Prob-k – Probabilistic Candidate Pruning

- **Prob-conservative**
  - A family of algorithms based on invariant:
    \[
    \sum_{c \in C} \text{score}(c, d) \leq \sum_{c \in C} \text{high}_i(c) + \sum_{\text{low}_i(c)}
    \]
  - If \( \text{worstscore}(d) < \min-k \), add \( d \) to top-k
  - If \( \text{bestscore}(d) \leq \min-k \), drop \( d \) from queue

- **Prob-progressive**
  - A queue per query & use bestscore as priority
  - Merge candidates by their best-scores
  - No dominating candidate in terms of score prediction
  - Update and rebuild entire queue periodically
  - Stop heuristically, if
    \[
    \text{size}(I) \text{ candidates} \leq \varepsilon
    \]

- **Prob-aggressive**
  - No queue
  - Efficient hash joins only
  - Consider virtual candidate \( d \), with \( E(d) = \emptyset \)
  - \( d \) dominates all yet unseen candidates
  - Stop heuristically, if
    \[
    \text{size}(I) \text{ candidates} \leq \varepsilon
    \]

**TopX – Efficient Support for XML IR**

- Supports XPath 2.0 Full-Text and NEXI query languages
- Probabilistic candidate pruning
- Combined aggregated score- and XML-specific selectivity predictor
- XML-specific scoring model
- Okapi BM-25 extension for semistructured data using individual element statistics
- Supports multiple index structures
- Pre/Post-Order or Data Guides
- Dynamic query rewriting wrt. to individual predicate selectivities

**Experimental Setup**

- Aquaint News Corpus: 528,155 docs; 86,000,000 tuples; 1.9 GB
- 50 “hard” queries taken from the TREC Robust track 2004

**Experiments**

- **Prob-k Pruning Performance**
  - Incremental Merge vs. Static Extensions
  - M: \( 1 \leq 118 \), k: \( 10^5 \) for MAP
  - k = 1,000 for MAP

- **Expansion Performance**
  - Incremental Merge vs. Static Extensions
  - M: \( 1 \leq 118 \), k: \( 10^5 \) for MAP
  - k = 1,000 for MAP