S2.
- Nulls in a source are replaced using COALESCE.
- Fuses complementing tuples, but only for inter-source duplicates
- Priorization possible: Removes conflicting tuples from right relation.

```sql
(SELECT R.A, COALESCE(R.B, S.B), R.C, S.D
 FROM R LEFT OUTER JOIN S ON R.A = S.A )
UNION
(SELECT S.A, COALESCE(S.B, R.B), R.C, S.D
 FROM R RIGHT OUTER JOIN S ON R.A = S.A )
```
Merge and Prioritized Merge

A is real-world ID

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>e</td>
<td>f</td>
<td>g</td>
</tr>
<tr>
<td>m</td>
<td>n</td>
<td>o</td>
</tr>
<tr>
<td>m</td>
<td>n</td>
<td>↓</td>
</tr>
<tr>
<td>q</td>
<td>r</td>
<td>s</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>↓</td>
</tr>
<tr>
<td>e</td>
<td>f</td>
<td>h</td>
</tr>
<tr>
<td>m</td>
<td>p</td>
<td>↓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>COAL(b,b)</td>
<td>c</td>
<td>↓</td>
</tr>
<tr>
<td>e</td>
<td>COAL(f,f)</td>
<td>g</td>
<td>h</td>
</tr>
<tr>
<td>m</td>
<td>COAL(n,p)</td>
<td>o</td>
<td>↓</td>
</tr>
<tr>
<td>m</td>
<td>COAL(n,p)</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>q</td>
<td>r</td>
<td>s</td>
<td>↓</td>
</tr>
</tbody>
</table>

|=| |=|

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>c</td>
<td>↓</td>
</tr>
<tr>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
</tr>
<tr>
<td>m</td>
<td>n</td>
<td>o</td>
<td>↓</td>
</tr>
<tr>
<td>m</td>
<td>n</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>m</td>
<td>n</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>q</td>
<td>r</td>
<td>s</td>
<td>↓</td>
</tr>
</tbody>
</table>

|=| |=|

Data Fusion | VLDB 2009 Tutorial | Luna Dong & Felix Naumann
Match Join

- Context: AURORA Project [YÖ99]
- Handles columns individually using projections (with IDs)
- Performs UNION on each column across all sources
- Reassembles using FULL OUTER JOINS
- Uses “conflict-tolerant query model” to query these possible worlds.

```sql
WITH OU(A,B,C,D) AS (  
    ( SELECT A, B, C, NULL AS D FROM U1 )
    UNION
    ( SELECT A, B, NULL AS C, D FROM U2 ) ),
B_V (A,B) AS ( SELECT DISTINCT A, B FROM OU ),
C_V (A,C) AS ( SELECT DISTINCT A, C FROM OU ),
D_V (A,D) AS ( SELECT DISTINCT A, D FROM OU ),
SELECT A, B, C, D
FROM B_V FULL OUTER JOIN C_V FULL OUTER JOIN D_V
ON B_V.A=C_V.A AND C_V.A=D_V.A
```
Match Join

- Conflict-tolerant query model
  - Chooses tuples from result of MatchJoin

- Three semantics
  - HighConfidence, RandomEvidence, PossibleAtAll

- Resolution functions
  - SUM, AVG, MAX, MIN, ANY, DISCARD

```
SELECT ID, Name[ANY], Age[MAX]
FROM MatchJoin(U1,U2)
WHERE Age>22
WITH PossibleAtAll
```
Grouping and Aggregation

- Outer union then group by real-world ID
- Aggregate all other columns using conflict resolving aggregate function
- Efficient implementations
- Catches inter- and intra-source duplicates
- Restricted to built-in aggregate-functions
  - MAX, MIN, AVG, VAR, STDDEV, SUM, COUNT

```sql
WITH OU AS (
    ( SELECT A, B, C, NULL AS D FROM U1 )
    UNION (ALL)
    ( SELECT A, B, NULL AS C, D FROM U2 )
),
SELECT A, MAX(B), MIN(C), SUM(D)
FROM OU
GROUP BY A
```
SQL extensions to resolve uncertainties and contradictions [BN05,BBB+05]

FUSE FROM implies OUTER UNION
- Removes subsumed and duplicate tuples by default

FUSE BY declares real-world ID

RESOLVE specifies conflict resolution function from catalog
- Default: COALESCE

Implemented on top of relational DBMS “XXL”

```
SELECT ID,
    RESOLVE(Title, Choose(IMDB)),
    RESOLVE(Year, Max), RESOLVE(Director),
    RESOLVE(Rating), RESOLVE(Genre, Concat)
FUSE FROM IMDB, Filmdienst
FUSE BY (ID)
ON ORDER Year DESC
```
## Summary of Operators

<table>
<thead>
<tr>
<th></th>
<th>Duplicates</th>
<th>Subsumed tuples</th>
<th>Complementing tuples</th>
<th>Contradictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Union, Outer Union</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Minimum Union</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Full Disjunction</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>(inter-source)</td>
</tr>
<tr>
<td>Complement Union</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Merge</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>(inter-source)</td>
</tr>
<tr>
<td>MatchJoin + CTQM</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Group By</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Fuse By</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
FuSem

- Tool to query and fuse data from diverse data sources [BDN07]
  - Based on HumMer project [BBB+05].
  - http://www.hpi.uni-potsdam.de/naumann/sites/fusem/
- Explore data and find interesting subsets
- Execute, explore and compare five different data fusion semantics, specified in their respective syntax:
  - SQL (and extensions, such as Subsumption)
  - Merge
  - MatchJoin
  - FuseBy
  - ConQuer
What else is there?

- **Consistent Query Answering**
  - Avoid conflicts and report only certain tuples
    - Those that appear in every repair [FFM05]

- **“Possible worlds” models**
  - Build all possible solutions, annotated with likelihood
    - Yes/No/Maybe [DeM89]
    - Probability value [LSS94]

- **Probabilistic databases [SD05]**
  - Extend algebra to produce probabilities
  - Extend query language to query and export probabilities
Overview

- Data fusion in the integration process
- Foundations of data fusion
  - Conflict resolution strategies and functions
  - Conflict resolution operators
- Advanced truth-discovery techniques
- Existing data fusion systems
- Open problems
Outline

- Data fusion in the integration process
- Foundations of data fusion
  - Conflict resolution strategies and functions
  - Conflict resolution operators
- Advanced truth-discovery techniques
- Data fusion in existing integration systems
- Open problems
Basic Strategies

conflict resolution strategies

- conflict ignorance
  - instance based: pass it on
  - metadata based: trust your friends
- conflict avoidance
  - instance based: take the information no gossiping
  - metadata based: cry with the wolves roll the dice
- conflict resolution
  - deciding: nothing is older than the news from yesterday
  - mediating: meet in the middle
Data sources are of different quality and we trust data from accurate sources more.
Intuitions

- Data sources are of different quality and we trust data from accurate sources more.
Basic Strategies

conflict resolution strategies

conflict ignorance

- instance based
  - deciding
  - mediating

conflict avoidance

- metadata based
  - trusting
  - meeting

conflict resolution

- instance based
  - deciding
  - mediating

- metadata based
  - trusting
  - meeting
Intuitions

- Data sources are of different quality and we trust data from accurate sources more
- The real world is dynamic and the true value often evolves over time
  - E.g., person affiliation, business contact phone
Data sources are of different quality and we trust data from accurate sources more.

The real world is dynamic and the true value often evolves over time.
- E.g., person affiliation, business contact phone.

Data sources can copy from each other and errors can be propagated quickly.
Data sources are of different quality and we trust data from accurate sources more.

The real world is dynamic and the true value often evolves over time.
- E.g., person affiliation, business contact phone.

Data sources can copy from each other and errors can be propagated quickly.

Consider accuracy of sources.

Consider freshness of sources.

Consider dependence between sources.
Advanced Truth-Discovery Techniques

- Data sources are of different quality and we trust data from accurate sources more.
- The real world is dynamic and the true value often evolves over time. E.g., person affiliation, business contact phone.
- Data sources can copy from each other and errors can be propagated quickly.

Consider accuracy of sources
Consider freshness of sources
Consider dependence between sources
Trust Accurate Sources

- Considering accuracy can often improve truth discovery

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker</td>
<td>MIT</td>
<td>Berkeley</td>
<td>MIT</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UWisc</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>AT&amp;T</td>
<td>BEA</td>
</tr>
<tr>
<td>Halevy</td>
<td>Google</td>
<td>Google</td>
<td>UW</td>
</tr>
</tbody>
</table>
Trust Accurate Sources

- Considering accuracy can often improve truth discovery

<table>
<thead>
<tr>
<th>Source</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker</td>
<td>MIT</td>
<td>Berkeley</td>
<td>MIT</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UWisc</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>AT&amp;T</td>
<td>BEA</td>
</tr>
<tr>
<td>Halevy</td>
<td>Google</td>
<td>Google</td>
<td>UW</td>
</tr>
</tbody>
</table>

S1 is more accurate; trusting it more can help find the correct affiliation for Carey.
**Trust Accurate Sources**

- Considering accuracy can often improve truth discovery

<table>
<thead>
<tr>
<th>Source</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker</td>
<td>MIT</td>
<td>Berkeley</td>
<td>MIT</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UWisc</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>AT&amp;T</td>
<td>BEA</td>
</tr>
<tr>
<td>Halevy</td>
<td>Google</td>
<td>Google</td>
<td>UW</td>
</tr>
</tbody>
</table>

S1 is more accurate; trusting it more can help find the correct affiliation for Carey.
Deciding authority based on link analysis and source popularity

- Survey: “Link analysis ranking: algorithms, theory, and experiments” [Borodin et al., 05]
- PageRank [Brin and Page, 98]
- Authority-hub analysis [Kleinberg, 98]
Assign a global trust rating to each data source based on its behavior in a P2P network

- TrustMe [Singh and Liu, 03]
- EigenTrust [Kamvar et al., 03]

\[
s_{ij} = \text{sat}(i, j) - \text{unsat}(i, j)
\]

\[
c_{ij} = \frac{\max(s_{ij}, 0)}{\sum_{j} \max(S_{ij}, 0)}
\]

\[
t_{ij} = \sum_{k} c_{ik} c_{kj}
\]
Compute accuracy of sources

- Corroborating answers from web sources [Wu and Marian, 07]
- TruthFinder [Yin et al., 07]
- Solomon [Dong et al., 09a]

\[
A(S) = \text{Avg}_{v \in \overline{V}(S)} P(v)
\]

\(\overline{V}(S)\) - values provided by S; \(P(v)\) - pr of value \(v\) being true

How to compute \(P(v)\)?
Apply Source Accuracy in Truth Discovery

[Yin et al., 07] [Dong et al., 09a]

- **Input:**
  - Object $O$
  - $\text{Dom}(O) = \{v_0, v_1, \ldots, v_n\}$
  - Observation $\Phi$ on $O$

- **Output:** $\Pr(v_i \text{ true} | \Phi)$ for each $i=0, \ldots, n$ (sum up to 1)

According to the Bayes Rule, we need to know $\Pr(\Phi | v_i \text{ true})$

- Assuming independence of sources, we need to know $\Pr(\Phi(S) | v_i \text{ true})$
  - If $S$ provides $v_i : \Pr(\Phi(S) | v_i \text{ true}) = A(S)$
  - If $S$ does not provide $v_i : \Pr(\Phi(S) | v_i \text{ true}) = (1 - A(S))/n$
Properties

- A value provided by more accurate sources has a higher probability to be true
- Assuming uniform accuracy, a value provided by more sources has a higher probability to be true

Consider value similarity

$$C^*(v) = C(v) + \rho \sum_{v' \neq v} C(v') \cdot \text{sim}(v, v')$$

Source accuracy

$$A(S) = \text{Avg}_{v \in \overline{V}(S)} P(v)$$

Value probability

$$P(v) = \frac{e^{C(v)}}{\sum_{v_0 \in D(O)} e^{C(v_0)}}$$

Source trustworthy

$$A'(S) = \ln \frac{nA(S)}{1 - A(S)}$$

Value confidence

$$C(v) = \sum_{S \in S(v)} A'(S)$$

Continue until source accuracy converges
## An Example

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker</td>
<td>MIT</td>
<td>Berkeley</td>
<td>MIT</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UWisc</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>AT&amp;T</td>
<td>BEA</td>
</tr>
<tr>
<td>Halevy</td>
<td>Google</td>
<td>Google</td>
<td>UW</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round 1</td>
<td>.69</td>
<td>.57</td>
<td>.45</td>
</tr>
<tr>
<td>Round 2</td>
<td>.81</td>
<td>.63</td>
<td>.41</td>
</tr>
<tr>
<td>Round 3</td>
<td>.87</td>
<td>.65</td>
<td>.40</td>
</tr>
<tr>
<td>Round 4</td>
<td>.90</td>
<td>.64</td>
<td>.39</td>
</tr>
<tr>
<td>Round 5</td>
<td>.93</td>
<td>.63</td>
<td>.40</td>
</tr>
<tr>
<td>Round 6</td>
<td>.95</td>
<td>.62</td>
<td>.40</td>
</tr>
<tr>
<td>Round 7</td>
<td>.96</td>
<td>.62</td>
<td>.40</td>
</tr>
<tr>
<td>Round 8</td>
<td>.97</td>
<td>.61</td>
<td>.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value Confidence</th>
<th>UCI</th>
<th>AT&amp;T</th>
<th>BEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round 1</td>
<td>1.61</td>
<td>1.61</td>
<td>1.61</td>
</tr>
<tr>
<td>Round 2</td>
<td>2.40</td>
<td>1.89</td>
<td>1.42</td>
</tr>
<tr>
<td>Round 3</td>
<td>3.05</td>
<td>2.16</td>
<td>1.26</td>
</tr>
<tr>
<td>Round 4</td>
<td>3.51</td>
<td>2.23</td>
<td>1.19</td>
</tr>
<tr>
<td>Round 5</td>
<td>3.86</td>
<td>2.20</td>
<td>1.18</td>
</tr>
<tr>
<td>Round 6</td>
<td>4.17</td>
<td>2.15</td>
<td>1.19</td>
</tr>
<tr>
<td>Round 7</td>
<td>4.47</td>
<td>2.11</td>
<td>1.20</td>
</tr>
<tr>
<td>Round 8</td>
<td>4.76</td>
<td>2.09</td>
<td>1.20</td>
</tr>
</tbody>
</table>
Data sources are of different quality and we trust data from accurate sources more.

The real world is dynamic and the true value often evolves over time. For example, person affiliation, business contact phone.

Data sources can copy from each other and errors can be propagated quickly.

Consider accuracy of sources.
Consider freshness of sources.
Consider dependence between sources.
A Dynamic World

- True values can evolve over time
  - A subtle third case: *out-of-date*

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker</td>
<td>(03, MIT)</td>
<td>(00, Berkeley)</td>
<td>(06, MIT)</td>
</tr>
<tr>
<td>Dewitt</td>
<td>(09, MSR)</td>
<td>(08, MSR)</td>
<td>(01, UWisc)</td>
</tr>
<tr>
<td>Bernstein</td>
<td>(00, MSR)</td>
<td>(00, MSR)</td>
<td>(01, MSR)</td>
</tr>
<tr>
<td>Carey</td>
<td>(09, UCI)</td>
<td>(05, AT&amp;T)</td>
<td>(06, BEA)</td>
</tr>
<tr>
<td>Halevy</td>
<td>(07, Google)</td>
<td>(05, Google)</td>
<td>(06, UW)</td>
</tr>
</tbody>
</table>
## A Dynamic World

- True values can evolve over time
  - A subtle third case: *out-of-date*
- Low-quality data can be caused by different reasons

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker ((\Theta, \text{Berkeley})), (02, <strong>MIT</strong>)</td>
<td>(03, MIT)</td>
<td><strong>(00, Berkeley)</strong></td>
<td>(01, Berkeley)</td>
</tr>
<tr>
<td>Dewitt ((\Theta, \text{UWisc})), (08, <strong>MSR</strong>)</td>
<td>(00, UWisc)</td>
<td>(00, UW)</td>
<td><strong>(01, UWisc)</strong></td>
</tr>
<tr>
<td>Bernstein ((\Theta, <strong>MSR</strong>))</td>
<td>(00, MSR)</td>
<td>(00, MSR)</td>
<td>(01, MSR)</td>
</tr>
<tr>
<td>Carey ((\Theta, \text{Propell})), (02, <strong>BEA</strong>), (08, <strong>UCI</strong>)</td>
<td>(04, BEA)</td>
<td><strong>(05, AT&amp;T)</strong></td>
<td><strong>(06, BEA)</strong></td>
</tr>
<tr>
<td>Halevy ((\Theta, \text{UW})), (05, <strong>Google</strong>)</td>
<td>(00, UW)</td>
<td>(00, UWisc)</td>
<td>(01, UWisc)</td>
</tr>
</tbody>
</table>

*ERR!* *OUT-OF-DATE!* *SLOW!*
Refine Accuracy of Sources [Dong et al., 09b]

- How many transitions are captured
- How quickly transitions are captured
- How many transitions are not mis-captured

Coverage = \#Captured/\#Capturable (e.g., ¼ = .25)

Exactness = 1 - \#Mis-Captured/\#Mis-Capturable (e.g., 1 - 2/5 = .6)

Freshness(\(\Delta\)) = \#(Captured w. length \(\leq \Delta\))/\#Captured (e.g., \(F(0) = 0, F(1) = 0, F(2) = 1/1 = 1\ldots\))
Freshness Measures in Other Work

- Other work on data freshness: Compare a materialized view with the original source
  - [Peralta, Ph.D. Thesis’06]: timeliness, currency
  - [Guo et al., 05]: completeness, consistency, currency
  - [Olston and Widom, 05]: divergence
  - [Labrinidis and Roussopoulos, 04]: QoD(freshness)
  - [Theodoratos and Bouzeghoub, 01]: consistency
  - [Cho and Garcia-Molina, 00]: freshness, age
Decide the initial value $v_0$

Decide the next transition $(t,v)$

Terminate when no more transition
Decide the initial value $v_0$

Decide the next transition $(t, v)$

- **Decide the initial value**: according to the Bayes Rule, we need to know
  - $\Pr(\Phi(S) \mid v_i)$ for each value $v_i$
    - If $S$ provides $v_i$: $E(S)C(S)$
    - If $S$ does not provide any value: $E(S)(1 - C(S))$
    - If $S$ provides another value: $(1 - E(S))/n$
  - $\Pr(\Phi(S) \mid \bot)$—the object does not exist initially
    - If $S$ does not provide any value: $E(S)$
    - If $S$ provides a value: $(1 - E(S))/(n+1)$

Terminate when no more transition
Decide the initial value $v_0$.

Terminate when no more transition.

Decide the next transition $(t,v)$.
Discover Evolving True Values

Decide the initial value $v_0$

Decide the next transition $(t,v)$

- **Decide the next transition $(t,v)$: according to the Bayes Rule, we need to know**
  - $Pr(\Phi(S) | (t_i, v_j))$ for each time $t_i$ and value $v_j$
    - If $S$ provides $v_j$ at time $t$: $E(S)C(S)F(S, t-t_i)$
    - If $S$ does not update any more: $E(S)(1-C(S)F(S, t_n-t_i))$
    - If $S$ makes a wrong update: $(1-E(S))/n(t_n-t')$ \((t_n—\text{the last obs point}, t’—\text{time of the prev update})$
  - $Pr(\Phi(S) | \text{no more transition})$: similarly computed
    - If $S$ does not update any more: $E(S)$
    - If $S$ makes a wrong update: $(1-E(S))/(n+1)(t_n-t’)$

Terminate when no more transition
## An Example

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Halevy</td>
<td>(00, UW)</td>
<td>(00, UWisc)</td>
<td>(01, UWisc)</td>
</tr>
<tr>
<td>(Θ, UW), (05, Google)</td>
<td>(07, Google)</td>
<td>(02, UW)</td>
<td>(06, UW)</td>
</tr>
</tbody>
</table>

### Affiliation for Halevy:

<table>
<thead>
<tr>
<th>Rnd</th>
<th>UWisc</th>
<th>UW</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rnd 1</td>
<td>2000</td>
<td>2002</td>
<td>2004</td>
</tr>
<tr>
<td>Rnd 2</td>
<td>2000</td>
<td></td>
<td>2004</td>
</tr>
<tr>
<td>Rnd 3</td>
<td>2000</td>
<td></td>
<td>2005</td>
</tr>
</tbody>
</table>
Advanced Truth-Discovery Techniques

- Data sources are of different quality and we trust data from accurate sources more.
- The real world is dynamic and the true value often evolves over time. E.g., person affiliation, business contact phone.
- Data sources can copy from each other and errors can be propagated quickly.

Consider accuracy of sources.

Consider freshness of sources.

Consider dependence between sources.
Copied Data Can Change Truth Discovery Results

- Previous methods assume source independence

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stonebraker</td>
<td>MIT</td>
<td>Berkeley</td>
<td>MIT</td>
<td>MIT</td>
<td>MS</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UWisc</td>
<td>UWisc</td>
<td>UWisc</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>AT&amp;T</td>
<td>BEA</td>
<td>BEA</td>
<td>BEA</td>
</tr>
<tr>
<td>Halevy</td>
<td>Google</td>
<td>Google</td>
<td>UW</td>
<td>UW</td>
<td>UW</td>
</tr>
</tbody>
</table>
10 sources voting for an object

- Count = 3
  - \(S_1\)
  - \(S_2\)
  - \(S_3\)

- Count = 5
  - \(S_4\)
  - \(S_5\)
  - \(S_6\)
  - \(S_7\)
  - \(S_8\)

- Count = 2
  - \(S_9\)
  - \(S_{10}\)
Voting w. Knowledge of Copying

- 10 sources voting for an object

Count = 1

Count = 2

Count = 3
10 sources voting for an object

How to compute vote count?

How to detect copying?
Opinion pooling: combine probability distribution from multiple experts

- Combination of opinions [Chang, Ph.D. thesis’85]
- Reconciliation of probability distributions [Lindley, 83]
- Updating of belief in the light of someone else’s opinion [French, 80]

Data fusion w. source dependence

- [Dong et al., 09a][Dong et al., 09b]

See Tomorrow’s talks in “Data Integration I”
Outline

- Data fusion in the integration process
- Foundations of data fusion
  - Conflict resolution strategies and functions
  - Conflict resolution operators
- Advanced truth-discovery techniques
- Data fusion in existing integration systems
- Open problems
Allows discussion of values between users
Typical ETL tools support rule-based fusion

- IIS (IBM Information Server)
- SSIS (Microsoft’s SQL Server Integration Services)
- Etc.

See details in survey [Bleiholder and Naumann, 08]
### Research DI Systems w. Awareness of Data Conflicts

<table>
<thead>
<tr>
<th>System</th>
<th>Conflict types</th>
<th>Methodology</th>
<th>Strategy</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multibase</td>
<td>Schematic, data</td>
<td>Resolution</td>
<td>Choose, Avg, Min, Max, Sum, ...</td>
<td>Manually, in query</td>
</tr>
<tr>
<td>Hermes</td>
<td>Schematic, data</td>
<td>Resolution</td>
<td>MostRecent, Choose</td>
<td>Manually, in mediator</td>
</tr>
<tr>
<td>Fusionplex</td>
<td>Schematic, object, data</td>
<td>Resolution</td>
<td>MostRecent, Min, Max, Avg, ...</td>
<td>Manually, in query</td>
</tr>
<tr>
<td>HumMer</td>
<td>Schematic, object, data</td>
<td>Resolution</td>
<td>MostAbstract, Vote, Min, ChooseDepen...</td>
<td>Manually, in query</td>
</tr>
<tr>
<td>Ajax</td>
<td>Schematic, object, data</td>
<td>Resolution</td>
<td>Various</td>
<td>Manually, in workflow definition</td>
</tr>
<tr>
<td>TSIMMIS</td>
<td>Schematic, data</td>
<td>Avoidance</td>
<td>Choose</td>
<td>Manually, rules in mediator</td>
</tr>
<tr>
<td>SIMS/Ariadne</td>
<td>Schematic, data</td>
<td>Avoidance</td>
<td>Choose</td>
<td>Automatically</td>
</tr>
<tr>
<td>Infomix</td>
<td>Schematic, data</td>
<td>Avoidance</td>
<td>onlyConsistentValue</td>
<td>Automatically</td>
</tr>
<tr>
<td>Hippo</td>
<td>Schematic, object, data</td>
<td>Avoidance</td>
<td>onlyConsistentValue</td>
<td>Automatically</td>
</tr>
<tr>
<td>ConQuer</td>
<td>Schematic, object, data</td>
<td>Avoidance</td>
<td>onlyConsistentValue</td>
<td>Automatically</td>
</tr>
<tr>
<td>Rainbow</td>
<td>Schematic, object, data</td>
<td>Avoidance</td>
<td>onlyConsistentValue</td>
<td>Automatically</td>
</tr>
<tr>
<td>Pegasus</td>
<td>Schematic, data</td>
<td>Ignorance</td>
<td>Escalate</td>
<td>Manually</td>
</tr>
<tr>
<td>Nimble</td>
<td>Unknown</td>
<td>Ignorance</td>
<td>Escalate</td>
<td>Manually</td>
</tr>
<tr>
<td>Carnot</td>
<td>Schematic</td>
<td>Ignorance</td>
<td>Escalate</td>
<td>Automatically</td>
</tr>
<tr>
<td>InfoSleuth</td>
<td>Schematic</td>
<td>Ignorance</td>
<td>Escalate</td>
<td>Unknown</td>
</tr>
<tr>
<td>Potter’s Wheel</td>
<td>Schematic</td>
<td>Ignorance</td>
<td>Escalate</td>
<td>Manually, transformation</td>
</tr>
</tbody>
</table>
Other DI Systems

- Research DI systems
  - Trio: including accuracy and lineage into data model
  - Information Manifold
  - Garlic
  - Disco (Distributed Information Search Component)
  - etc.

- Peer data management systems
  - Orchestra: allowing multiple viewpoints
  - Hyper: isolating the minimum amount of data to reach consistency

See details in survey [Bleïholder and Naumann, 08]
Outline

- Data fusion in the integration process
- Foundations of data fusion
  - Conflict resolution strategies and functions
  - Conflict resolution operators
- Advanced truth-discovery techniques
- Data fusion in existing integration systems
- Open problems
Open Problems

- Accuracy of fusion
- Efficiency of fusion
- Usability of fusion
- Interaction with other components of data integration
Challenge 1: Correlated values

E.g.1, (firstName, lastName) from 4 sources
- S1: (Xin, Dong)
- S2: (Xin Luna, Dong)
- S3: (Dong, Xin)
- S4: (Dong, Xin Luna)

E.g.2, (ISBN, authors) from 3 sources
- S1: (**1, Peter Loshin) (**2, Peter Loshin)
- S2: (**1, Pete Loshin)
- S3: (**1, Pete Loshin)

Current effort: ChooseDepending(val, col)

Directions: consider correlation at the attribute level and at the instance level.
Challenge 2: Different formatting styles

E.g., (ISBN, authors) from 4 sources

Src1
(*1, Pete Loshin)
(*2, Dennis Suhanovs)
(*3, Zhigang Xiang, Roy A Plastock)
(*4, Peter Aiken, David M Allen)...

Src2
(*1, Pete Loshin)
(*2, Dennis Suhanovs)
(*3, Zhigang Xiang, Roy Plastock)
(*4, Peter Aiken, David Allen)...

Src3
(*1, Pete Loshin)
(*2, Dennis Suhanovs)
(*3, Zhigang Xiang)
(*4, David Allen, Peter Aiken)...

Skip middle-names
**Challenge 2:** Different formatting styles

- E.g., (ISBN, authors) from 4 sources

**Src1**
- (**1, Pete Loshin**)
- (**2, Dennis Suhanovs**)
- (**3, Zhigang Xiang, Roy A Plastock**)
- (**4, Peter Aiken, David M Allen**)...  

**Src2**
- (**1, Pete Loshin**)
- (**2, Dennis Suhanovs**)
- (**3, Zhigang Xiang, Roy Plastock**)
- (**4, Peter Aiken, David Allen**)...  

**Src3**
- (**1, Pete Loshin**)
- (**2, Dennis Suhanovs**)
- (**3, Zhigang Xiang**)
- (**4, David Allen, Peter Aiken**)...  

**Src4**
- (**1, Pete Loshin**)
- (**2, Dennis Suhanovs**)
- (**3, Zhigang Xiang**)
- (**4, Peter Aiken**)...  

**Current effort:** consider value similarity

**Directions:** consider formatting styles used by each source.
Challenge 3: Source profiling

Current effort: accuracy (coverage, exactness, freshness)

Data properties can be different for different categories of data
- Source A is a vertical source on restaurants
- Source B knows very well about NYC

Data properties can evolve over time
- Source C improves its data over time

Directions: partition data into different portions and profile on each portion
## Challenge 4: Incremental fusion

When we have more data sources (e.g., Src4) or lose some data sources, shall we do data fusion from scratch?

<table>
<thead>
<tr>
<th>Src1</th>
<th>Src2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(**1, Pete Loshin)</td>
<td>(**1, Pete Loshin)</td>
</tr>
<tr>
<td>(**2, Dennis Suhanovs)</td>
<td>(**2, Dennis Suhanovs)</td>
</tr>
<tr>
<td>(**3, Zhigang Xiang, Roy A Plastock)</td>
<td>(**3, Zhigang Xiang, Roy Plastock)</td>
</tr>
<tr>
<td>(**4, Peter Aiken, David M Allen)</td>
<td>(**4, Peter Aiken, David Allen)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Src3</th>
<th>Src4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(**1, Pete Loshin)</td>
<td>(**1, Pete Loshin)</td>
</tr>
<tr>
<td>(**2, Dennis Suhanovs)</td>
<td>(**2, Dennis Suhanovs)</td>
</tr>
<tr>
<td>(**3, Zhigang Xiang)</td>
<td>(**3, Zhigang Xiang)</td>
</tr>
<tr>
<td>(**4, David Allen, Peter Aiken)</td>
<td>(**4, Peter Aiken)</td>
</tr>
</tbody>
</table>

When more data come, shall we start from scratch?

**Directions:** maintain metadata or statistics, retain data lineage.
Challenge 5: Runtime fusion

- In some applications fusing data upfront is infeasible

Directions: maintain source profiles by sampling; emphasize efficiency.
**Challenge 6:** Personalized fusion

- Express preference on certain sources
- Emphasize certain property; e.g., up-to-date vs. high coverage
- Use certain formats; e.g., full author list vs. only first author

**Current effort:**

- Function `choose(src)`
- Operator Prioritized-Merge

**Directions:**

- A language to express such user preferences
- Algorithms for efficient execution.
Challenge 7: User feedback

- Correct certain errors

Directions:

- Critical questions that can best improve the fusion results
- A way for users to browse source data and fusion results, and correct mistakes
- Quickly fixing errors and propagation to related items
Challenge 8: Data Lineage

- Legal requirement
- Application requirement: e.g., fusing two customers
- HCI requirement: HOW did you merge the data? And WHY?

Directions:
- Effective representation of lineage information
- Explanation of merging decisions
- Effective way to find disappeared data items
- Reversibility and repeatability
Challenge 9: Fuse data w. different schemas

E.g., Contact information from three sources

- S1: (pid = "1", work phone = "1234", home phone = "8765", mobile phone = "4321")
- S2: (pid = "1", daytime phone = "1234", evening phone = "4321")
- S3: (pid = "1", phone = "4321")

Directions: Combine data fusion w. schema matching
Challenge 10: Distinguish wrong values from alternative representations of correct values
- E.g., A quiz
Which type of listing are they?

A: are the same business

B: are different businesses sharing the same phone#

C: are different businesses, only one with correct phone#
**Challenge 10**: Distinguish wrong values from alternative representations of correct values

- E.g., A quiz

**Directions**: Combine data fusion w. record linkage
Conclusions

- **Foundations**
  - Strategies and functions
  - Operators

- **Advanced techniques**
  - Consider accuracy
  - Consider freshness
  - Consider dependence

- **Open problems**
  - Accuracy
  - Efficiency
  - Usability
  - Interaction with other components of DI
References

Survey

Foundations of Fusion
Advanced truth-discovery techniques


