

	LUDWIG- MAXIMILIANS- UNIVERSITÄT MÜNCHEN	 DEPARTMENT INSTITUTE FOR INFORMATICS	 DATABASE SYSTEMS GROUP	MCIS 2009 Workshop
---	---	--	--	--------------------



  

## Probabilistic Ranking in Uncertain Vector Spaces

Thomas Bernecker, Hans-Peter Kriegel,  
Matthias Renz and Andreas Zuefle


Ludwig-Maximilians-Universität München  
Munich, Germany  
[www.dbs.ifi.lmu.de](http://www.dbs.ifi.lmu.de)



	<b>Outline</b>	
---	----------------	---


- **Introduction**
- **Uncertainty Model**
  - modelling uncertain vector data
  - probabilistic query processing
- **Probabilistic Similarity Ranking**
  - probabilistic ranking definition
  - approaches for efficient ranking processing
- **Summary**

M. Renz: Probabilistic Ranking in Uncertain Vector Spaces 2




DATABASE  
SYSTEMS  
GROUP

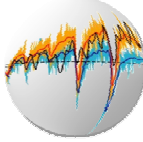
## Introduction




- modern database applications require efficient and effective query methods for
 



spatial-




temporal- and



multimedia data.
- often vague and imprecise attributes due to
  - incomplete data, imprecise monitoring, data manipulation for privacy preserving, inexact prediction/estimation, etc.


→ **uncertain databases**

M. Renz: Probabilistic Ranking in Uncertain Vector Spaces
3



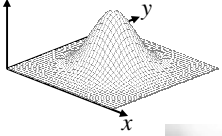
DATABASE  
SYSTEMS  
GROUP

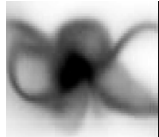
## Introduction




- Types of uncertain databases
  - Tuple uncertainty
    - uncertain relational databases
    - tuples with confidence
    - e.g. *Trio*, MayDBMS, ...
  - Attribute uncertainty
    - uncertain feature vectors
    - uncertain (similarity) distances
  - Spatial uncertainty
    - uncertain spatial extensions

ID	NAME	CONF
p1	john	0.6
p2	fred	0.3
p3	mary	0.7
...	...	...






M. Renz: Probabilistic Ranking in Uncertain Vector Spaces
4



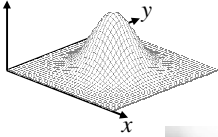
**Introduction**

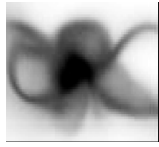


- **Types of uncertain databases**
  - Tuple uncertainty
    - uncertain relational databases
    - tuples with confidence
    - e.g. *Trio*, *MayDBMS*, ...
  - Attribute uncertainty
    - uncertain feature vectors
    - uncertain (similarity) distances
  - Spatial uncertainty
    - uncertain spatial extensions


ID	NAME	CONF
p1	john	0.6
p2	fred	0.3
p3	mary	0.7
...	...	...






M. Renz: Probabilistic Ranking in Uncertain Vector Spaces

5

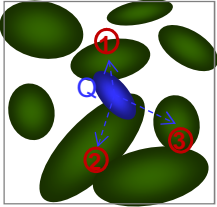


**Introduction**




- **Ranking Queries in Uncertain Data**
  - given:
    - database with uncertain vectors
    - (uncertain) query object(s)  $Q$
  - queries:
 



ranking query
  - challenges:
    - uncertain distances, uncertain query results


M. Renz: Probabilistic Ranking in Uncertain Vector Spaces

6



DATABASE  
SYSTEMS  
GROUP


## Outline



- Introduction
- **Uncertainty Model**
  - modelling uncertain vector data
  - probabilistic query processing
- Probabilistic Similarity Ranking
  - probabilistic ranking definition
  - approaches for efficient ranking processing
- Summary


M. Renz: Probabilistic Ranking in Uncertain Vector Spaces

7

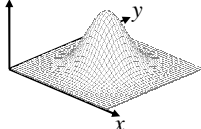


DATABASE  
SYSTEMS  
GROUP

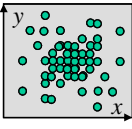
## Modelling Uncertainty in Feature Spaces



- Uncertain Vector Data
  - vector data in  $d$ -dimensional space  $\mathcal{R}^d$
  - objects are represented by
    - multiple  $d$ -dimensional vectors
    - that are mutually exclusive
    - a confidence value is assigned to each vector
  - types of uncertain vector object representations




pdf (continuous)




vector samples (discrete)

M. Renz: Probabilistic Ranking in Uncertain Vector Spaces

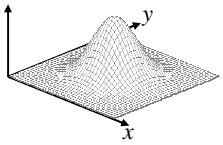
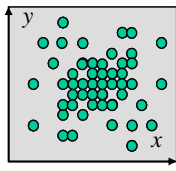
8



**Modelling Uncertainty in Feature Spaces**




- Representation of Uncertain Vectors
  - continuous representation  
(conform to *attribute uncertainty* model)
    - attributes associated with a pdf
    - defined within a spec. interval
    - often used with standard prob. dist.,  
e.g. gaussian-, uniform distribution
  - discrete representation  
(conforms to *x-relation* model)
    - set of vectors with probabilities
    - probabilities define discrete prob. dist.





M. Renz: Probabilistic Ranking in Uncertain Vector Spaces

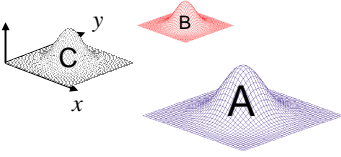
9

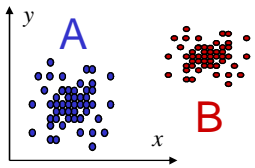


**Modelling Uncertainty in Feature Spaces**

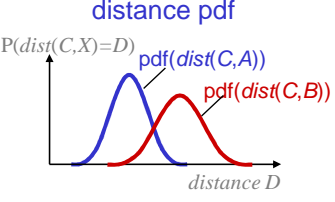


- probabilistic similarity distance:
  - uncertain object → uncertain similarity distance

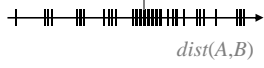




distance pdf




set of (possible) distances  
{..., (d<sub>i</sub>, p<sub>i</sub>), ...}




M. Renz: Probabilistic Ranking in Uncertain Vector Spaces

10




## Modelling Uncertainty in Feature Spaces




- probabilistic query predicate evaluation:
  - traditional (similarity) queries:
    - evaluation of query predicate  $\Phi_{sim} \in \{true, false\}$
  - probabilistic (similarity) queries:
    - evaluation of probability  $P(\Phi_{sim} = true) \in [0,1]$
  - probabilistic similarity query results:
 
$$P(\Phi_{sim} = true) = \sum_{w \in W' \subseteq W} P(w)$$
    - $W$  := set of all possible worlds (possible distances)
    - $W'$  := worlds in which query predicate  $\Phi_{sim} = true$
    - $P(w)$  := probability that world  $w$  is true

M. Renz: Probabilistic Ranking in Uncertain Vector Spaces
11




## Outline




- Introduction
- Uncertainty Model
  - modelling uncertain vector data
  - probabilistic query processing
- **Probabilistic Similarity Ranking**
  - **probabilistic ranking definition**
  - **approaches for efficient ranking processing**
- Summary

M. Renz: Probabilistic Ranking in Uncertain Vector Spaces
12



DATABASE  
SYSTEMS  
GROUP


## Probabilistic Similarity Ranking



- Ranking Queries
  - very important for similarity search applications
  - give the most relevant answers first
  - are more flexible than  $\mathcal{E}$ -range and NN queries
  
- probabilistic ranking queries
  - results are associated with confidence values
  - in contrast to  $\mathcal{E}$ -range / NN queries
    - no unique query predicate


M. Renz: Probabilistic Ranking in Uncertain Vector Spaces

13

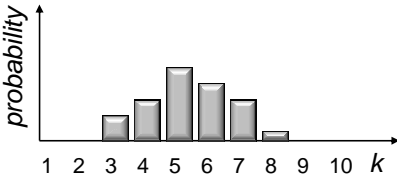


DATABASE  
SYSTEMS  
GROUP

## Probabilistic Similarity Ranking



- Output of probabilistic ranking:
  - for each object: discrete pdf over ranking positions
$$prob\_ranked_q: \mathcal{D} \times \{1, \dots, N\} \rightarrow [0..1]$$




k	probability
3	0.10
4	0.15
5	0.25
6	0.20
7	0.15
8	0.05

  - $prob\_ranked_q(o, k)$  reports the probability that object  $o$  is exactly the  $k^{th}$ -nearest-neighbor of the query object  $q$


M. Renz: Probabilistic Ranking in Uncertain Vector Spaces

14



DATABASE  
SYSTEMS  
GROUP

## Probabilistic Similarity Ranking



LMU

• Example: Probabilistic Ranking Output

□  $a_1$  (0.2)

□  $a_2$  (0.8) •  $q$

★  $b_1$  (0.5)

★  $b_2$  (0.2) ★  $b_3$  (0.3)

vector space

△  $c_1$  (0.2)

△  $c_2$  (0.2)


△  $c_3$  (0.6)

$A$	(0.46)	1
$B$	(0.54) (0.54)	2
$C$	(0.12) (0.12)	3
	(0.88)	

probabilistic ranking output


– probabilistic rank assignment is expensive

M. Renz: Probabilistic Ranking in Uncertain Vector Spaces
15



DATABASE  
SYSTEMS  
GROUP

## Probabilistic Similarity Ranking



LMU

– straight-forward approach


- instantiation of all possible worlds:
- complexity:  $O(S^N)$ 
  - S: number of vector instances per object
  - N: number of uncertain objects

➔ not applicable for large databases

– basic idea:


- iterative processing of the vector instances
- consider each instance only once
- apply distance browsing centered at query point  $q$

M. Renz: Probabilistic Ranking in Uncertain Vector Spaces
16

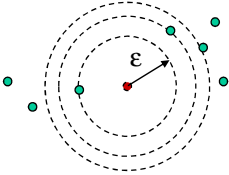


DATABASE  
SYSTEMS  
GROUP

## Probabilistic Similarity Ranking



- Iterative Probability Computation
  - ranking applied on vector instances
  - radial sweep with increasing range  $\epsilon$




	A	B	C	D
$P_o$	0.2	0.4	0.1	0.0

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


M. Renz: Probabilistic Ranking in Uncertain Vector Spaces

17



DATABASE  
SYSTEMS  
GROUP

## Probabilistic Similarity Ranking




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

$$P(o_{i,j}, k)_{DB} = \sum_{\substack{\sigma \subseteq DB \setminus \{o_i\} \\ |\sigma| = k-1}} \prod_{o \in DB \setminus (\sigma \cup \{o_i\})} \begin{cases} P(o) & \text{if } o \in \sigma \\ (1 - P(o)) & \text{else} \end{cases}$$


- problem:
  - for each  $k$ ,  $\binom{N}{k-1}$  possible subsets  $\sigma$  must be considered
  - very expensive for larger databases

M. Renz: Probabilistic Ranking in Uncertain Vector Spaces

18

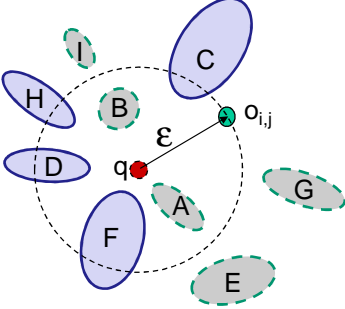


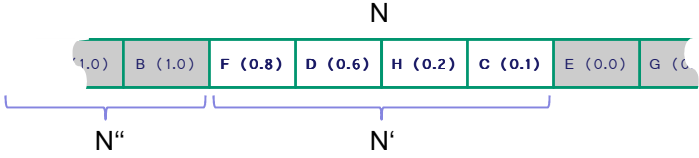
**Probabilistic Similarity Ranking**



– 1. Approach:


- apply only relevant objects
- prune objects that are beyond  $\epsilon$ :
  - reduce  $DB \rightarrow DB'$  ( $|DB'| \ll |DB|$ )
  - $DB' = \{o \in DB: 0 < P(o) < 1\}$
- $P(a_{i,j}, k)_{DB} = P(a_{i,j}, k - N'')_{DB'}$






M. Renz: Probabilistic Ranking in Uncertain Vector Spaces

19

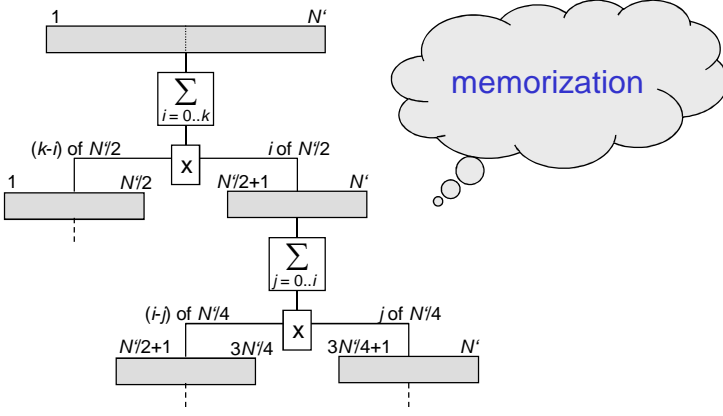


**Probabilistic Similarity Ranking**



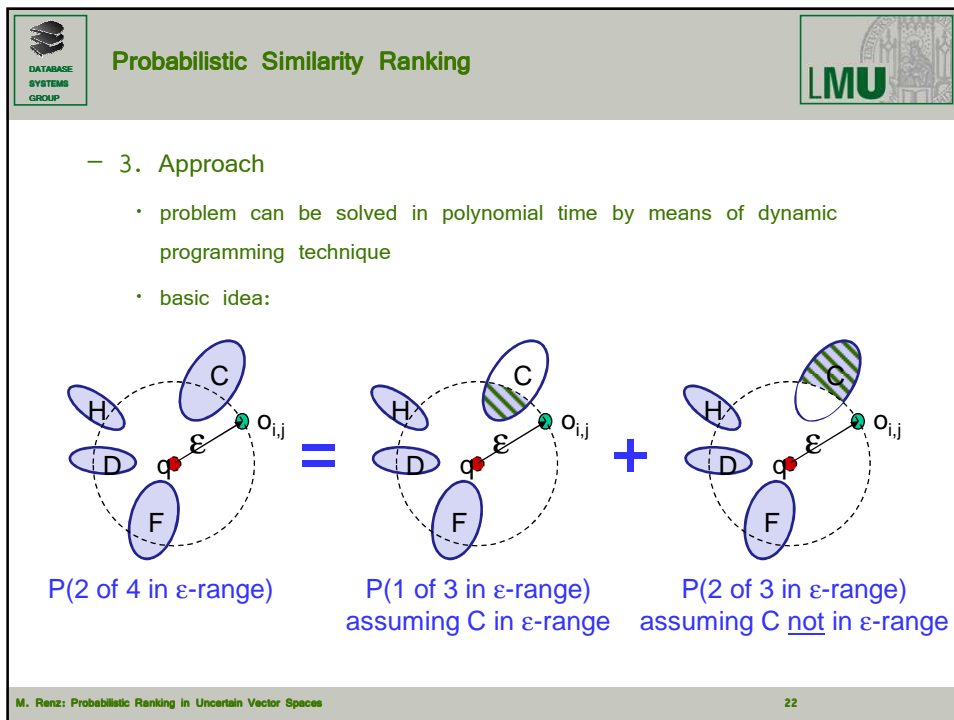
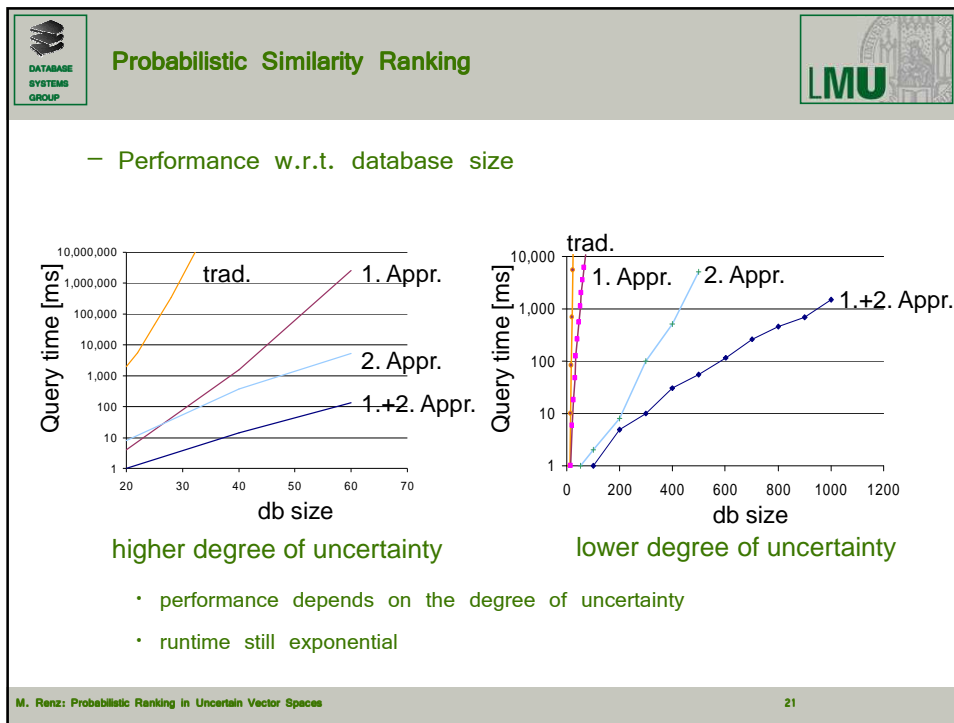
– 2. Approach: avoid redundant computations


(Divide & Conquer) [Kriegel et al. SSDBM'08]



M. Renz: Probabilistic Ranking in Uncertain Vector Spaces


20





DATABASE  
SYSTEMS  
GROUP

## Probabilistic Similarity Ranking



– 3. Approach:

- problem can be solved in polynomial time by means of dynamic programming technique:
- recursive function:


$$P(o_{i,j}, k)_{DB} = P(o_{i,j}, k-1)_{DB \setminus \{o_k\}} \cdot P_i(o_k) + P(o_{i,j}, k)_{DB \setminus \{o_k\}} (1 - P_i(o_k))$$

where

$$P(o_{i,j}, 0)_{\{\}} = 1 \quad \text{and}$$


$$P(o_{i,j}, k)_{DB} = 0 \quad \text{if } k > |DB|$$

M. Renz: Probabilistic Ranking in Uncertain Vector Spaces
23

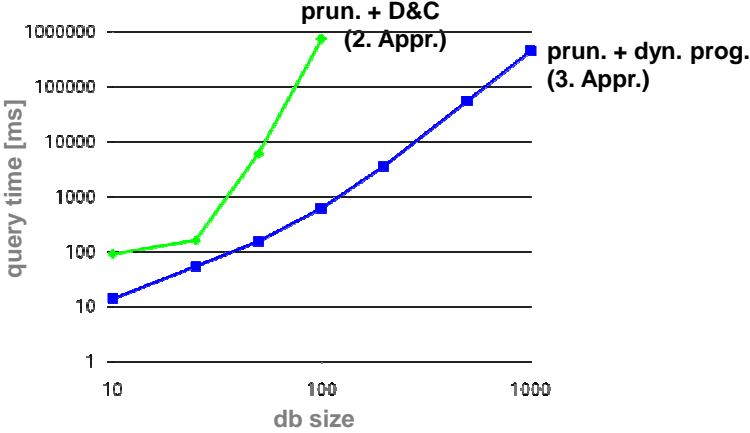


DATABASE  
SYSTEMS  
GROUP

## Probabilistic Similarity Ranking




• Experiments:




db size	prun. + D&C (2. Appr.) [ms]	prun. + dyn. prog. (3. Appr.) [ms]
10	~100	~15
20	~150	~30
50	~1000	~100
100	~10000	~300
200	~100000	~1000
500	~1000000	~3000
1000	-	~10000

M. Renz: Probabilistic Ranking in Uncertain Vector Spaces
24




## Outline




- Introduction
- Uncertainty Model
  - modelling uncertain vector data
  - probabilistic query processing
- Probabilistic Similarity Ranking
  - probabilistic ranking definition
  - approaches for efficient ranking processing
- **Summary**

M. Renz: Probabilistic Ranking in Uncertain Vector Spaces25



## Summary



- approaches to accelerate probabilistic ranking queries in vector spaces
- assumption:
  - objects are mutually independent
  - discrete uncertainty representations
- support by
  - traditional access methods
  - multi-step query processing techniques
- very high speed-up factor using Dyn. Prog.

M. Renz: Probabilistic Ranking in Uncertain Vector Spaces26

