Advanced Data Management (CSCI 490/680)

Data Fusion

Dr. David Koop
Databases

• Databases:
  - Have been around for years
  - Organize data by tables, allow powerful queries
  - Most support concurrency: allowing multiple users to work with the database at once
  - Provide many features to ensure data integrity, security

• Database Management Systems (DBMS): software that manages databases and facilitates adding, updating, and removing data as well as queries over the data

• Main language used to interact with databases: Structured Query Language (SQL)
Football Game Data

• Have each game store the id of the home team and the id of the away team (one-to-one)
• Have each player store the id of the team he plays on (many-to-one)
Concatenation

• Take two data frames with the same columns and add more rows
• `pd.concat([data-frame-1, data-frame-2, ...])`
• Default is to add rows (axis=0), but can also add columns (axis=1)
• Can also concatenate Series into a data frame.
• `concat` preserves the index so this can be confusing if you have two default indices (0,1,2,3…)—they will appear twice
  - Use `ignore_index=True` to get a 0,1,2…
Merges (aka Joins)

- Want to join the two tables based on the location and date
- Location and date are the keys for the join
- Merges are ordered: there is a left and a right side

<table>
<thead>
<tr>
<th>Game</th>
<th>Weather</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>wId</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Id</th>
<th>Location</th>
<th>Date</th>
<th>Home</th>
<th>Away</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Boston</td>
<td>9/2</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>1</td>
<td>Boston</td>
<td>9/9</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>Cleveland</td>
<td>9/16</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>San Diego</td>
<td>9/23</td>
<td>21</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>wId</th>
<th>City</th>
<th>Date</th>
<th>Temp</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Boston</td>
<td>9/2</td>
<td>72</td>
</tr>
<tr>
<td>1</td>
<td>Boston</td>
<td>9/3</td>
<td>68</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>7</td>
<td>Boston</td>
<td>9/9</td>
<td>75</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>21</td>
<td>Boston</td>
<td>9/23</td>
<td>54</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>36</td>
<td>Cleveland</td>
<td>9/16</td>
<td>81</td>
</tr>
</tbody>
</table>

No data for San Diego
Types of Joins

- Inner: intersection of keys (match on both sides)
- Outer: union of keys (if there is no match on other side, still include with NaN to indicate missing data)
- Left: always have rows from left table (no unmatched right data)
- Right: like left, but with no unmatched left data
Data Merging in Pandas

- `pd.merge(left, right, ...)`
- Default merge: join on matching column names
- Better: specify the column name(s) to join on via `on` kwarg
  - If column names differ, use `left_on` and `right_on`
  - Multiple keys: use a list
- `how` kwarg specifies type of join ("inner", "outer", "left", "right")
- Can add suffixes to column names when they appear in both tables, but are not being joined on
- Can also merge using the index by setting `left_index` or `right_index` to True
Data Integration

```sql
select title, startTime
from Movie, Plays
where Movie.title=Plays.movie AND
      location="New York" AND
      director="Woody Allen"
```

Sources S1 and S3 are relevant, sources S4 and S5 are irrelevant, and source S2 is relevant but possibly redundant.

Sources:

- **S1**: Movies: name, actors, director, genre
- **S2**: Cinemas: place, movie, start
- **S3**: Cinemas in NYC: cinema, title, startTime
- **S4**: Cinemas in SF: location, movie, startingTime
- **S5**: Reviews: title, date, grade, review

References:
[AH Doan et al., 2012]
Data Integration

- Lots of data sources, how do we answer questions where we need to access data from more than one?
- Schema matching
- Problem of heterogeneity
- AI-Complete problem: difficulty is the same as making computers as intelligent as people
- Two techniques:
  - Mediation
  - Data Warehouses
Data Integration Application: Biomedical

- Phenotype
- Gene
- Sequenceable Entity
- Structured Vocabulary
- Protein
- Nucleotide Sequence
- Experiment
- Microarray Experiment

OMIM
HUGO
Swiss-Prot
GO
Entrez
GEO
Gene-Clinics
Locus-Link

[A. Doan et al., 2012]
Data Warehouses: Offline Replication

- Determine physical schema
- Define a database with this schema
- Define procedural mappings in an “ETL tool” to import the data and clean it.
- Periodically copy all of the data from the data sources
  - Note that the sources and the warehouse are basically independent at this point

[A. Doan et al., 2012]
Virtual Data Warehouses

Independence of:
• source & location
• data model, syntax
• semantic variations
• ...
Integrated Schema Example

[A. Doan et al., 2012]
Why is Data Integration Hard?

• Systems-level reasons:
  - Managing different platforms
  - SQL across multiple systems is not so simple
  - Distributed query processing

• Logical reasons:
  - Schema (and data) heterogeneity

• ‘Social’ reasons:
  - Locating and capturing relevant data in the enterprise.
  - Convincing people to share (data fiefdoms)
   • Security, privacy and performance implications

[A. Doan et al., 2012]
Assignment 3

- Data wrangling with
  - Trifacta Wrangler
  - pandas

- Same hurdat2 data

- Start now!

- Due Tuesday, March 3
Record Linkage Motivation

- Often data from different sources need to be integrated and linked
  - To allow data analyses that are impossible on individual databases
  - To improve data quality
  - To enrich data with additional information

- Lack of unique entity identifiers means that linking is often based on personal information

- When databases are linked across organisations, maintaining privacy and confidentiality is vital

- The linking of databases is challenged by data quality, database size, and privacy concerns

[P. Christen, 2019]
Motivating Example

- Preventing the outbreak of epidemics requires monitoring of occurrences of unusual patterns of symptoms, ideally in real time.
- Data from many different sources will need to be collected (including travel and immigration records; doctors, emergency and hospital admissions; drug purchases; social network and location data; and possibly even animal health data).

Map of the 2019–20 coronavirus outbreak (as of 25 February 2020):
- Region of origin
- Confirmed cases reported
- Suspected cases reported

[P. Christen, 2019], image: [Pharexia, Wikipedia]
Record Linkage

P. Christen
Record Linkage Process

1. Database A
   - Data pre-processing
   - Indexing / Searching
   - Comparison

2. Database B
   - Data pre-processing
   - Indexing / Searching
   - Comparison

3. Classification
4. Non-matches
5. Potential Matches
6. Evaluations
7. Potential Matches

[P. Christen, 2019]
Record Linkage Techniques

- **Deterministic matching**
  - Rule-based matching (complex to build and maintain)

- **Probabilistic record linkage [Fellegi and Sunter, 1969]**
  - Use available attributes for linking (often personal information, like names, addresses, dates of birth, etc.)
  - Calculate match weights for attributes

- **“Computer science” approaches**
  - Based on machine learning, data mining, database, or information retrieval techniques
  - Supervised classification: Requires training data (true matches)
  - Unsupervised: Clustering, collective, and graph based

[P. Christen, 2019]
Data Matching & Data Fusion

- **Google Thinks I’m Dead (I know otherwise.)** [R. Abrams, NYTimes, 2017]

- Not only Google, but also Alexa:
  - "Alexa replies that Rachel Abrams is a sprinter from the Northern Mariana Islands (which is true of someone else)."
  - "He asks if Rachel Abrams is deceased, and Alexa responds yes, citing information in the Knowledge Graph panel."

---

[Image of a screenshot showing a Google search result for Rachel Abrams with a note indicating the death date is incorrect.]
Data Integration and Data Fusion

- Data Integration: focus on integrating data from different sources
- When sources are orthogonal, no problems
- What happens when two sources provide the same type of information and they conflict?
- Data Fusion: create a single object while resolving conflicting values
Data Fusion—
Resolving Data Conflicts in Integration

X. L. Dong and F. Naumann
Data Fusion Summary

• Conflict resolution strategies
• "Truth-discovery" techniques
  - Accuracy
  - Freshness
  - Dependence
• Fusion Issues
  - Accuracy
  - Efficiency
  - Usability
  - How fusion fits with the rest of data integration?
Data Conflicts

[L. Dong and F. Naumann, 2009]
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> Scheth & Larson </author>
<year> 1990 </year>
</publication>

Schema Mapping Data Transformation Duplicate Detection Data Fusion

[L. Dong and F. Naumann, 2009]
Information Integration

L. Dong and F. Naumann, 2009

Source A

Source B

Federated Database Systems

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

Preserve lineage

Schema Mapping

Data Transformation

Duplicate Detection

Data Fusion

[L. Dong and F. Naumann, 2009]
Data Fusion

• 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.
Conflict Resolution Strategies

[Conflict Ignorance]
- PASS IT ON

[Conflict Avoidance]
- instance based
  - TAKE THE INFORMATION NO GOSSIPING
  - deciding
  - CRY WITH THE WOLVES ROLL THE DICE
  - mediating
  - MEET IN THE MIDDLE

[Conflict Resolution]
- instance based
  - deciding
  - CRY WITH THE WOLVES ROLL THE DICE
  - mediating
  - MEET IN THE MIDDLE
- metadata based
  - trusting your friends
  - mediating
  - NOTHING IS OLDER THAN THE NEWS FROM YESTERDAY

[L. Dong and F. Naumann, 2009]
Integrating Conflicting Data: The Role of Source Dependence

X. L. Dong, L. Berti-Equille, and D. Srivastava
Discussion

• What is the paper's main contribution?
• Do you buy the argument? Any issues with the experiments?
• Can you think of any scenarios where the proposed technique will fail?
• Questions?