ProUD: Probabilistic Ranking in Uncertain Databases

Matthias Renz,
Thomas Bernecker and Hans-Peter Kriegel
Ludwig-Maximilians-Universität München
Munich, Germany
www.dbs.ifi.lmu.de

Outline

- **Introduction**
- Ranking on Uncertain Data
- Probabilistic Ranking Algorithm
- Experimental Evaluation
- Summary

Introduction

- Ranking Queries
  - are very important for similarity search
  - give the most relevant answers first
  - are more flexible than distance range and *k*NN queries
- Applications:
  - person identification
  - similarity search in multimedia databases
  - …

➔ often uncertainty in data

Introduction

- uncertain data caused by
  - continuously changing data
  - concurring data descriptions
  - measurement inaccuracies
- uncertain data leads to uncertain query results

  ➔ probabilistic queries

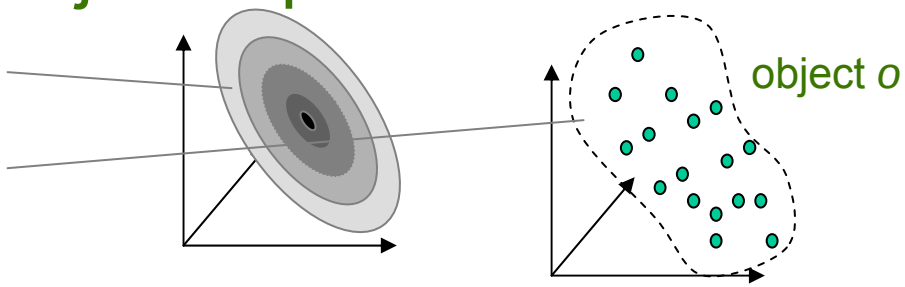
  (e.g. probabilistic ranking)
- results are associated with confidence values

Outline

- Introduction
- **Ranking on Uncertain Data**
- Probabilistic Ranking Algorithm
- Experimental Evaluation
- Summary

Ranking on Uncertain Data

- Positionally Uncertain Data
  - vector data in *d*-dimensional space $\Re^d$
  - no unique position in $\Re^d$
  - objects are represented by
    - **multiple *d*-dimensional vectors**
    - **that are mutually exclusive**
    - **a confidence value is assigned to each vector**
  - types of uncertain object representations
    - **pdf (continuous)**
    - **<u>discrete samples</u>**



Ranking on Uncertain Data

- Distance Computation for Uncertain Objects
  - two uncertain objects

    $o_i = \{o_{i,1},\ldots,o_{i,m}\}$ and $o_j=\{o_{j,1},\ldots,o_{j,m}\}$
  - distance between o_i and o_j:

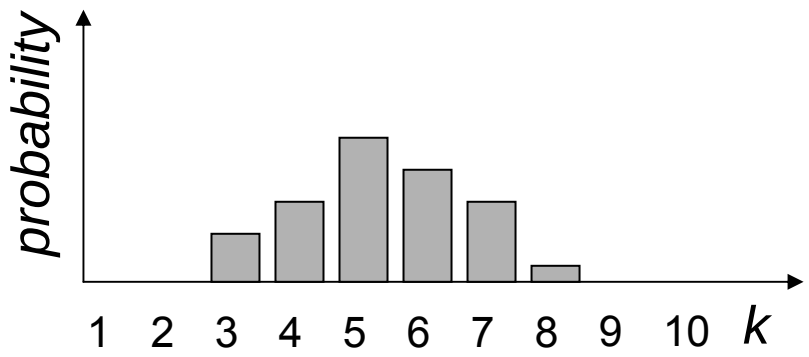
    $$d_{uncertain}(o_i, o_j) = \{dist(o_{i,m}, o_{j,n}) \mid 1 \le m \le M, 1 \le n \le M\}$$
  - probability that the distance between o_i and o_j is less than $\varepsilon \in \Re_0^+$:

    $$P(d_{uncertain}(o_i, o_j) \le \varepsilon) = \frac{|\{d \in d_{uncertain}(o_i, o_j) : d \le \varepsilon\}|}{|d_{uncertain}(o_i, o_j)|}$$

Ranking on Uncertain Data

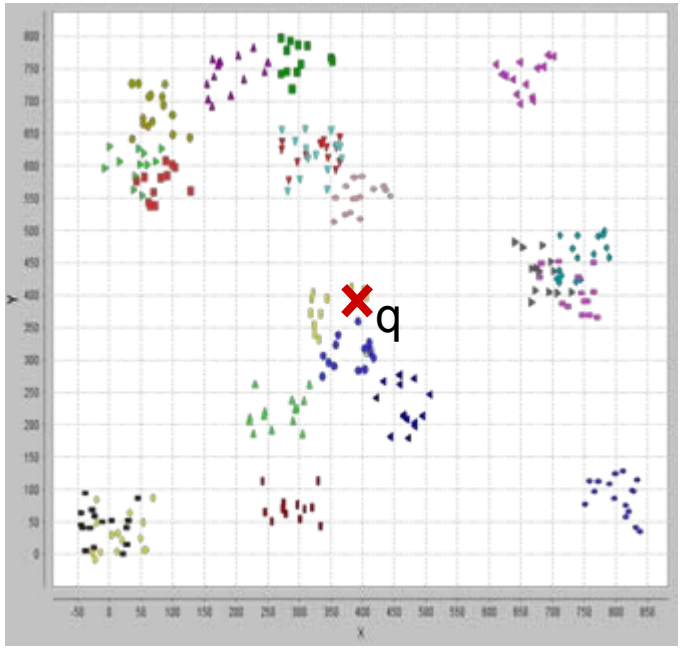
- Probabilistic Ranking
  - function $prob\_ranked_q$: $\mathcal{D} \times \{1,..,N\} \rightarrow [0..1]$
  - $prob\_ranked_q(o,k)$ reports the probability that object o is the kth-nearest-neighbor of the query point q

    ➔ for each object:

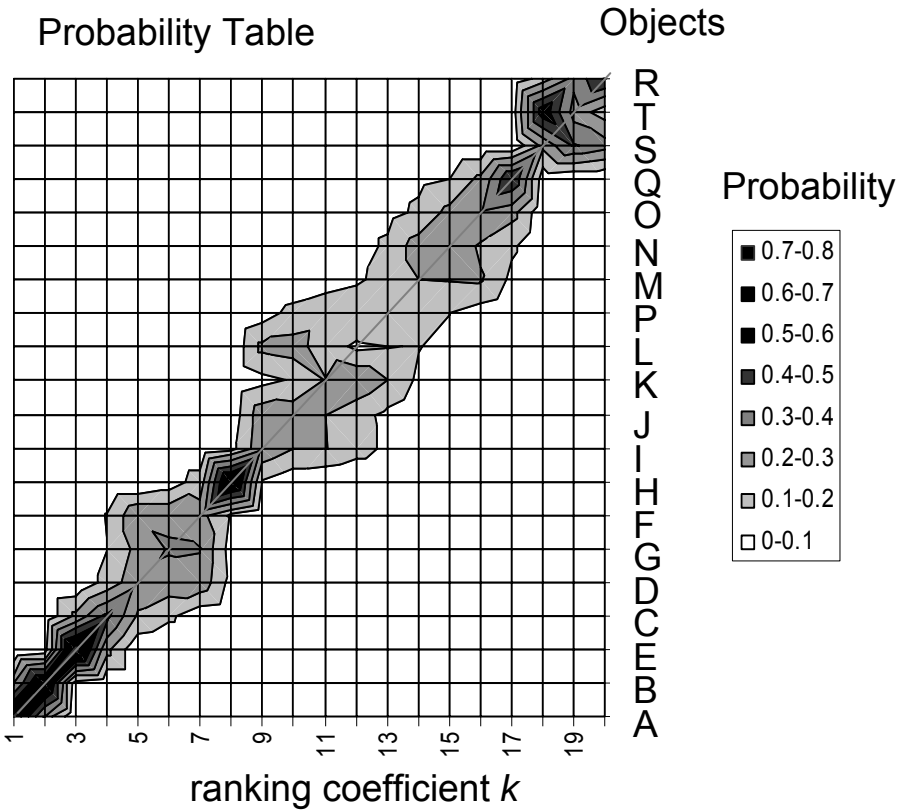


Ranking on Uncertain Data

- Probabilistic Ranking Output:
  - Example:



vector space



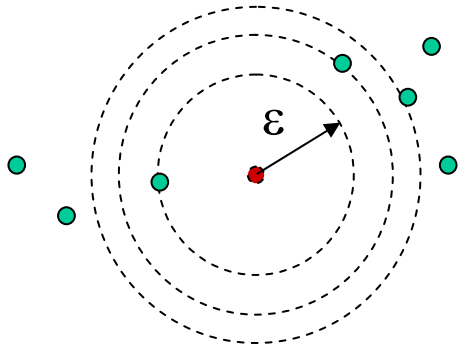
probabilistic ranking output

Outline

- Introduction
- Ranking on Uncertain Data
- **Probabilistic Ranking Algorithm**
- Experimental Evaluation
- Summary

Probabilistic Ranking Algorithm

- Iterative Probability Computation
  - ranking query on the object samples



radial sweep with increasing range ε

  - during the radial sweep: maintain for each object *o* the probability $P_o = P(d_{uncertain}(o,q) \leq \varepsilon)$
  - for each accessed sample $o_{i,j}$, compute the probability $P(o_{i,j},k)$ that exactly (k-1) objects $o \neq o_i$ are within the ε-range, for k = 1..N.

Probabilistic Ranking Algorithm

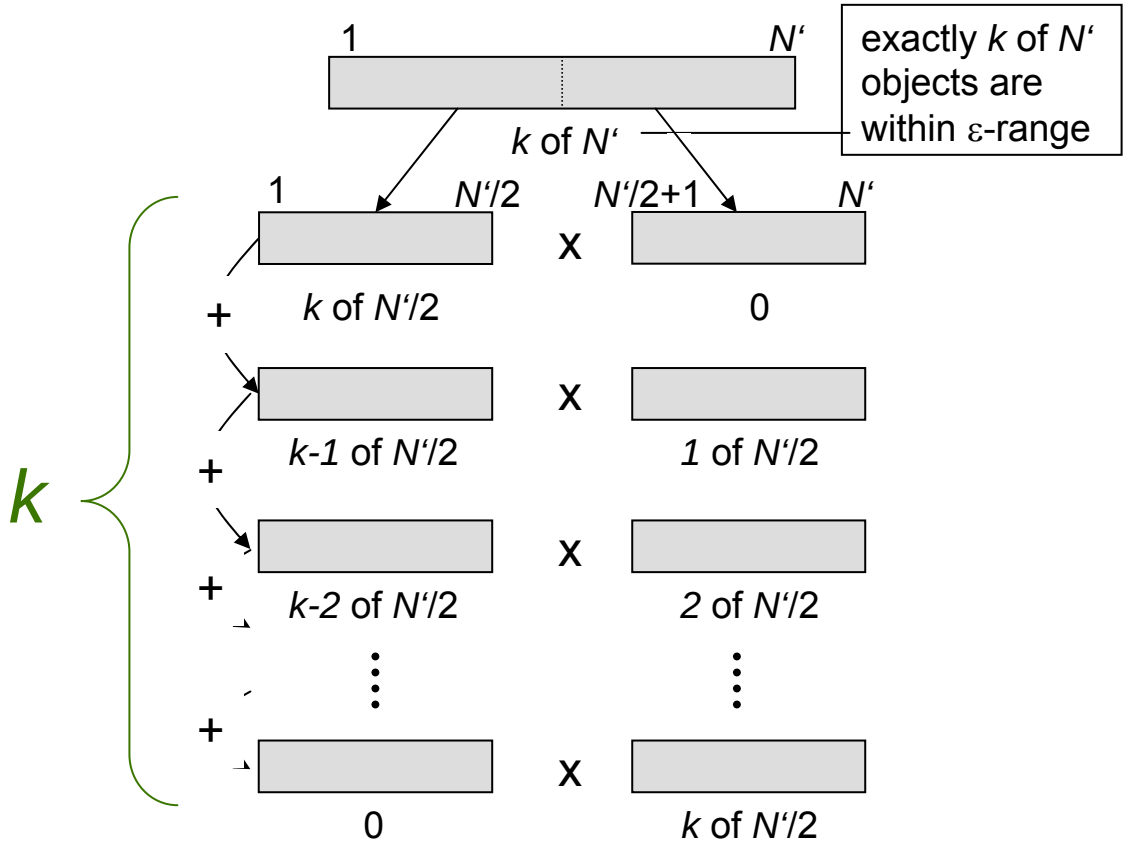
- computation of $P(o_{i,j},k)$:

$$P(o_{i,j},k) = \sum_{\sigma_k(i) \in S_k} \prod_{\substack{l=1..N \\ l \neq i}} \begin{cases} P_{o_l} & if\ o_l \in \sigma_k(i) \\ (1-P_{o_l}) & else \end{cases}$$

- problem: a lot of possibilities to select the set $\sigma_k(i)$ of (k-1) objects out of N objects
- reduce $N \rightarrow N'$ by pruning objects that are beyond ε

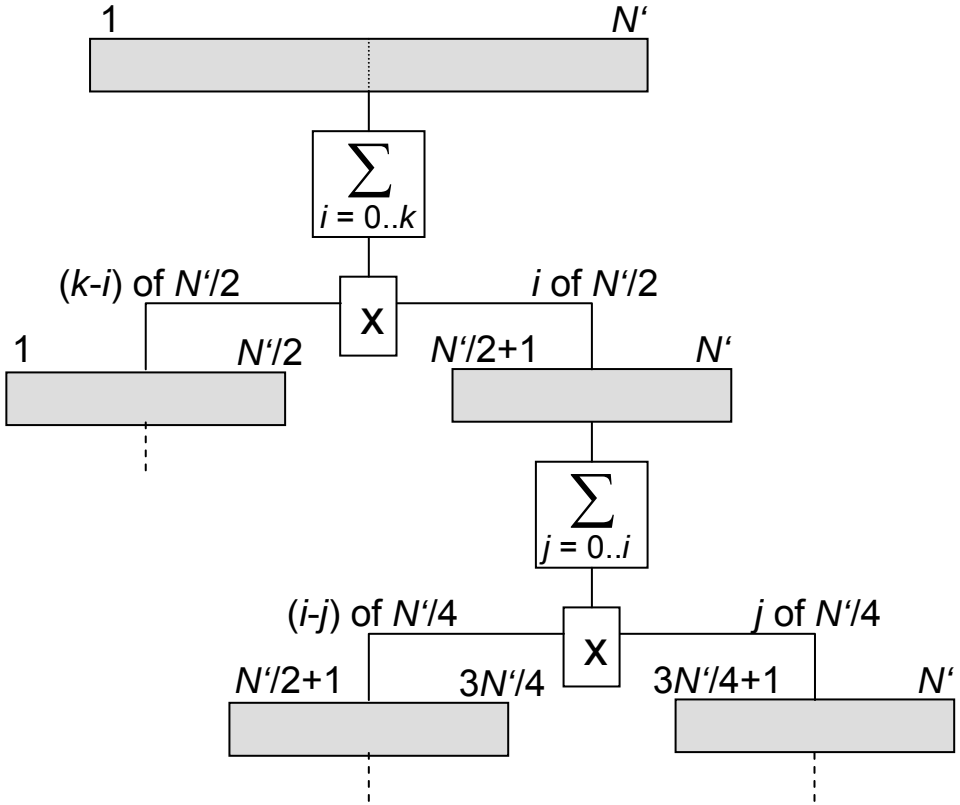
Probabilistic Ranking Algorithm

- Accelerated Probability Computation
  - divide and conquer approach



Probabilistic Ranking Algorithm

- Accelerated Probability Computation
  - divide and conquer approach

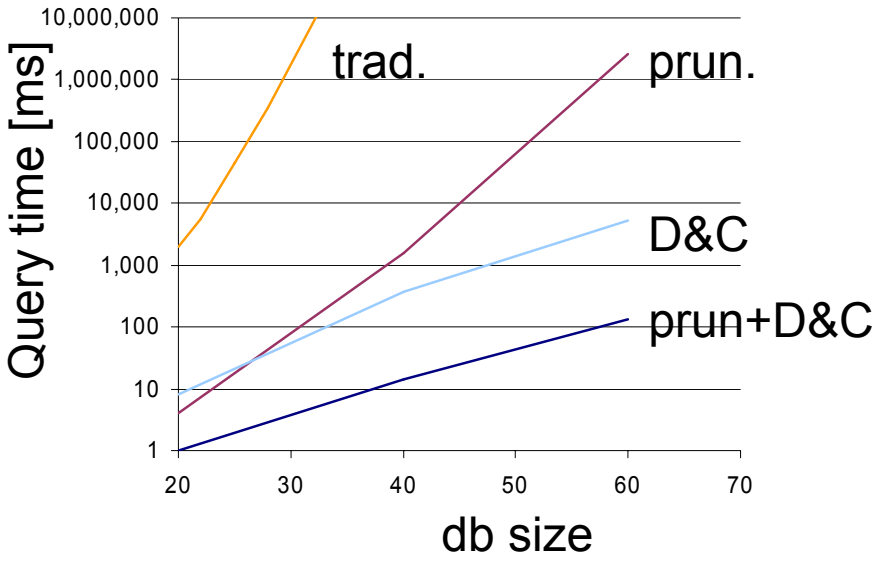


Outline

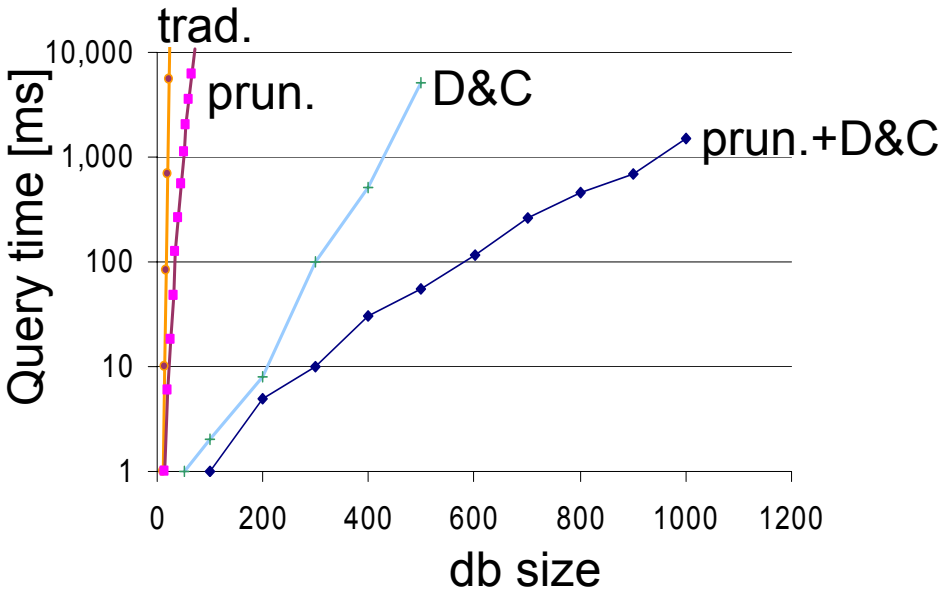
- Introduction
- Ranking on Uncertain Data
- Probabilistic Ranking Algorithm
- **Experimental Evaluation**
- Summary

Experimental Evaluation

- Performance w.r.t. database size



higher degree of uncertainty

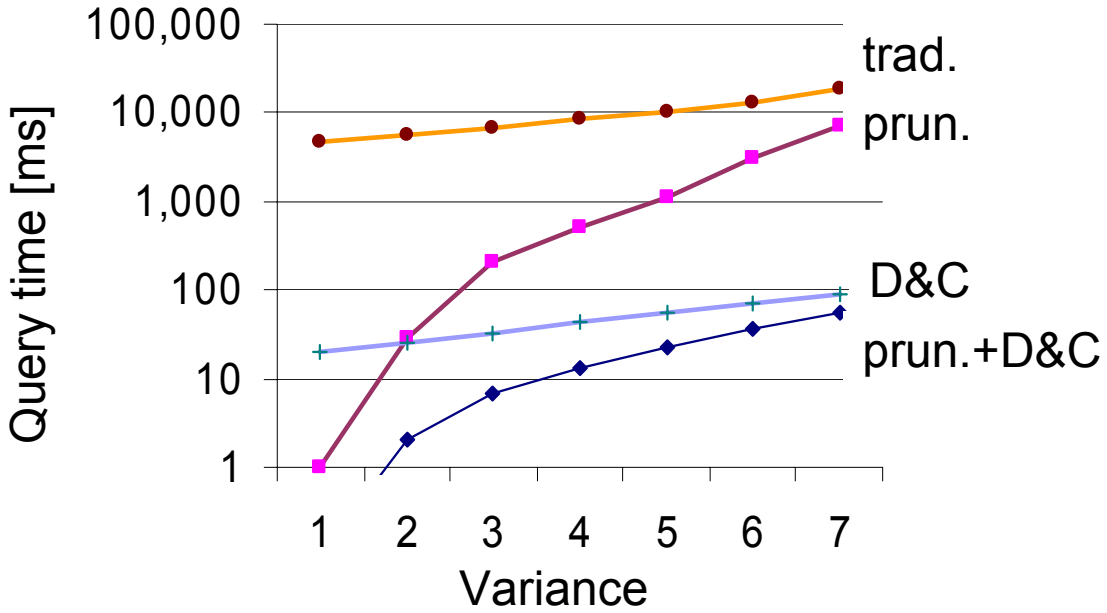


lower degree of uncertainty

– performance depends on the degree of uncertainty (object variance)

Experimental Evaluation

- Performance w.r.t. degree of uncertainty



  – speed-up by pruning degenerates for higher object variances (object uncertainty)

Outline

- Introduction
- Ranking on Uncertain Data
- Probabilistic Ranking Algorithm
- Experimental Evaluation
- **Summary**

Summary

- approach to accelerate probabilistic ranking queries
- very high speed-up factor (2 orders of magnitude)

- Future Work:
  - comparison to the latest top-*k* query approaches

Thank you,

for your attention,

any questions?