Emerging Topics in Data Integration

Traditional Data Integration[1]

Schema Alignment > Record Linkage > Data Fusion

Usually done using pipeline architecture with three major steps:

1. **Schema Alignment**: Find attributes with the same meaning.
2. **Record Linkage**: Find records that refer to the same distinct entity.
3. **Data Fusion**: Decide the true value for an item with multiple sources.

Challenges

<table>
<thead>
<tr>
<th>Volume</th>
<th>Velocity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huge volume of data and large number of data sources.</td>
<td>Dynamic data sources with frequently changing information.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variety</th>
<th>Veracity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heterogeneous data sources and evolving schemas and representations.</td>
<td>Significant differences in data quality e.g. in coverage, accuracy, and timeliness.</td>
</tr>
</tbody>
</table>

Emerging Topics

**Volume**

- **Huge volume of data and large number of data sources.**

**Velocity**

- **Dynamic data sources with frequently changing information.**

**Variety**

- **Heterogeneous data sources and evolving schemas and representations.**

**Veracity**

- **Significant differences in data quality e.g. in coverage, accuracy, and timeliness.**

**Best Effort Schema Alignment**

**Goal**: Start with best effort solution with pay-as-you-go improvements[3].

- **Probabilistic Schema[4]**: Clustering of mapped attributes annotated with probability of them being true (p-schema).
- **Best Effort Queries[4]**: Queries return approximate answers based on p-schema.
- **Pay-as-you-go User Feedback[5]**: Improve mapping using user feedback. Maximize benefit by finding best candidates for users to decide on.

**Crowdsourcing**

“Hands-off Crowdsourcing”[2]

- End-to-end workflow for record linkage without external intervention.
- Achieves high accuracy with low costs.

**Future Work**

- Impact of data quality on crowdsourcing results.
- How to apply crowdsourcing to recent algorithmic innovations.

**Source Profiling**

**Goal**: Discover sources that are relevant and have sufficient quality.

**Bellman System[8]**: Surface data quality issues, find linked attributes, discover join paths, ...

**Database Summarizing[9]**: Identify domains and main tables. Cluster tables based on strength and importance of a table.

**Future Work**:

- Incremental profiling.
- Profiling for non-relational sources.

**References**

(3) Franklin et al. From databases to dataspaces: a new abstraction for information management. ACM SIGMOD Rec., 34 (4), 2005
(8) Dasu et al. Mining database structures: or, how to build a data quality browser. In Proc. ACM SIGMOD Int. Conf. on Management of Data, 2002
(9) Yang et al. Summarizing relational databases. Proc. VLDB Endowment, 2009