Database design and implementation
CMPSCI 645

Lecture 02: Project overview
Overview of deliverables

- Form groups (week 3)
- Project proposal (week 5)
- Literature survey (week 7)
- Midterm status report (week 10)
- Project presentations (last week of classes)
- Project report
Data cleaning

- Cost of data errors:
  - Poor data quality costs the US economy more than $600 billion per year [Eckerson’02]
  - Erroneous price data in retail databases costs US consumers $2.5 billion each year [Fan et al.’08]
  - Consumes 30-80% of the development time and budget in data warehouses [Fan et al.’08]
Using FDs for cleaning

<table>
<thead>
<tr>
<th>CC</th>
<th>AC</th>
<th>PN</th>
<th>NM</th>
<th>STR</th>
<th>CT</th>
<th>ZIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>908</td>
<td>1111111</td>
<td>Mike</td>
<td>Tree Ave.</td>
<td>MH</td>
<td>07974</td>
</tr>
<tr>
<td>01</td>
<td>908</td>
<td>1111111</td>
<td>Rick</td>
<td>Tree Ave.</td>
<td>MH</td>
<td>07974</td>
</tr>
<tr>
<td>01</td>
<td>212</td>
<td>2222222</td>
<td>Joe</td>
<td>5th Ave</td>
<td>NYC</td>
<td>01202</td>
</tr>
<tr>
<td>01</td>
<td>908</td>
<td>2222222</td>
<td>Jim</td>
<td>Elm Str.</td>
<td>MH</td>
<td>07974</td>
</tr>
<tr>
<td>44</td>
<td>131</td>
<td>3333333</td>
<td>Ben</td>
<td>High St.</td>
<td>EDI</td>
<td>EH4 1DT</td>
</tr>
<tr>
<td>44</td>
<td>131</td>
<td>4444444</td>
<td>Ian</td>
<td>High St.</td>
<td>EDI</td>
<td>EH4 1DT</td>
</tr>
<tr>
<td>44</td>
<td>908</td>
<td>4444444</td>
<td>Ian</td>
<td>Port Pl</td>
<td>MH</td>
<td>W1B 1JH</td>
</tr>
<tr>
<td>01</td>
<td>131</td>
<td>2222222</td>
<td>Sean</td>
<td>3rd Str.</td>
<td>UN</td>
<td>01202</td>
</tr>
</tbody>
</table>

Examples:
FD: STR → ZIP

\[ cFD: [CC, ZIP] \rightarrow STR, (44, _ || _ ) \]

Challenge 1: discover FDs from data
Challenge 2: use FDs to clean the data

credit: Wenfei Fan
Data fusion

American Airlines Flight Number 119 (AA119)

FLIGHT TRACKER

Aircraft: Boeing 737-800 (twin-jet) (B738/Q - track or photos)
Origin: Terminal A / Gate 32 / Newark Liberty Intl (EWR - track or info)
Destination: Terminal 4 / Gate 42B / Los Angeles Intl (LAX - track or info)
Aircraft: American Airlines "American"

Route: "The flights between these airports"
Date: 2011年12月08日 (Thursday)
Duration: 5 hours 43 minutes
20 minutes left
5 hours 23 minutes

Status: "En Route" (2,284 sm done, 3,082 sm to go)
Distance: Direct: 2,451 sm. Planned: 2,458
Fare: $51.99 to $3,569.19, average: $241.96 (airline insight)
Cabin: First Dinner: Economy: Food for sale

Scheduled Estimated Actual
6:22p - 6:32p
Dec 8 - Dec 8

Arrives: Los Angeles (LAX) View real-time airport conditions
Gate: 42B

Scheduled Estimated Actual
9:54p - 9:47p
Dec 8 - Dec 8

Arrives: Los Angeles (LAX) View real-time airport conditions
Gate: 42B

credit: Luna Dong, Divesh Srivastava
Ideas?

- Majority voting?
- What if websites copy information?
- Can you account for correlations?
- What if there are malicious contributors / spammers?
Data disambiguation

Bogdan Alexe

2014


2013

[c18] Tyler Baldwin, Yunyao Li, Bogdan Alexe, Ioana Roxana Stanoiu: Automatic Term Ambiguity Detection. ACL 2013: 804-809


2012


[c13] Bogdan Alexe, Mauricio A. Hernández, Kirsten Hildrum, Rajasekar Krishnamurthy, Georgia Koutrika, Meenakshi Nagarajan, Haggai Roitman,
MapMerge: Correlating Independent Schema Mappings

Bogdan Alexe  
UC Santa Cruz

Mauricio Hernández  
IBM Almaden

Lucian Popa  
IBM Almaden

Wang-Chiew Tan  
IBM Almaden & UC Santa Cruz

Preference-aware Integration of Temporal Data

Bogdan Alexe  
IBM Almaden
balexe@us.ibm.com

Mary Roth  
IBM Almaden and UCSC
torkroth@us.ibm.com

Wang-Chiew Tan  
UCSC
tan@cs.ucsc.edu

---

Exploiting spatial overlap to efficiently compute appearance distances between image windows

Bogdan Alexe  
ETH Zurich

Viviana Petrescu  
ETH Zurich

Vittorio Ferrari  
ETH Zurich

Abstract

We present a computationally efficient technique to compute the distance of high-dimensional appearance descriptor vectors between image windows. The method exploits the relation between appearance distance and spatial overlap. We derive an upper bound on appearance distance given the spatial overlap of two windows.
Ideas?

- Affiliation information?
- Co-authorship graph?
- Paper content?
- Crowdsourcing?
Data diagnosis
Post-factum cleaning with uncertainty

<table>
<thead>
<tr>
<th>Data</th>
<th>Transformations</th>
<th>Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accelerometer</td>
<td>Periodicity $p$</td>
<td>Sensor data</td>
</tr>
<tr>
<td>GPS</td>
<td>HasSignal? $h$</td>
<td>true</td>
</tr>
<tr>
<td>Cell Tower</td>
<td>Speed $s$</td>
<td>false</td>
</tr>
<tr>
<td>Audio</td>
<td>Rate of Change $r$</td>
<td>false</td>
</tr>
<tr>
<td>Light</td>
<td>Avg. Strength $a$</td>
<td>true</td>
</tr>
<tr>
<td></td>
<td>Zero crossing rate $z$</td>
<td>false</td>
</tr>
<tr>
<td></td>
<td>Spectral roll-off $c$</td>
<td>false</td>
</tr>
<tr>
<td></td>
<td>Avg. Intensity $i$</td>
<td>true</td>
</tr>
</tbody>
</table>

| | Is Walking? $M(p > P_w, R_s < r < R_w, \neg h \lor (s < S_w))$ |
| | Is Driving? $M(p < P_d, r > R_d, h, s > S_d)$ |
| | Alone? $(A_2 \geq a > A_1) \lor ((a > A_2) \land (z > Z)) \lor ((a > A_3) \land (z < Z) \land (c > C))$ |
| | Is Indoor? $M(\neg h, i < I_i)$ |
| | Is Meeting? $M(\neg h, i < I_m, a > A_m, z > Z_m)$ |

| Sensors may be faulty or inhibited |
| It is not straightforward to spot such errors in the provenance |

| 0.016 | True | 0.067 | 0 | 0.4 | 0.004 | 0.86 | 0.036 | 10 |
| 0.0009 | False | 0 | 0 | 0.2 | 0.0039 | 0.81 | 0.034 | 68 |
| 0.005 | True | 0.19 | 0 | 0.03 | 0.003 | 0.75 | 0.033 | 17 |
| 0.0008 | True | 0.003 | 0 | 0.1 | 0.003 | 0.8 | 0.038 | 18 |

What caused these errors?
Post-factum cleaning with uncertainty

- What if we don’t know for sure that there is an error?
- What if we know for sure that there is an error, but we don’t know where?
Query debugging

Example: find all actors with Bacon number 2.

```sql
SELECT a3.fname, a3.lname
FROM Actor a0, Casts c0, Casts c1,
     Casts c2, Casts c3, Actor a3
WHERE a0.fname = 'Kevin' AND a0.lname = 'Bacon'
  AND c0.pid = a0.id AND c0.mid = c1.mid AND
  c1.pid = c2.pid AND c2.mid = c3.mid AND
  c3.pid = a3.id
  AND NOT (a3.fname = 'Kevin' and a3.lname = 'Bacon')
  AND NOT EXISTS (SELECT xc1.pid
                  FROM Actor xa0, Casts xc0, Casts xc1
                  WHERE xa0.fname = 'Kevin' AND
                       xa0.lname = 'Bacon'
                       AND xa0.id = xc0.pid AND
                       xc0.mid = xc1.mid AND xc1.pid = a3.id)
GROUP BY a3.id, a3.fname, a3.lname;
```
Ideas?

- Data sampling?
- Visualizations?
- Different query languages?
- Query by example?
  - Instead of writing a query, give examples of the results
- Can you query with natural language?
Query containment/equivalence

- Equivalence: \( q_1 \equiv q_2 \iff q_1(D) = q_2(D), \forall D \)

- Containment: \( q_1 \subseteq q_2 \iff q_1(D) \subseteq q_2(D), \forall D \)

- Query rewrite with views.

- Idea: use SAT solvers

NP-hard
Databases for applications

- A declarative language for animation
  - Create characters by “joining” existing components
  - Find rigs (skeletons) for a given character
Databases for applications

- Navigating and querying large image data

credit: Sloan digital sky survey
Probabilistic databases

- What are the probabilities in the join results?
- Can you find query classes for which inference is tractable?

Researchers:

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
<th>P</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fred</td>
<td>U. Washington</td>
<td>$p_1^1 = 0.3$</td>
<td>$X_1 = 1$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U. Wisconsin</td>
<td>$p_1^2 = 0.2$</td>
<td>$X_1 = 2$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Y! Research</td>
<td>$p_1^3 = 0.5$</td>
<td>$X_1 = 3$</td>
<td></td>
</tr>
<tr>
<td>Sue</td>
<td>U. Washington</td>
<td>$p_2^1 = 1.0$</td>
<td>$X_2 = 1$</td>
<td></td>
</tr>
<tr>
<td>John</td>
<td>U. Wisconsin</td>
<td>$p_3^1 = 0.7$</td>
<td>$X_3 = 1$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U. Washington</td>
<td>$p_3^2 = 0.3$</td>
<td>$X_3 = 2$</td>
<td></td>
</tr>
<tr>
<td>Frank</td>
<td>Y! Research</td>
<td>$p_4^1 = 0.9$</td>
<td>$X_4 = 1$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M. Research</td>
<td>$p_4^2 = 0.1$</td>
<td>$X_4 = 2$</td>
<td></td>
</tr>
</tbody>
</table>

Services:

<table>
<thead>
<tr>
<th>Name</th>
<th>Conference</th>
<th>Role</th>
<th>P</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fred</td>
<td>VLDB</td>
<td>Session Chair</td>
<td>$q_1 = 0.2$</td>
<td>$Y_1 = 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PC Member</td>
<td>$q_2 = 0.8$</td>
<td>$Y_2 = 1$</td>
</tr>
<tr>
<td>John</td>
<td>SIGMOD</td>
<td>PC Member</td>
<td>$q_3 = 0.7$</td>
<td>$Y_3 = 1$</td>
</tr>
<tr>
<td></td>
<td>VLDB</td>
<td>PC Member</td>
<td>$q_4 = 0.7$</td>
<td>$Y_4 = 1$</td>
</tr>
<tr>
<td>Sue</td>
<td>SIGMOD</td>
<td>Chair</td>
<td>$q_5 = 0.5$</td>
<td>$Y_5 = 1$</td>
</tr>
</tbody>
</table>

credit: Dan Suciu
Provenance-based content authenticity

- Provenance: sources and process that derives new data.

  - Openly authored
  - Mistaken contributions
  - Malicious contributions

- Idea:
  - Use history to assess trustworthiness
  - Analyze edit behavior
  - Analyze citations
  - Propagate trust
Data auditing

- Example: medical records
  - Who should have access?

- Query language to model security violations
  - Query execution engine for this language

- Trust propagation
  - What is data gets tainted?
  - What other data is affected?
  - Which data is safe?
Data understanding

Why??

Support to understand surprising results

What if the results are not surprising? Can DBs add context?

Explain by example
Other projects

- ACM SIGMOD programming contest

- PVLDB experiment and analysis track
  - thorough evaluation of existing approaches