

Probabilistic Similarity Join on Uncertain Data

DASFAA 2006

Hilton Hotel, Singapore

Hans-Peter Kriegel, Peter Kunath, Martin Pfeifle,
Matthias Renz

 Database Group

Institute for Computer Science
University of Munich, Germany

Outline of the Talk



1. Introduction
- 2. Non-Probabilistic Approach**
3. Probabilistic Approach
4. Experimental Evaluation
5. Conclusions and Future Work

Similarity Join

Distance-Range Join (ε - join)

Let R and S be two sets, $\varepsilon \geq 0$,
and let $d : R \times S \rightarrow \mathbb{R}_0^+$ be a distance function.

The distance range join of R and S is the set
 $R \bowtie_{\varepsilon} S := \{(r, s) \in R \times S : d(r, s) \leq \varepsilon\}$

→ join algorithms are often based on index structures
(e.g. R-tree spatial join)

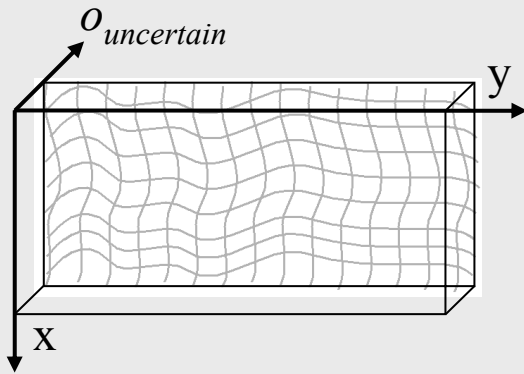
Uncertain Object Representation

Let $o \in O \subseteq \mathbb{R}^d$ be an object from a database.

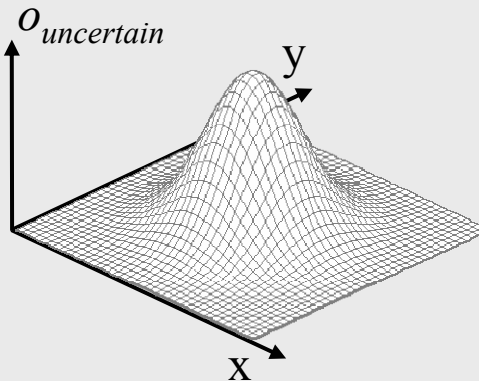
An uncertain object representation is a function

$o_{uncertain}: \mathbb{R}^d \rightarrow \mathbb{R}_0^+ \cup \infty$, for which holds: $\iint_{\mathbb{R}^d} o_{uncertain}(v) dv = 1$

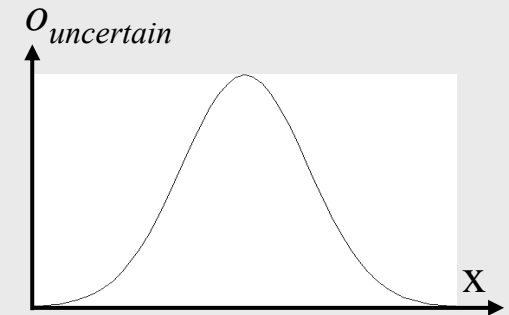
Uncertain Object Descriptions



Feature vectors



Moving objects



Sensor data

Distance Function

→ traditional join algorithms require distance functions which express the similarity between two objects by exactly one numerical value

Probabilistic Distance Function

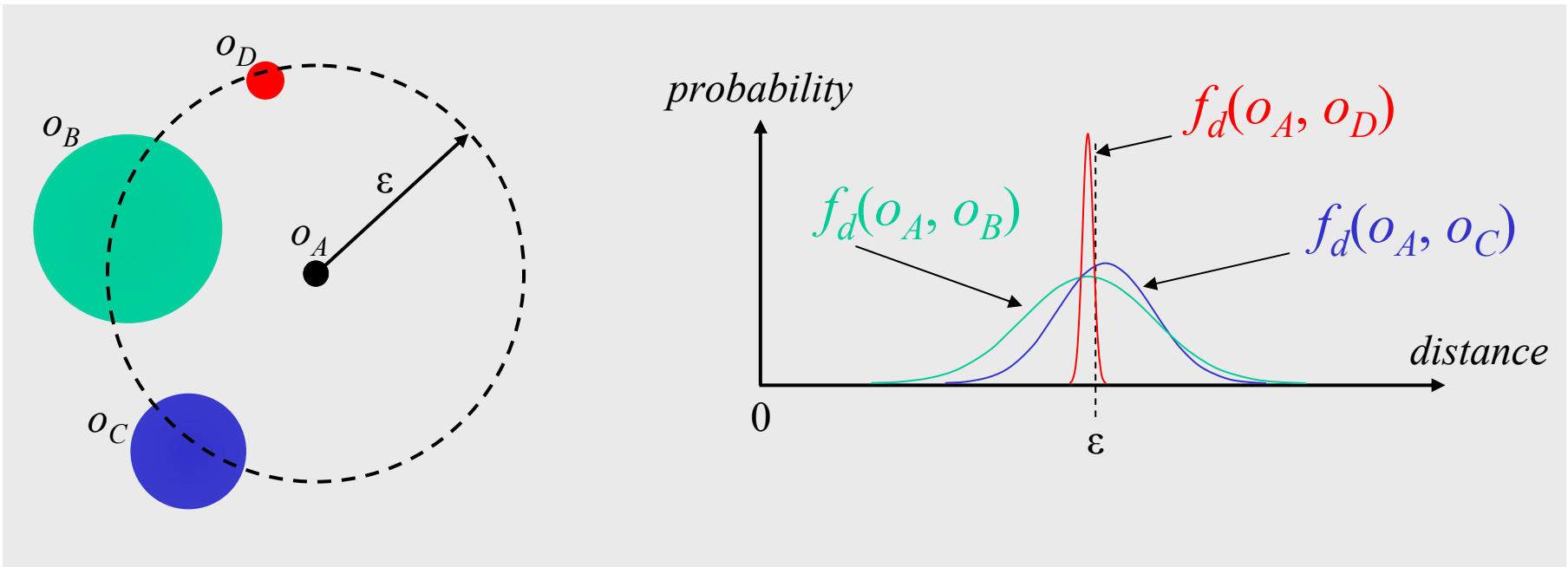
Let $d : O \times O \rightarrow \mathbb{R}_0^+$ be a distance function, and let $P(a \leq d(o, o') \leq b)$ denote the probability that $d(o, o')$ is between a and b .

Then a probabilistic density function f_d is called a probabilistic distance function if the following condition holds:

$$P(a \leq d(o, o') \leq b) = \int_a^b f_d(o, o')(x) dx$$

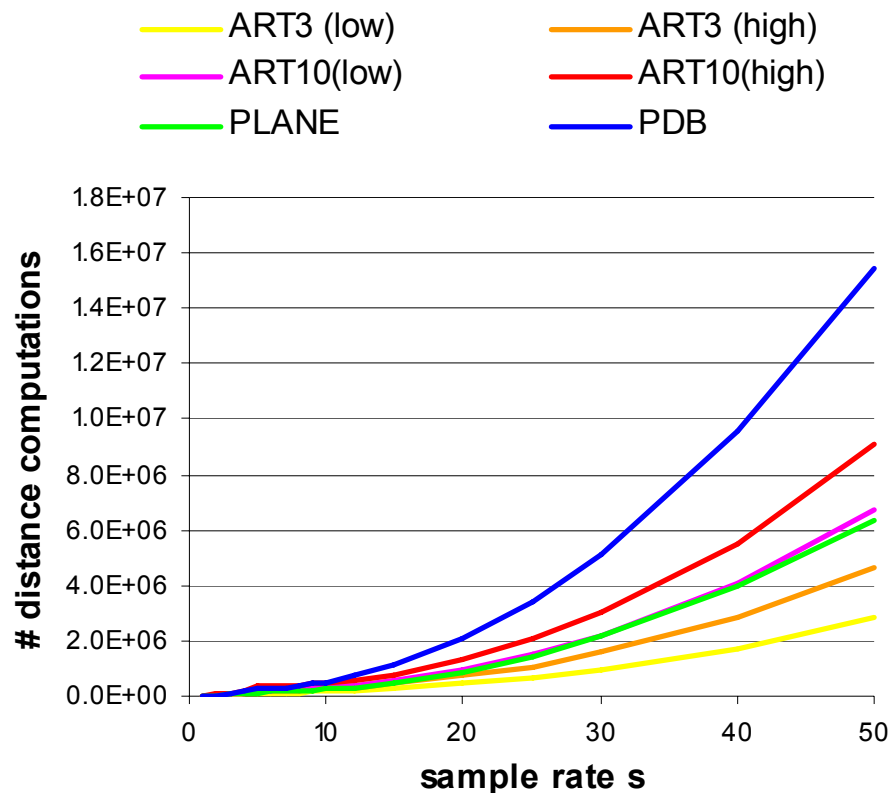
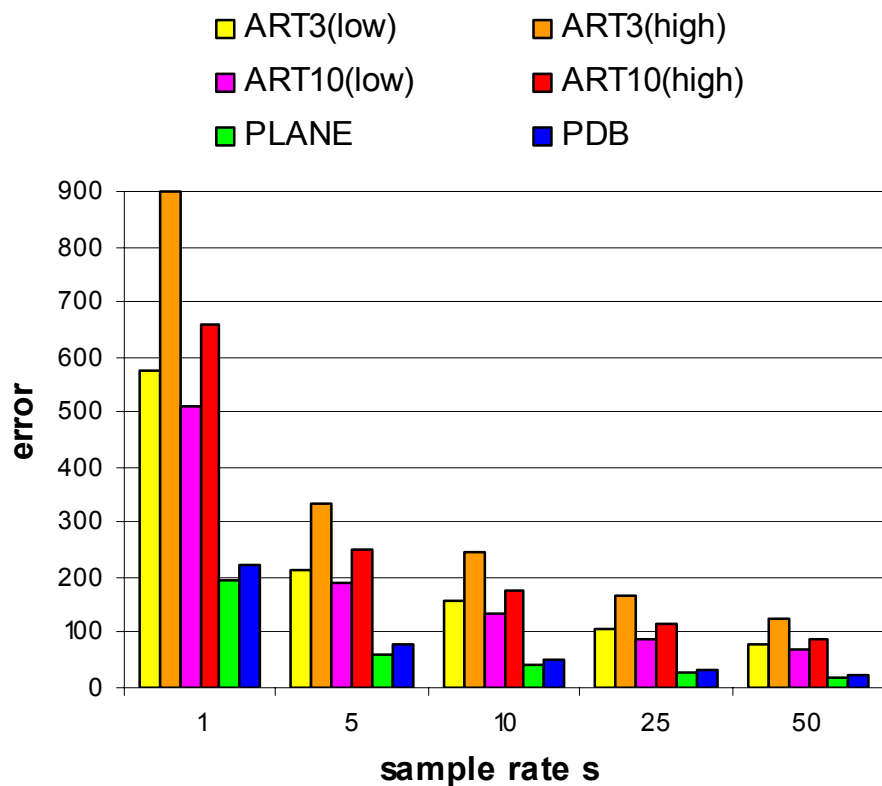
Distance-Range Join Example

distance-range join based on the expected distance



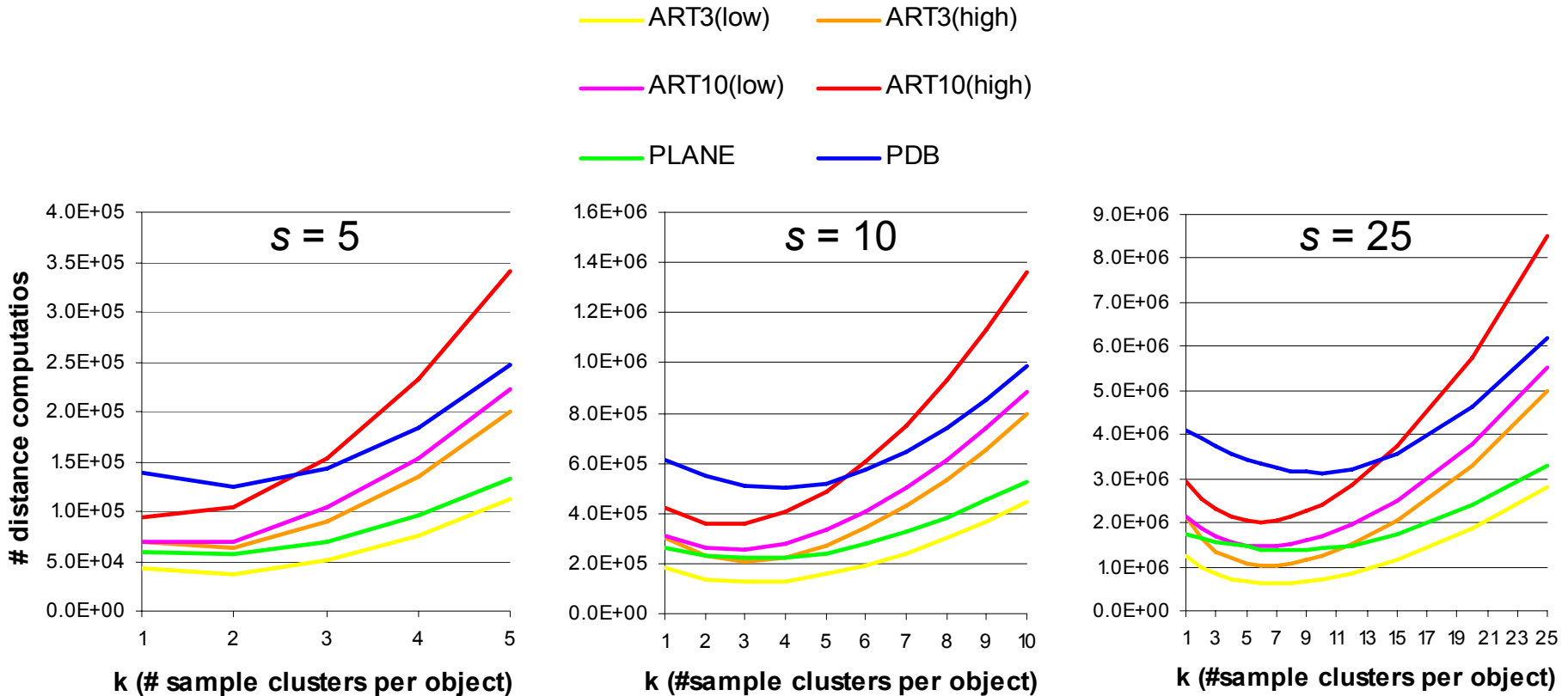
→ similarity joins based on the expected distance fail to produce meaningful results

Influence of the Sample Rate s



→ a sample rate of $s = 25$ leads to a good trade-off between cost and quality of the result set

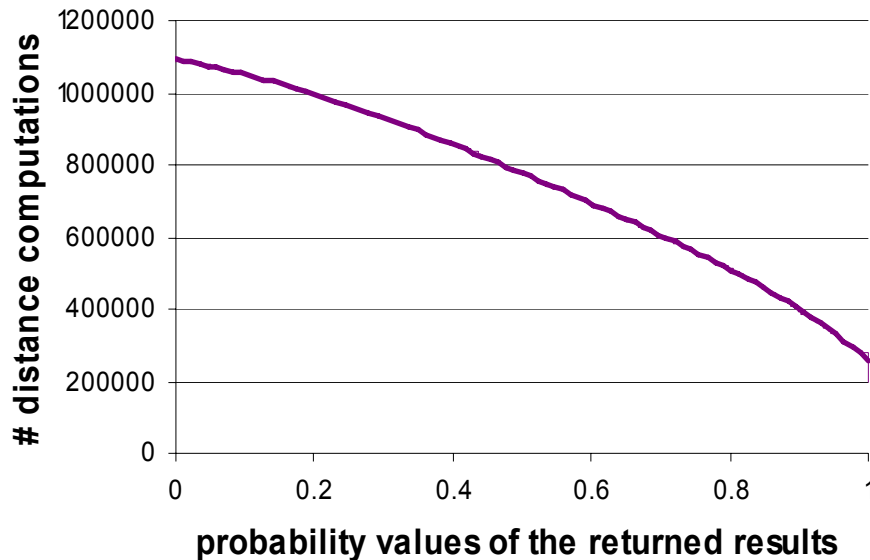
Performance for varying Number of Clusters k



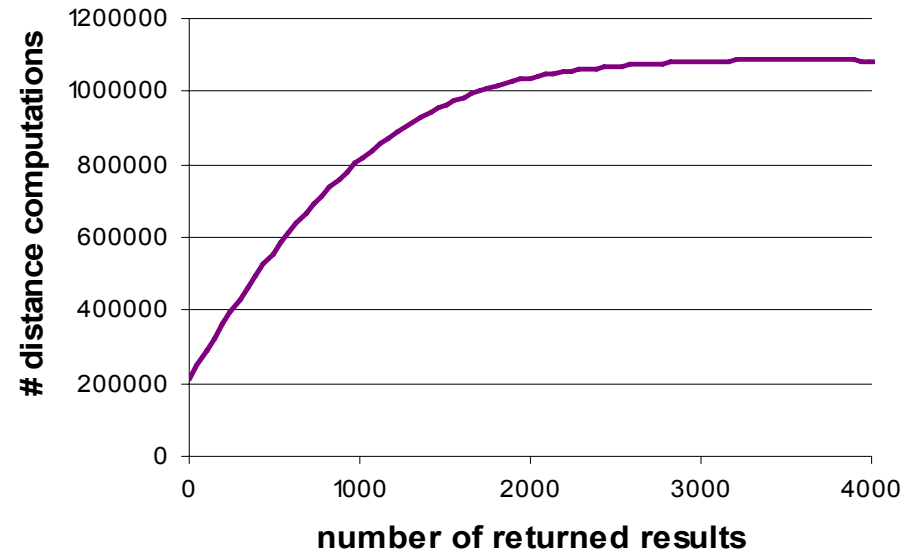
→ the optimal setting for k depends on the sample rate s

Performance of the ranked Distance-Range Join for ART3(high)

Performance with respect to the result set probability



Performance with respect to the number of returned results



→ the proposed incremental join processing is particularly useful when the user is interested in a small portion of the result set

Conclusions and Future Work

Our Approach:

- Assigns to each object pair a probability value indicating the likelihood that it belongs to the result set
- Effective computation based on the concept of monte-carlo sampling
- The incremental probabilistic distance-range join allows to report the results very early

Future Work:

- Extend our approach to other similarity join predicates, e.g. nearest neighbor predicate
- Apply our approach to data mining algorithms, e.g. clustering and classification

