29th International Conference on Very Large Data Bases

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

Zaiqing Nie
Subbarao Kambhampati
Thomas Hernandez

Arizona State University, USA

Query Processing in Data Integration

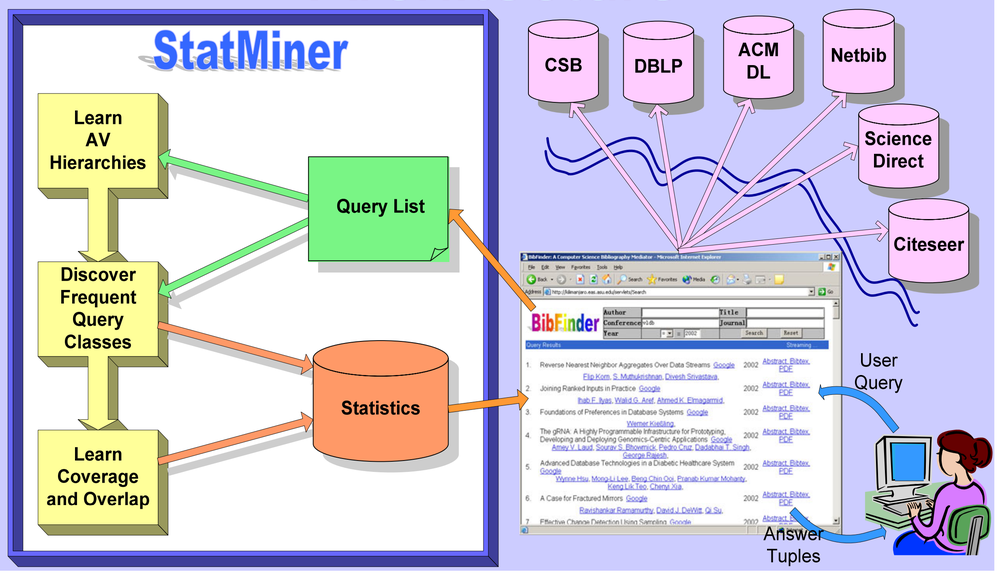
The problem:
- Integrated web sources are autonomous, incomplete, and partially overlapping
- Calling every possible source is inefficient and impolite

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

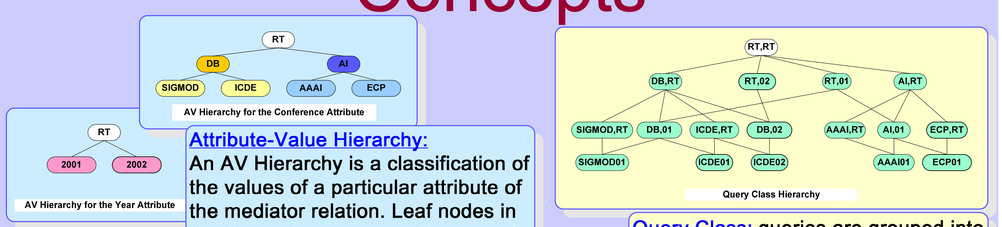
The BibFinder/StatMiner approach:
- Learn AV Hierarchies and Query Classes to gather statistics about coverage of individual sources w.r.t. classes of queries
- Using these statistics, only call the most relevant sources which have minimal overlap

The result:
More tuples are obtained faster from only relevant sources and without flooding every integrated sources

BibFinder/StatMiner Architecture



Concepts

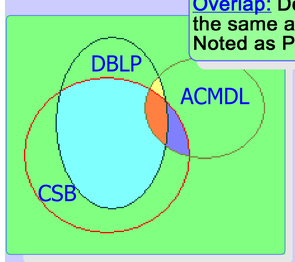


Attribute-Value Hierarchy:
An AV Hierarchy is a classification of the values of a particular attribute of the mediator relation. Leaf nodes in the hierarchy correspond to concrete values bound in a query.

Query Class: queries are grouped into classes by computing cartesian products over the AV Hierarchies. A query class is a set of queries that all share a set of assignments of particular attributes to specific values.

Coverage: probability that a random answer tuple for query Q belongs to source S. Noted as P(S|Q).

Overlap: Degree to which sources contain the same answer tuples for query Q. Noted as $P(S_1 \wedge S_2 \wedge \ldots \wedge S_k | Q)$.



Query List: the mediator maintains an XML log of all user queries, along with their access frequency, number of total distinct answers obtained, and number of answers from each source set which has answers for the query.

Query	Frequency	Distinctive Answers	Overlap (Coverage)	
Author="andy king"	106	46	DBLP	35
			CSB	23
			CSB, DBLP	12
			DBLP, Science	3
			Science	3
			CSB, DBLP, Science	1
			CSB, Science	1
Author="fayyad" Title="data mining"	1	27	CSB	16
			DBLP	16
			CSB, DBLP	7
			ACMdl	5
			ACMdl, CSB	3
			ACMdl, DBLP	3
			ACMdl, CSB, DBLP	2
			Science	1

StatMiner

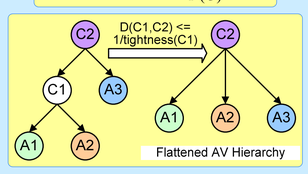
Learning AV Hierarchies

- Attribute values are extracted from the query list.
- Clustering similar attribute values leads to finding similar selection queries based on the similarity of their answer distributions over the sources.

$$d(Q1,Q2) = \sqrt{\sum_i [P(\hat{S}_i \mid Q1) - P(\hat{S}_i \mid Q2)]^2}$$

- The AV Hierarchies are generated using an agglomerative hierarchical clustering algorithm.
- They are then flattened according to their tightness.

$$tightness(C) = \frac{1}{\sum_{Q \in C} \frac{P(Q)}{P(C)} d(Q,C)}$$



Discovering Frequent Query Classes

- Candidate frequent query classes are identified using the anti-monotone property.
- Classes which are infrequently mapped are then removed.

Learning Coverage and Overlap

Coverage and overlap statistics are computed for each frequent query class using a modified Apriori algorithm.

$$P(\hat{S} \mid C) = \frac{\sum_{Q \in C} P(\hat{S} \mid Q) P(Q)}{P(C)}$$

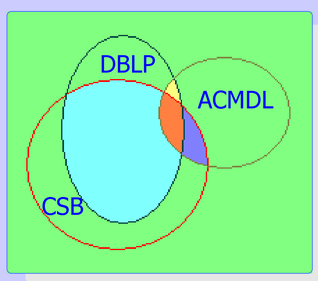
Using the Learned Statistics

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 mapped classes.

3. Data sources are ranked and called in order of relevance using the estimated statistics.
   In particular:
   - The most relevant source has highest coverage
   - The next best source has highest *residual* coverage

As a result, the maximum number of tuples are obtained while the least number of sources are called.



Example:
Here, CSB has highest coverage, followed by DBLP. However, since ACMDL has higher residual coverage than DBLP, the top 2 sources that would be called are CSB and ACMDL.

Effects of Learned Statistics on BibFinder

Purpose of the experiments:
- Analysis of space consumption
- Estimation of the accuracy of the learned statistics
- Evaluation of the effectiveness of those statistics in BibFinder.

Query planning algorithms used in the experiments:
- Random Select (RS): without any stats.
- Simple Greedy (SG): only coverage stats.
- Greedy Select (GS): coverage and overlap stats.

Precision of a plan: fraction of sources in the estimated plan which are the actual top sources.