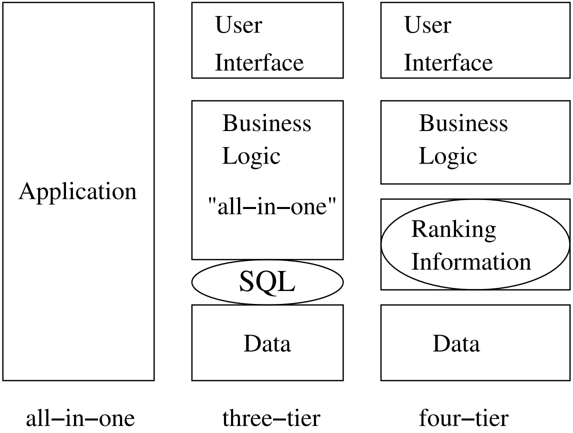
Modelling Retrieval Models in a Probabilistic Relational Algebra with a new Operator: The Relational Bayes

CWI, Amsterdam, December 2007
Thomas Roelleke
Queen Mary University of London

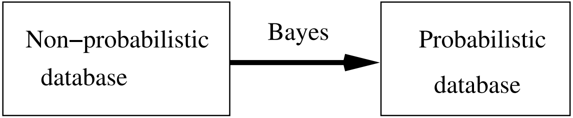
Outline

- Introduction
- Background: ProbDB, IR-on-DB, Retrieval Models, PRA
- The Relational Bayes
    - DB+IR Toy Database:
      person(Name, City, Nationality); coll(Term, DocId);
    - Motivation
    - City_Nationality and Nationality_City
    - Syntax and Examples
- Modelling Retrieval Models: TF-IDF, BIR, LM
- Implementation
- Summary and Outlook

Data and information independence



Generation of probabilistic databases



Layers

External layer	Information sorted by relevance
Logical layer	Probabilistic relations Probability estimation: Bayes[]()
Physical layer	Relational model/algebra; SQL

Probabilistic Databases

[Cavallo and Pittarelli, 1987]: Relational and probabilistic databases, information content (Shannon), probabilistic data dependencies

[Fuhr and Roelleke, 1997]: Probabilistic relational algebra for the integration of IR and DB: intensional vs extensional semantics, event expressions, solve Norbert's "db AND NOT ir OR ir AND NOT db"

[Dalvi and Suciu, 2004]: Efficient query evaluation: intensional semantics and possible worlds semantics, safe-plan optimisation algorithm

IR-on-DB

[Schek and Pistor, 1982]: Data Structures for an Integrated Database Management and Information Retrieval System

[Agrawal et al., 2002, Hristidis and Papakonstantinou, 2002]: DBXplorer, DISCOVER: keyword search

[Schefe, 1983]: Natuerlichsprachiger Zugang zu Datenbanken?

[Chaudhuri et al., 2004, Chaudhuri et al., 2006]: Probabilistic ranking of database query results (based on BIR model)

[Ercegovac et al., 2005]: TEXTURE benchmark (automatic scaling of benchmark; three competing systems)

[Cornacchia and de Vries, 2007]: A parametrised search system

(Probabilistic) Retrieval Models

[Robertson and Sparck Jones, 1976, Croft and Harper, 1979]:
BIR, “the probabilistic model”

[Wong and Yao, 1995, Roelleke et al., 2006]: $P(d \rightarrow q)$ and
matrix framework

[Ponte and Croft, 1998, Hiemstra, 2000, Lafferty and Zhai, 2003]:
language modelling (LM)

[Roelleke, 2003, Robertson, 2004, Robertson, 2005, Roelleke and Wa
“probability of being informative”, on theoretical
arguments for idf, on event spaces, a parallel
derivation of probabilistic models

[Motro, 1988, Motro, 1990, Fuhr, 1990, Bosc and Pivert, 1994]: vague queries, fuzzy

[Barbara et al., 1992, Lee, 1992, Macleod, 1991]: probabilistic relational modelling, probability *AGGREGATION*, text retrieval with SQL

[Niemi and Järvelin, 1995, Fuhr and Roelleke, 1996]: NF2 (non-first-normal-form): [Schek and Scholl, 1986]: relation-valued attributes

[Robert Ross, 2002]: probabilistic aggregates: the aggregates underline the difference between "normal" attributes and tuple probabilities

[Grossman and Frieder, 2004]: implement TF-IDF VSM in SQL

[Roelleke and Fuhr, 1996, Fuhr et al., 1998, Lalmas et al., 2002]: POOL SIGIR 06, "Dolores" SIGIR 98, POOL (in Intelligent Exploration of the Web 02), Logic in IR,

Coll	
Term	**DocId**
sailing	doc1
boats	doc1
sailing	doc2
sailing	doc2
boats	doc2
sailing	doc3
east	doc3
coast	doc3
sailing	doc4
boats	doc5

$\Longrightarrow$

probColl: “tf”		
Prob	**Term**	**DocId**
0.5	sailing	doc1
0.5	boats	doc1
0.66	sailing	doc2
0.33	boats	doc2
0.33	sailing	doc3
0.33	east	doc3
0.33	coast	doc3
1.0	sailing	doc4
1.0	boats	doc5

```
-- PSQL
INSERT INTO probQuery VALUES
  0.4 ('sailing', 'q2'),
  0.6 ('boats', 'q2');

-- Query
CREATE VIEW retrieved AS
  SELECT DocId
  FROM probQuery, probColl
  WHERE probQuery.Term = probColl.Term;
```

non-distinct	
Prob	DocId
$0.4 \cdot 0.5$	doc1
$0.6 \cdot 0.5$	doc1
$0.4 \cdot 0.66$	doc2
$0.6 \cdot 0.33$	doc2
$0.4 \cdot 0.33$	doc3
$0.4 \cdot 1.0$	doc4
$0.6 \cdot 1.0$	doc5

distinct (aggregated)	
Prob	DocId
1.0	doc1
1.0	doc2
$0.4 \cdot 0.33$	doc3
0.4	doc4
0.6	doc5

Aggregation of non-distinct tuples:

```
SELECT DISJOINT DocId FROM retrieved;
```

- Where do the probabilities in probColl ("tf") and probQuery come from?
- How can the estimation be described algebraically?

person		
Name	City	Nationality
Peter	London	German
Paul	London	Irish
Mary	London	Irish
Thomas	Dortmund	German
Thomas	London	German
Thomas	Hamburg	German
Hany	London	Egyptian
Hany	London	Polish
Jun	London	Chinese
Zhi	London	Chinese

coll	
Term	DocId
sailing	doc1
boats	doc1
sailing	doc2
sailing	doc2
boats	doc2
sailing	doc3
east	doc3
coast	doc3
sailing	doc4
boats	doc5

City_Nationality		
Prob	City	Nationality
0.500000	"London"	"German"
1.000000	"London"	"Irish"
0.250000	"Dortmund"	"German"
0.250000	"Hamburg"	"German"
1.000000	"London"	"Egyptian"
1.000000	"London"	"Polish"
1.000000	"London"	"Chinese"

```
CREATE VIEW City_Nationality AS
  SELECT DISJOINT City, Nationality
  FROM person | DISJOINT(Nationality);
```

Nationality_City		
0.250000	"London"	"German"
0.250000	"London"	"Irish"
1.000000	"Dortmund"	"German"
1.000000	"Hamburg"	"German"
0.125000	"London"	"Egyptian"
0.125000	"London"	"Polish"
0.250000	"London"	"Chinese"

```
-- PSQL
CREATE VIEW Nationality_City AS
  SELECT DISJOINT City, Nationality
  FROM person | DISJOINT(City);

# PRA
Nationality_City =
  Project DISJOINT[$2,$3](Bayes[$2](person))
```

```
# PRA
Bayes <estAssumption>
 [<evidenceKey>](<prae>)
```

```
-- PSQL
SELECT ... FROM ... WHERE
 ASSUMPTION <estAssumption>
 EVIDENCE KEY <evidenceKey>
```

$$
\begin{aligned}
\text{evidence} &= \text{Project estAssumption[i1..in](a)} \\
T &:= T_a \\
P(\tau) &= \begin{cases} \dfrac{P_a(\tau)}{P_{\text{evidence}}(\tau[i_1..i_n])} \\ \dfrac{\log P_a(\tau)}{\log P_{\text{evidence}}(\tau[i_1..i_n])} \end{cases}
\end{aligned}
$$

In 2003 (early Bayes), Bayes result was distinct; in 2004, revision led to $T := T_a$.

Motivation: "Seamlessly" integrated DB+IR technology

- Support document/context and tuple retrieval
- Support free-text and semantic/data retrieval
- Support the flexible modelling of *ALL* retrieval models
- Support the high-level (abstract) modelling of general and specific retrieval tasks (ad-hoc retrieval, classification, summarisation, structured document retrieval, hypertext retrieval, multimedia retrieval, ...)
- Support text, XML, and SQL

Retrieval Models: Notation Notation ([Roelleke et al., 2006]):

$n_D(t,x)$	number of *Documents* ...
$n_L(t,x)$	number of *Locations* ...
$P_D(t,x) := \frac{n_D(t,x)}{N_D(x)}$	document-based term probability
$P_L(t,x) := \frac{n_L(t,x)}{N_L(x)}$	location-based term probability

...

Retrieval Models: RSV's

$$RSV_{\text{TF-IDF}}(d,q) := \sum_{t \in q} tf(t,d) \cdot idf(t,c) \tag{1}$$

$$RSV_{\text{BIR}}(d,q) := \sum_{t \in d \cap q} \log \frac{P(t|r)}{P(t|\bar{r})} \tag{2}$$

$$P(q|d) := \prod_{t \in q} [\delta \cdot P(t|d) + (1-\delta) \cdot P(t|c)] \tag{3}$$

Retrieval Models: Relationships, Rewritings

$$RSV_{\mathrm{BIR}}(d,q) := \sum_{t \in d \cap q} idf(t,\bar{r}) - idf(t,r) \tag{4}$$

$$RSV_{\mathrm{LM}}(d,q) := \log \frac{P(q|d)}{\prod_{t \in q}(1-\delta) \cdot P(t|c)} \tag{5}$$

$$= \sum_{t \in q} \log \left[1 + \frac{\delta \cdot P(t|d)}{(1-\delta) \cdot P(t|c)}\right] \tag{6}$$

eqn 4: BIR and IDF:
[Robertson, 2004, de Vries and Roelleke, 2005] eqn 6: LM:
[Hiemstra, 2000]
Investigating the relationships important for “good” design of probabilistic relational modelling.

```
-- inverse document frequency
CREATE VIEW idf AS
    SELECT Term FROM coll
    ASSUMPTION MAX INFORMATIVE
    EVIDENCE KEY ();

-- query term weighting
CREATE VIEW wQuery AS
    SELECT Term, QueryId FROM Query, idf
    WHERE Query.Term = idf.Term;

-- normalisation
CREATE VIEW norm_wQuery AS
    SELECT Term, QueryId FROM wQuery
    EVIDENCE KEY (QueryId);
```

```
-- retrieve documents
CREATE VIEW std_tf_idf_retrieve AS
    SELECT DISJOINT DocId, QueryId
    FROM norm_wQuery, tf
    WHERE norm_wQuery.Term = tf.Term;
```

```
# tf := P(t|d)
tf =
 Project DISJOINT(
  Bayes DISJOINT[$2](coll));

# idf(t,c) :=
# -log P(t|c) /
# max_idf(c)
idf =
 Bayes MAX_IDF[](
  Project[$1](coll));
```

```
-- P(t occurs | d)
CREATE VIEW tf AS
  SELECT DISJOINT Term, DocId
  FROM coll |
       DISJOINT (DocId);

-- P(t informs | c)
CREATE VIEW idf AS
  SELECT Term
  FROM coll
  ASSUMPTION MAX_IDF
  EVIDENCE KEY ();
```

```
-- idf in collection / non-relevant:
CREATE VIEW idf_c AS
    SELECT Term FROM coll
    ASSUMPTION MAX_IDF
    EVIDENCE KEY ();

-- idf in relevant:
CREATE VIEW idf_r AS
    SELECT Term FROM relColl
    ASSUMPTION MAX_IDF
    EVIDENCE KEY ();
```

```
CREATE VIEW docModel AS
    SELECT Term, DocId FROM lambda1, p_t_d;

CREATE VIEW collModel AS
    SELECT Term, DocId FROM lambda2, p_t_c, retrieved;

-- combine document and collection models
CREATE VIEW lm1_mix AS
    docModel UNION DISJOINT collModel;

-- retrieve documents
CREATE VIEW lm1_retrieve AS
    SELECT SUM_LOG DocId, QueryId
    FROM Query, lm1_mix
    WHERE Query.Term = lm1_mix.Term;
```

Implementation

- Retrieval of $P(t|c)$ probabilities in $O(1)$ (special index)
- Incremental update facility
- $P(t|d)$ probabilities/views are materialised off-line; future research
- Difficulty: System A runs SQL, System B runs PSQL: How to compare?
- top-k and early-response processing

Summary

- Probabilistic DB, IR-on-DB, PRA, Retrieval Models
- The magnificent five (Select, Project, Join/Multiply, Unite, Subtract): describe probability *AGGREGATION*
- The relational Bayes: describe probability *ESTIMATION*
- DB+IR requires "relaxed" probability theory
  - *idf*-based ("informativeness") probabilities
  - here and there, relax the boundaries of probabilistic modelling and rethink the genuine formulation of IR models
- Scalability: $O(1)$ retrieval of Bayes tuples for $P(t|c)$, given Bayes-oriented indexing structure

Outlook

- Database/tuple ranking: modelling and application of retrieval models to SELECT-FROM-WHERE; entropy-based ranking
- Design and verification of probabilistic logical programs
- Optimisation: Semantic, algebraic, and processing (Hengzhi Wu)
- High-level languages: RDF-SPARQL → PSQL/PRA (Hany Azzam)
- POLIS: Probabilistic Object-oriented logic for information summarisation (Fred Forst)
- POLAR: Probabilistic Object-oriented annotation-based retrieval (Ingo Frommholz)

Agrawal, S., Chaudhuri, S., and Das, G. (2002).
Dbxplorer: A system for keyword-based search over relational databases.
In *ICDE*, pages 5–16.

Barbara, D., Garcia-Molina, H., and Porter, D. (1992).
The management of probabilistic data.
IEEE Transactions on Knowledge and Data Engineering, 4(5):487–502.

Bosc, P. and Pivert, O. (1994).
Fuzzy queries and relational databases.
In *Proceedings of the 1994 ACM Symposium on Applied Computing*, pages 170–174. ACM Press.

Cavallo, R. and Pittarelli, M. (1987).
The theory of probabilistic databases.
In *Proceedings of the 13th International Conference on Very Large Databases*, pages 71–81, Los Altos, California. Morgan Kaufman.

Chaudhuri, S., Das, G., Hristidis, V., and Weikum, G. (2004).
Probabilistic ranking of database query results.
In *VLDB*, pages 888–899.

Chaudhuri, S., Das, G., Hristidis, V., and Weikum, G. (2006).
Probabilistic information retrieval approach for ranking of database query results.
ACM Trans. Database Syst., 31(3):1134–1168.

Cornacchia, R. and de Vries, A. P. (2007).
A parameterised search system.
In *ECIR*, pages 4–15.

Croft, W. and Harper, D. (1979).
Using probabilistic models of document retrieval without relevance information.
Journal of Documentation, 35:285–295.

Croft, W. B., Moffat, A., van Rijsbergen, C. J., Wilkinson, R., and Zobel, J., editors (1998).
Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York. ACM.

Dalvi, N. N. and Suciu, D. (2004).
Efficient query evaluation on probabilistic databases.
In Nascimento, M. A., Özsu, M. T., Kossmann, D., Miller, R. J., Blakeley, J. A., and Schiefer, K. B., editors, *VLDB*, pages 864–875. Morgan Kaufmann.

de Vries, A. and Roelleke, T. (2005).
Relevance information: A loss of entropy but a gain for idf?
In *ACM SIGIR*, Salvador, Brazil.

Ercegovac, V., DeWitt, D. J., and Ramakrishnan, R. (2005).
The texture benchmark: Measuring performance of text queries on a relational dbms.
In *VLDB*, pages 313–324.

Fuhr, N. (1990).
A probabilistic framework for vague queries and imprecise information in databases.
In McLeod, D., Sacks-Davis, R., and Schek, H., editors, *Proceedings of the 16th International Conference on Very Large Databases*, pages 696–707, Los Altos, California. Morgan Kaufman.

Fuhr, N., Gövert, N., and Roelleke, T. (1998).
Dolores: A system for logic-based retrieval of multimedia objects.

In [Croft et al., 1998], pages 257–265.

Fuhr, N. and Roelleke, T. (1996).
A probabilistic NF2 relational algebra for integrated information retrieval and database systems.
In Tanik, M. M., Bastani, F. B., Gibson, D., and Fielding, P. J., editors, *Proceedings of the 2nd World Conference on Integrated Design and Process Technology (IDPT)*, pages 17–30, Austin, Texas. Society for Design and Process Science (SDPS).

Fuhr, N. and Roelleke, T. (1997).
A probabilistic relational algebra for the integration of information retrieval and database systems.
ACM Transactions on Information Systems (TOIS), 14(1):32–66.

Grossman, D. A. and Frieder, O. (2004).
Information Retrieval. Algorithms and Heuristics, 2nd ed., volume 15 of *The Information Retrieval Series*.
Springer.

Hiemstra, D. (2000).
A probabilistic justification for using tf.idf term weighting in information retrieval.
International Journal on Digital Libraries, 3(2):131–139.

Hristidis, V. and Papakonstantinou, Y. (2002).
Discover: Keyword search in relational databases.
In *VLDB*, pages 670–681.

Lafferty, J. and Zhai, C. (2003).
Probabilistic Relevance Models Based on Document and Query Generation, chapter 1.
Kluwer.

Lalmas, M., Roelleke, T., and Fuhr, N. (2002).
Intelligent hypermedia retrieval.
In Szczepaniak, P. S., Segovia, F., and Zadeh, L. A., editors, *Intelligent Exploration of the Web*. Springer-Verlag Group (Physica-Verlag).

Lee, S. (1992).
An extended relational database model for uncertain and imprecise information.
In *Proceedings of the 18th VLDB Conference*, pages 211–220, Los Altos, California. Morgan Kaufman.

Macleod, I. (1991).
Text retrieval and the relational model.
Journal of the American Society for Information Science, 42(3):155–165.

Motro, A. (1988).
Vague: A user interface to relational databases that permits vague queries.
ACM Transactions on Office Information Systems, 6(3):187–214.

Motro, A. (1990).
Accommodating imprecision in database systems: Issues and solutions.
Sigmod record, 19(4):69.

Niemi, T. and Järvelin, K. (1995).
A straightforward NF2 relational interface with applications in information retrieval.
Information Processing and Management, 31(2):215–231.

Ponte, J. and Croft, W. (1998).
A language modeling approach to information retrieval.
In [Croft et al., 1998], pages 275–281.

Robert Ross, V. S. Subrahmanian, J. G. (2002).
Probabilistic aggregates.
In *13th International Symposium on Methodologies for Intelligent Systems (ISMIS), Lyon, France*, Foundations of Intelligent Systems. Springer.

Robertson, S. (2004).
Understanding inverse document frequency: On theoretical arguments for idf.
Journal of Documentation, 60:503–520.

Robertson, S. (2005).
On event spaces and probabilistic models in information retrieval.
Information Retrieval, 8(2):319–329.

Robertson, S. and Sparck Jones, K. (1976).
Relevance weighting of search terms.
Journal of the American Society for Information Science, 27:129–146.

Roelleke, T. (2003).
A frequency-based and a Poisson-based probability of being informative.
In *ACM SIGIR*, pages 227–234, Toronto, Canada.

Roelleke, T. and Fuhr, N. (1996).
Retrieval of complex objects using a four-valued logic.
In Frei, H.-P., Harmann, D., Schäuble, P., and Wilkinson, R., editors, *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–214, New York. ACM.

Roelleke, T., Tsikrika, T., and Kazai, G. (2006).

A general matrix framework for modelling information retrieval.
Journal on Information Processing & Management (IP&M), Special Issue on Theory in Information Retrieval, 42(1).

Roelleke, T. and Wang, J. (2006).
A parallel derivation of probabilistic information retrieval models.
In *ACM SIGIR*, pages 107–114, Seattle, USA.

Schefe, P. (1983).
Natürlichsprachiger zugang zu datenbanken?
Angewandte Informatik, (10):419–423.

Schek, H.-J. and Pistor, P. (1982).
Data structures for an integrated database management and information retrieval system.
In *Proceedings of the 8th International Conference on Very Large Data Bases*, pages 197–207, Los Altos, California. Morgan Kaufman.

Schek, H.-J. and Scholl, M. (1986).
The relational model with relation-valued attributes.
Information systems, 2:137–147.

Wong, S. and Yao, Y. (1995).
On modeling information retrieval with probabilistic inference.
ACM Transactions on Information Systems, 13(1):38–68.