Query Processing in Data Integration

**The problem:**
- Integrated web sources are autonomous, incomplete, and partially overlapping.
- Fetching every possible source is inefficient and time-consuming.

**The solution:**
- Determine which sources are more relevant for a particular query.
- Determine what order the relevant sources should be called in.
- Learn statistics about individual sources and queries.

**The BibFinder/StatMiner approach:**
- Learn AV Hierarchy and QPQ overlap statistics to gather evidence about coverage of individual sources w.r.t. classes of queries.

**The result:**
- More tuples are obtained faster from only relevant sources and without fetching every integrated source.

Concepts

- **AV hierarchy:** A classification of the values of a particular attribute of the mediator relation. Leaf nodes in the hierarchy represent concrete values found in a query.
- **Candidate:** A probable answer to a query. Noted as \( C \).
- **QpQP:** Degree to which sources contain the same answer tuples for query \( Q \).

Using the Learned Statistics

1. A new user query is mapped to a set of least general query classes.
2. The mediator estimates the statistics for the query using a weighted sum of the statistics of the matched classes.
3. Data sources are ranked and called in order of relevance using the estimated statistics.

**Effects of Learned Statistics on BibFinder**

- DBLP has highest coverage, followed by CSB. However, since ACMDL is the highest residual coverage, it would be called last.
- DBLP has highest overlap, followed by CSB. However, since ACMDL has the highest residual overlap, it would be called last.

BibFinder/StatMiner Architecture

- **StatMiner**
  - **Learn AV Hierarchy**
  - **Learn QPQ overlap**
  - **Discover Frequent Query Classes**

BibFinder/StatMiner : Effectively Mining and Using Coverage and Overlap Statistics in Data Integration

Zaiqing Nie
Subbarao Kambhampati
Thomas Hernandez
Arizona State University, USA

Project Havasu

**Overlap (Coverage)**

- **DBLP, Science:** 3
- **ACMdl, DBLP:** 3
- **CSB, Science:** 1
- **ACMdl, CSB:** 3
- **CSB, DBLP:** 7
- **Science:** 3
- **ACMdl, CSB:** 3
- **ACMdl, CSB:** 5
- **DBLP, Science:** 1

**Using the Learned Statistics**

- **DBLP:** 16
- **CSB:** 23
- **ACMdl:** 5
- **Science:** 3
- **DBLP:** 35
- **DBLP:** 35
- **CSB:** 16
- **CSB:** 16

**Memory Consumption (bytes)**

- **Number of distinct answers:**
  - 1,000,000: 28
  - 1,200,000: 38
  - 1,400,000: 48
  - 1,600,000: 53

**Precision**

- **0.4, 0.5, 0.7, 0.8, 0.9:**
  - **Precision:**
    - **0.03, 0.13, 0.23, 0.33, 0.43, 0.53, 0.63, 0.73:**

Aptron algorithms.