Autoplex: Automated Discovery of Content for Virtual Databases

Jacob Berlin and Amihai Motro

Proceedings of CoopIS 2001, Sixth IFCIS International Conference on Cooperative Information Systems, pages 108-122.

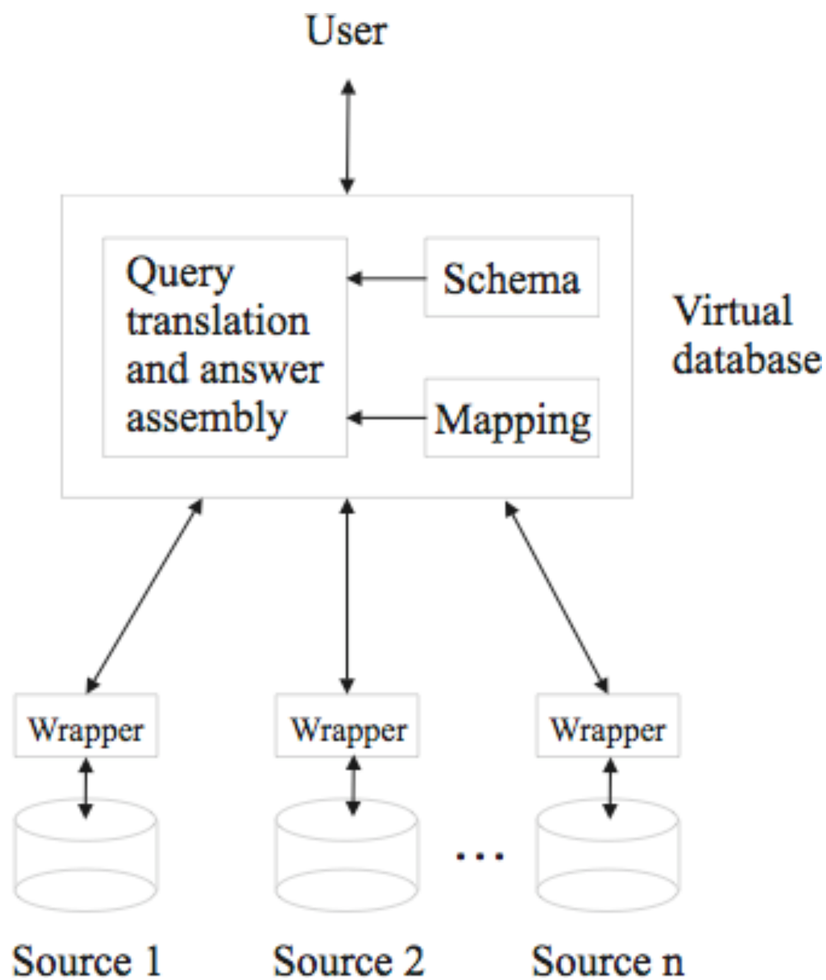
Outline

- Virtual databases
- The problem and our overall approach
- Basic assumptions and formal statement of the problem
- Autoplex architecture: Leaner and classifier
- The discovery (classification) methodology
- Learner details (skipped)
- Classifier details (skipped)
- Validation methodology
- Implementation and experimentation
- Conclusions

Virtual Databases

- Integration of information from multiple information sources
  - Provide flexible and efficient access to multiple sources.
  - Sources are independent, distributed, heterogeneous, overlapping.
- A common approach: *virtual databases*:
  - A single *global scheme* models the information contained in the entire collection of sources (or much of it).
  - The global scheme is *mapped* into the schemes of the member databases.
  - Global *queries* are translated (using the information in the mapping) to queries on the member databases.
  - The answers are combined to form an answer to the global query.
  - The process is *transparent* to the users of the system.

Typical Architecture



Problem

- Current systems *do not scale up* to environments of very large number of sources:
  - Incorporating new member databases is a manual process which is *complex and costly*.
  - Hence, the current paradigm is useful only when the number of member databases is *small and stable*.

Our Approach

- We developed a system, called Autoplex, that *discovers* member schemes and incorporates them “automatically” into the global scheme.
- Based on Bayesian learning, Autoplex acquires probabilistic knowledge from examples that *have already been integrated* into the virtual database.
- It then uses this knowledge to discover content contributions in new, previously unseen sources.

Basic Assumptions

- Autoplex adopts the Multiplex framework for virtual databases.
  - The mapping is a list *contributions*.
  - Each contribution is a pair of views: (*global view*, *local view*).
  - The local view is a *materialization* of the global view.
- For complexity reasons, Autoplex restricts the view expressions:
  - The global view is a single relation.
  - The local view is a selection-projection of a single relation.

Statement of the Problem

Given:

- A relation scheme $R = (X_1, \ldots, X_n)$.
  - This is the virtual database.
  - Each column X_i is labeled *required* or *optional*.
- A set of contribution *examples*, each consisting of
  - A relation scheme $S = (Y_1, \ldots, Y_k)$
  - A relation instance s of scheme S.
  - A selection-projection expression e that defines a *contribution* to R.
- A new previously unseen
  - Relation scheme $T = (Z_1, \ldots, Z_m)$.
  - A relation instance t of scheme T.
  - This is the *candidate* relation.

Determine:

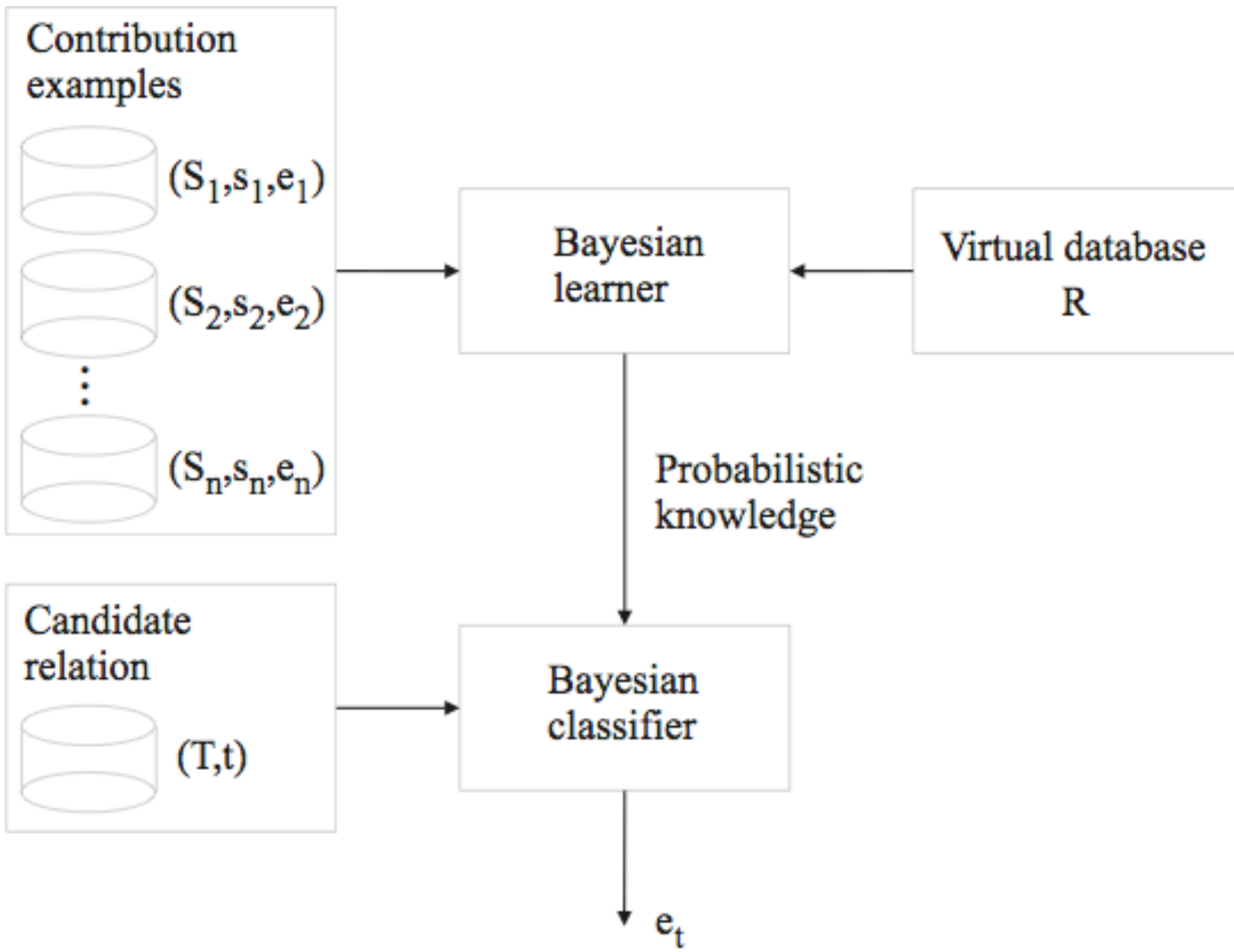
- Does T contain an acceptable contribution to R?
  - If so, find the expression e_t that defines it.
  - An *acceptable* contribution: Satisfies all the required columns of R while exceeding a predetermined threshold.

Autoplex Architecture

Two main components:

- *Learner*:
  - Input: the virtual scheme *R* and the contribution examples *S* (items 1 and 2 above).
  - Acquires probabilistic knowledge on features of the examples.
  - Stores this knowledge on secondary storage for future use.
- *Classifier*:
  - Input: A scheme *T* and instance *t* (item 3 above).
  - Uses the acquired knowledge to infer a *selection-projection* view that defines a contribution of *T* to *R*.

Autoplex Architecture



Classification Methodology

Classification consists of 4 steps:

1. Consider each column *Zi* of *T* and each column *Xi* of *R* and determine the probability that *Zi* is an instance of *Xi*.
2. Find an assignment of the local columns to the global columns that maximizes total column probabilities.
   - At this point (if successful) we found the best *projection* of T.
3. Prune the rows of the projection to retain only rows that “resemble” rows in the examples.
4. Partition the instance *t* into two sets of rows:
   - Those to be included in the contribution, and those to be excluded.

   Obtain at intensional (predicate) description of the included rows.
   - Use a classification tree algorithm.

The final result is a selection-projection expression that extracts a contribution from *T* (or *false*, if an acceptable contribution could not be found).

Classification Methodology (*Cont.*)

- Our approach is a *compromise* between
  - More powerful searching (that will discover more and better contributions), and
  - The need to keep the problem tractable.

- Two examples of our compromise:
  1. Our search for an expression is "greedy":
     - We search for a projection followed by a selection.
     - We might do better if some rows were removed first!
  2. Better projections could be found if we allowed value transformations first
     - e.g., unit conversions.

Validation Methodology

Autoplex output can be viewed as four Boolean decisions:

1. *Column Mapping*: for each combination of candidate column and virtual column, decide whether they match.
2. *Table mapping*: For each combination of candidate table and virtual table, decide whether they match.
3. *Tuple partitioning*: For each tuple in the candidate table, decide whether to assign it to the contributing set.
4. *Tuple selection*: (after the predicate was inferred from the partition) For each tuple, decide whether it satisfies the selection predicate.

Validation Methodology (*Cont.*)

- In each decision, the output falls into four disjoint categories.:
  - *A*. *True positives*: Decision is *true* and correct answer is *true*.
  - *B*. *False negatives*: Decision is *false* and correct answer is *false*
  - *C*. *False positives*: Decision is *true* but correct answer is *false*.
  - *D*. *False negatives*: Decision is *false* but correct answer is *true*.
- The ratio $|A|/(|A|+|C|)$ measures
  - The proportion of true positives among the cases thought to be positive.
  - The accuracy of Autoplex when it decides *true*.
  - The *soundness* of the content discovered.
- The ratio $|A|/(|A|+|B|)$ measures
  - The proportion of positives detected among the complete set of positives.
  - The ability of Autoplex to detect positives.
  - The *completeness* of the discovery process.

Implementation and Experimentation

- The methods developed were implemented in Java (though not integrated with a complete virtual database system).
- For the experiment, a virtual database was defined with 3 relations on computer retail information.
- Data from 15 retailers was collected off-line and imported into relational tables (the candidates).
- These 15 sources were mapped (by "expert") onto the 3 virtual relations with a total of 21 mappings.
- *Stratified threefold cross-validation*:
  - The 21 mappings were partitioned into 3 "folds".
  - Two folds were used for learning and one for testing (the "expert" mapping was assumed correct).
  - The experiment was repeated 3 times (for the 3 possible combinations of folds).

Experimentation (*Cont.*)

Decision Category	**Soundness**	**Completeness**
Column mapping	0.81	0.81
Table mapping	1.00	0.86
Table partitioning	0.89	0.94
Tuple selection	0.90	0.94

Conclusion

- A novel approach aimed at reducing the cost and complexity of incorporating new sources into the global system.
- Initial results encouraging.
- Some research issues:
  – Support more general views:
    - Allow local ad global views that involve *joins*.
    - Discover content that becomes suitable after an appropriate *transformation*.
  – Use intensional information (e.g., constraints):
    - With extensional features, discoveries were based on “similarity” to example data.
    - With intensional features, discoveries would also be based on satisfaction of constraints.