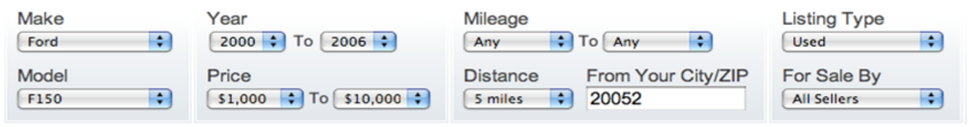
Breaking the Top-k Barrier of Hidden Web Databases

Saravanan Thirumuruganathan[1], Nan Zhang[2] and Gautam Das[1]
[1]University of Texas at Arlington and Qatar Computing Research Institute, [2]George Washington University

Hidden Web Databases



- Large part of hidden web
- *Form* input interface, top-k output constraint and limits over number of queries issued.

Motivation

Figure: Web Aggregator for Flight Search

- Hidden Web Databases to query from : FlightStats (On-time record), SeatGuru (Leg room), US DOT (Lost luggage record), TSA (Security wait time) etc
- Need to break top-k barrier!

GetNext for Hidden Web Databases

- **Goal :** Given the top-h tuples ($h \geq k$), retrieve the tuple with rank $h+1$ using public interface of hidden web database while minimizing query cost
- Possibly unknown static ranking function
- Challenging for hidden web databases

Approach

- **Idea :** Issue related queries and infer next ranked tuple from results
- *Direct* domination between tuples can be found with a *single* query
- Leverage rank related information disclosed by each query over its matched top-k and non top-k tuples

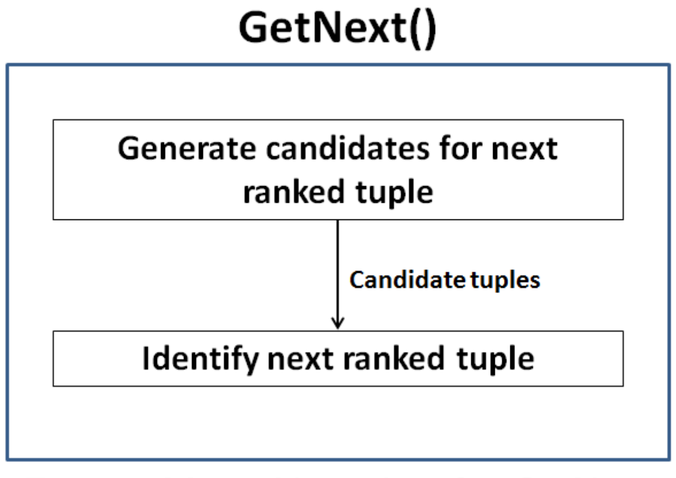


Figure: Meta Algorithm for GetNext

Candidate Generation

- Given top-h tuples, produce a *minimal* set of candidates *guaranteed* to contain next ranked tuple
- Partition top-h tuples using queries that jointly *cover* entire database
- *Dominance directed* graph from query results

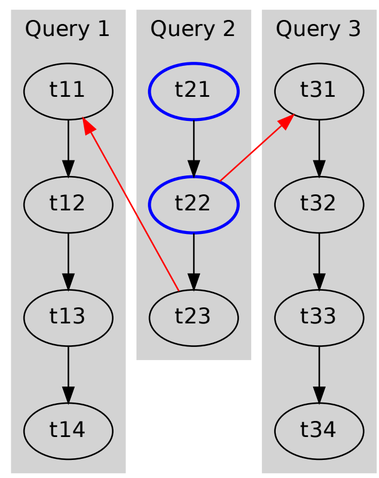


Figure: An example dominance graph with t21,t22 are next ranked tuples

Candidate Testing

- Identify the next ranked tuple from candidate set
- **Idea :** Next ranked tuple t can only be dominated by top-h tuples
- A **beyond-h query** is one that matches t and
  - return atleast one non top-h tuple and
  - Any query with subset of its predicates return only tuples from top-h
- Mapping between beyond-h queries and minimal infrequent itemsets

Experimental Results

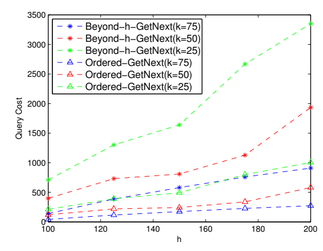


Figure: Varying h and k

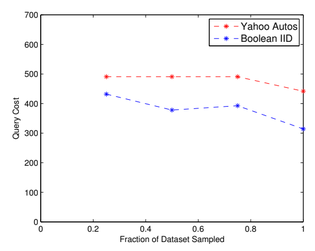


Figure: Varying Database Size

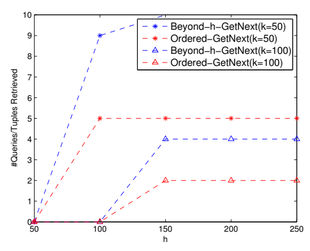


Figure: Retrieving Top-250 Amazon DVD Titles

Selected References

- D. J. Haglin and A. M. Manning, "On minimal infrequent itemset mining", in ICDM, 2007.
- N. Bruno, L. Gravano, and A. Marian, "Evaluating top-k queries over web-accessible databases" in ICDE, 2002.
- I. Ilyas, G. Beskales, and M. Soliman, "A survey of top-k query processing techniques in relational database systems" ACM Computing Surveys, vol. 40, 2008.