Bridging Information Retrieval and Databases

Tutorial at the PROMISE Winter School 2013

Norbert Fuhr

February 6, 2013

Introduction

IR and Databases
The Logic View

Retrieval

- DB: given query q, find objects o with $o \rightarrow q$
- IR: given query q, find documents d with high values of $P(d \rightarrow q)$
- DB is a special case of IR!
  (in a certain sense)

IR and Databases
The Logic View

Retrieval

- DB: given query q, find objects o with $o \rightarrow q$
- IR: given query q, find documents d with high values of $P(d \rightarrow q)$
- DB is a special case of IR!
  (in a certain sense)

This tutorial: Focusing on the logic view

- Inference
- Vague predicates
- Query language expressiveness

Inference

- IR with the Relational Model
- The Probabilistic Relational Model
- Interpretation of probabilistic weights
- Extensions
  - Disjoint events
  - Relational Bayes
  - Probabilistic rules

Relational Model
Projection

index

DOCNO	TERM
1	ir
1	db
2	ir
3	db
3	oop
4	ir
4	ai
5	db
5	oop

topic
ir
db
oop
ai

Projection: what is the collection about?

```
topic(T) :- index(D,T).
```

Relational Model
Selection

index

DOCNO	TERM
1	ir
1	db
2	ir
3	db
3	oop
4	ir
4	ai
5	db
5	oop

aboutir
1
2
4

Selection: which documents are about IR?

```
aboutir(D) :- index(D,ir).
```

Relational Model
Join

index

DOCNO	TERM
1	ir
1	db
2	ir
3	db
3	oop
4	ir
4	ai
5	db
5	oop

author

DOCNO	NAME
1	smith
2	miller
3	johnson
4	firefly
4	bradford
5	bates

irauthor
smith
miller
firefly
bradford

Join: who writes about IR?

```
irauthor(A):- index(D,ir) & author(D,A).
```

Relational Model
Union

index

DOCNO	TERM
1	ir
1	db
2	ir
3	db
3	oop
4	ir
4	ai
5	db
5	oop

irordb
1
2
3
4
5

Union: which documents are about IR or DB?

```
irordb(D) :- index(D,ir).
irordb(D) :- index(D,db).
```

Relational Model
Difference

index

DOCNO	TERM
1	ir
1	db
2	ir
3	db
3	oop
4	ir
4	ai
5	db
5	oop

irnotdb
2
4

Difference: which documents are about IR, but not DB?

```
irnotdb(D) :- index(D,ir) & not(index(D,db)).
```

The Probabilistic Relational Model

[Fuhr & Roelleke 97] [Suciu et al 11]

index

β	DOCNO	TERM
0.8	1	IR
0.7	1	DB
0.6	2	IR
0.5	3	DB
0.8	3	OOP
0.9	4	IR
0.4	4	AI
0.8	5	DB
0.3	5	OOP

The Probabilistic Relational Model

[Fuhr & Roelleke 97] [Suciu et al 11]

index

β	DOCNO	TERM
0.8	1	IR
0.7	1	DB
0.6	2	IR
0.5	3	DB
0.8	3	OOP
0.9	4	IR
0.4	4	AI
0.8	5	DB
0.3	5	OOP

Which documents are about DB?

```
aboutdb(D) :- index(D,db).
```

The Probabilistic Relational Model

[Fuhr & Roelleke 97] [Suciu et al 11]

index

β	DOCNO	TERM
0.8	1	IR
0.7	1	DB
0.6	2	IR
0.5	3	DB
0.8	3	OOP
0.9	4	IR
0.4	4	AI
0.8	5	DB
0.3	5	OOP

aboutdb

0.7	1
0.5	3
0.8	5

Which documents are about DB?

```
aboutdb(D) :- index(D,db).
```

The Probabilistic Relational Model

[Fuhr & Roelleke 97] [Suciu et al 11]

index

β	DOCNO	TERM
0.8	1	IR
0.7	1	DB
0.6	2	IR
0.5	3	DB
0.8	3	OOP
0.9	4	IR
0.4	4	AI
0.8	5	DB
0.3	5	OOP

aboutdb

0.7	1
0.5	3
0.8	5

aboutirdb

0.8*0.7	1

Which documents are about DB?

```
aboutdb(D) :- index(D,db).
```

Which documents are about IR and DB?

```
aboutirdb(D) :- index(D,ir) & index(D,db).
```

Extensional vs. intensional semantics

docterm

β	DOC	TERM
0.9	d1	ir
0.5	d1	db

link

β	S	T
0.7	d2	d1

```
about(D,T) :- docTerm(D,T).
about(D,T) :- link(D,D1) & about(D1,T)
q(D) :- about(D,ir) & about(D,db).
```

Extensional vs. intensional semantics

docterm

β	DOC	TERM
0.9	d1	ir
0.5	d1	db

link

β	S	T
0.7	d2	d1

```
about(D,T) :- docTerm(D,T).
about(D,T) :- link(D,D1) & about(D1,T)
q(D) :- about(D,ir) & about(D,db).
```

extensional semantics:

weight of derived fact as function of weights of subgoals

$$P(\texttt{q(d2)}) = P(\texttt{about(d2,ir)}) \cdot P(\texttt{about(d2,db)}) = (0.7 \cdot 0.9) \cdot (0.7 \cdot 0.5)$$

Extensional vs. intensional semantics

docterm

β	DOC	TERM
0.9	d1	ir
0.5	d1	db

link

β	S	T
0.7	d2	d1

```
about(D,T) :- docTerm(D,T).
about(D,T) :- link(D,D1) & about(D1,T)
q(D) :- about(D,ir) & about(D,db).
```

extensional semantics:

weight of derived fact as function of weights of subgoals

$$P(\mathtt{q(d2)}) = P(\mathtt{about(d2,ir)}) \cdot P(\mathtt{about(d2,db)}) = (0.7 \cdot 0.9) \cdot (0.7 \cdot 0.5)$$

Problem

"improper treatment of correlated sources of evidence" [Pearl 88]
$\rightarrow$ extensional semantics only correct for tree-shaped inference structures

Intensional semantics

weight of derived fact as function of weights of underlying ground facts

Intensional semantics

weight of derived fact as function of weights of underlying ground facts

Method: Event keys and event expressions

docterm

β	κ	DOC	TERM
0.9	dT(d1,ir)	d1	ir
0.5	dT(d1,db)	d1	db

link

β	κ	S	T
0.7	l(d2,d1)	d2	d1

Intensional semantics

weight of derived fact as function of weights of underlying ground facts

Method: Event keys and event expressions

docterm

β	κ	DOC	TERM
0.9	dT(d1,ir)	d1	ir
0.5	dT(d1,db)	d1	db

link

β	κ	S	T
0.7	l(d2,d1)	d2	d1

```
?- docTerm(D,ir) & docTerm(D,db).
```

gives

```
d1 [dT(d1,ir) & dT(d1,db)]
```

Intensional semantics

weight of derived fact as function of weights of underlying ground facts

Method: Event keys and event expressions

docterm

β	κ	DOC	TERM
0.9	dT(d1,ir)	d1	ir
0.5	dT(d1,db)	d1	db

link

β	κ	S	T
0.7	l(d2,d1)	d2	d1

```
?- docTerm(D,ir) & docTerm(D,db).
```

gives

`d1 [dT(d1,ir) & dT(d1,db)]` $0.9 \cdot 0.5 = 0.45$

Event keys and event expressions

docterm

β	κ	DOC	TERM
0.9	dT(d1,ir)	d1	ir
0.5	dT(d1,db)	d1	db

link

β	κ	S	T
0.7	l(d2,d1)	d2	d1

```
about(D,T) :- docTerm(D,T).
about(D,T) :- link(D,D1) & about(D1,T)
?- about(D,ir) & about(D,db).
```

gives

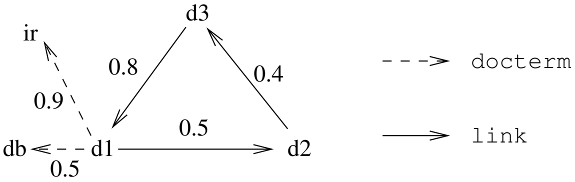
`d1 [dT(d1,ir) & dT(d1,db)]` $0.9 \cdot 0.5 = 0.45$

`d2 [l(d2,d1) & dT(d1,ir) & l(d2,d1) & dT(d1,db)]`

$0.7 \cdot 0.9 \cdot 0.5 = 0.315$

Recursion

```
about(D,T) :- docTerm(D,T).
about(D,T) :- link(D,D1) & about(D1,T).
```




```
?- about(D,ir)
d1 [dT(d1,ir) | l(d1,d2) & l(d2,d3) & l(d3,d1) &
     dT(d1,ir) | ...]                                  0.900
d3 [l(d3,d1) & dT(d1,ir)]                              0.720
d2 [l(d2,d3) & l(d3,d1) & dT(d1,ir)]                   0.288

?- about(D,ir) & about(D,db)
d1 [dT(d1,ir) & dT(d1,db)]                             0.450
d3 [l(d3,d1) & dT(d1,ir) & l(d3,d1) & dT(d1,db)]       0.360
```

Computation of probabilities for event expressions

1. transformation of expression into disjunctive normal form
2. application of sieve formula:
   - simple case of 2 conjuncts: $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

Computation of probabilities for event expressions

1. transformation of expression into disjunctive normal form
2. application of sieve formula:
   - simple case of 2 conjuncts: $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$
   - general case:
     c_i – conjunct of event keys

$$P(c_1 \vee \ldots \vee c_n) = \sum_{i=1}^{n} (-1)^{i-1} \sum_{1 \le j_1 < \ldots < j_i \le n} P(c_{j_1} \wedge \ldots \wedge c_{j_i}).$$

- $\rightsquigarrow$ exponential complexity

Computation of probabilities for event expressions

1. transformation of expression into disjunctive normal form
2. application of sieve formula:
   - simple case of 2 conjuncts: $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$
   - general case:
     c_i – conjunct of event keys

$$P(c_1 \ \vee \ldots \vee \ c_n) = \sum_{i=1}^{n} (-1)^{i-1} \sum_{1 \le j_1 < \ldots < j_i \le n} P(c_{j_1} \ \wedge \ldots \wedge \ c_{j_i}).$$

- $\rightsquigarrow$ exponential complexity
- $\rightsquigarrow$ use only when necessary for correctness

Computation of probabilities for event expressions

1. transformation of expression into disjunctive normal form
2. application of sieve formula:
    - simple case of 2 conjuncts: $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$
    - general case:
      c_i – conjunct of event keys

$$P(c_1 \vee \ldots \vee c_n) =$$
$$\sum_{i=1}^{n} (-1)^{i-1} \sum_{1 \leq j_1 < \ldots < j_i \leq n} P(c_{j_1} \wedge \ldots \wedge c_{j_i}).$$

- $\rightsquigarrow$ exponential complexity
- $\rightsquigarrow$ use only when necessary for correctness
- see [Dalvi & Suciu 07]

Possible worlds semantics

```
0.9 docTerm(d1,ir).
```

$P(W_1) = 0.9$: {`docTerm(d1,ir)`}
$P(W_2) = 0.1$: {}

```
0.6 docTerm(d1,ir). 0.5 docTerm(d1,db).
```

Possible interpretations:

I_1: $P(W_1) = 0.3$: {`docTerm(d1,ir)`}
$P(W_2) = 0.3$: {`docTerm(d1,ir), docTerm(d1,db)`}
$P(W_3) = 0.2$: {`docTerm(d1,db)`}
$P(W_4) = 0.2$: {}

I_2: $P(W_1) = 0.5$: {`docTerm(d1,ir)`}
$P(W_2) = 0.1$: {`docTerm(d1,ir), docTerm(d1,db)`}
$P(W_3) = 0.4$: {`docTerm(d1,db)`}

I_3: $P(W_1) = 0.1$: {`docTerm(d1,ir)`}
$P(W_2) = 0.5$: {`docTerm(d1,ir), docTerm(d1,db)`}
$P(W_3) = 0.4$: {}

```
0.6 docTerm(d1,ir). 0.5 docTerm(d1,db).
```

Possible interpretations:

I_1: $P(W_1) = 0.3$: $\{\texttt{docTerm(d1,ir)}\}$
$P(W_2) = 0.3$: $\{\texttt{docTerm(d1,ir), docTerm(d1,db)}\}$
$P(W_3) = 0.2$: $\{\texttt{docTerm(d1,db)}\}$
$P(W_4) = 0.2$: $\{\}$

I_2: $P(W_1) = 0.5$: $\{\texttt{docTerm(d1,ir)}\}$
$P(W_2) = 0.1$: $\{\texttt{docTerm(d1,ir), docTerm(d1,db)}\}$
$P(W_3) = 0.4$: $\{\texttt{docTerm(d1,db)}\}$

I_3: $P(W_1) = 0.1$: $\{\texttt{docTerm(d1,ir)}\}$
$P(W_2) = 0.5$: $\{\texttt{docTerm(d1,ir), docTerm(d1,db)}\}$
$P(W_3) = 0.4$: $\{\}$

probabilistic logic:

$0.1 \leq P(\texttt{docTerm}(\texttt{d1}, \texttt{ir})\&\texttt{docTerm}(\texttt{d1}, \texttt{db})) \leq 0.5$

```
0.6 docTerm(d1,ir). 0.5 docTerm(d1,db).
```

Possible interpretations:

I_1: $P(W_1) = 0.3$: {docTerm(d1,ir)}
$P(W_2) = 0.3$: {docTerm(d1,ir), docTerm(d1,db)}
$P(W_3) = 0.2$: {docTerm(d1,db)}
$P(W_4) = 0.2$: {}

I_2: $P(W_1) = 0.5$: {docTerm(d1,ir)}
$P(W_2) = 0.1$: {docTerm(d1,ir), docTerm(d1,db)}
$P(W_3) = 0.4$: {docTerm(d1,db)}

I_3: $P(W_1) = 0.1$: {docTerm(d1,ir)}
$P(W_2) = 0.5$: {docTerm(d1,ir), docTerm(d1,db)}
$P(W_3) = 0.4$: {}

probabilistic logic:

$0.1 \leq P(\mathtt{docTerm}(\mathtt{d1}, \mathtt{ir}) \& \mathtt{docTerm}(\mathtt{d1}, \mathtt{db})) \leq 0.5$

probabilistic Datalog with independence assumptions:

$P(\mathtt{docTerm}(\mathtt{d1}, \mathtt{ir}) \& \mathtt{docTerm}(\mathtt{d1}, \mathtt{db})) = 0.3$

Disjoint events

β	City	State
0.7	Paris	France
0.2	Paris	Texas
0.1	Paris	Idaho

Disjoint events

β	City	State
0.7	Paris	France
0.2	Paris	Texas
0.1	Paris	Idaho

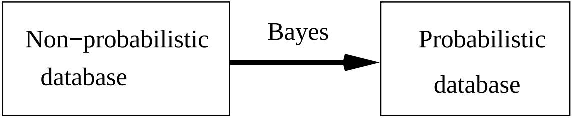
Interpretation:

$P(W_1) = 0.7$: $\{$`cityState(paris, france)`$\}$
$P(W_2) = 0.2$: $\{$`cityState(paris, texas)`$\}$
$P(W_3) = 0.1$: $\{$`cityState(paris, idaho)`$\}$

Relational Bayes

[Roelleke et al. 07]

Role of the relational Bayes: Generation of a probabilistic database



Relational Bayes
Example: P(Nationality | City)

nationality_and_city	
Nationality	City
"British"	"London"
"British"	"London"
"British"	"London"
"Scottish"	"London"
"French"	"London"
"German"	"Hamburg"
"German"	"Hamburg"
"Danish"	"Hamburg"
"British"	"Hamburg"
"German"	"Dortmund"
"German"	"Dortmund"
"Turkish"	"Dortmund"
"Scottish"	"Glasgow"

$\Longrightarrow$ Bayes[$City]()

nationality_city		
P(Nationality|City)	Nationality	City
0.600	"British"	"London"
0.200	"Scottish"	"London"
0.200	"French"	"London"
0.500	"German"	"Hamburg"
0.250	"Danish"	"Hamburg"
0.250	"British"	"Hamburg"
0.667	"German"	"Dortmund"
0.333	"Turkish"	"Dortmund"
1.000	"Scottish"	"Glasgow"

```
# P(Nationality | City):
nationality_city SUM(Nat, City) :-
        nationality_and_city (Nat, City) | (City);
```

Relational Bayes
Example: $P(t|d)$

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

p_t_d_space(Term, DocId) :- term(Term, DocId) | (DocId);		
$P(t\|d)$	Term	DocId
0.50	sailing	doc1
0.50	boats	doc1
0.33	sailing	doc2
0.33	boats	doc2
0.33	sailing	doc2
0.33	east	doc3
0.33	coast	doc3
0.33	sailing	doc3
1.00	sailing	doc4
1.00	boats	doc5

p_t_d SUM(Term, DocId) :- term(Term, DocId) | (DocId);		
$P(t\|d)$	Term	DocId
0.50	sailing	doc1
0.50	boats	doc1
0.67	sailing	doc2
0.33	boats	doc2
0.33	east	doc3
0.33	coast	doc3
0.33	sailing	doc3
1.00	sailing	doc4
1.00	boats	doc5

Probabilistic rules

Rules for deterministic facts:

```
0.7 likes-sports(X) :- man(X).
0.4 likes-sports(X) :- woman(X).
man(peter).
```

Probabilistic rules

Rules for deterministic facts:

```
0.7 likes-sports(X) :- man(X).
0.4 likes-sports(X) :- woman(X).
man(peter).
```

Interpretation:

$P(W_1) = 0.7$: {`man(peter), likes-sports(peter)`}

$P(W_2) = 0.3$: {`man(peter)`}

Probabilistic rules
Rules for uncertain facts:

```
#   gender is disjoint on the first attribute
0.7 l-sports(X)          :- gender(X,male).
0.4 l-sports(X)          :- gender(X,female).
0.5 gender(X,male)    :- human(X).
0.5 gender(X,female) :- human(X).
human(jo).
```

Probabilistic rules

Rules for uncertain facts:

```
#   gender is disjoint on the first attribute
0.7 l-sports(X)          :- gender(X,male).
0.4 l-sports(X)          :- gender(X,female).
0.5 gender(X,male)    :- human(X).
0.5 gender(X,female) :- human(X).
human(jo).
```

Interpretation:

$P(W_1) = 0.35$: {gender(jo,male), l-sports(jo)}
$P(W_2) = 0.15$: {gender(jo,male)}
$P(W_3) = 0.20$: {gender(jo,female), l-sports(jo)}
$P(W_4) = 0.30$: {gender(jo,female)}

```
?- l-sports(jo)
```

Probabilistic rules

Rules for uncertain facts:

```
#   gender is disjoint on the first attribute
0.7 l-sports(X)          :- gender(X,male).
0.4 l-sports(X)          :- gender(X,female).
0.5 gender(X,male)    :- human(X).
0.5 gender(X,female) :- human(X).
human(jo).
```

Interpretation:

$P(W_1) = 0.35$: {gender(jo,male), l-sports(jo)}
$P(W_2) = 0.15$: {gender(jo,male)}
$P(W_3) = 0.20$: {gender(jo,female), l-sports(jo)}
$P(W_4) = 0.30$: {gender(jo,female)}

`?- l-sports(jo)` $P(W_1) + P(W_3) = 0.55$

Probabilistic rules
Rules for independent events

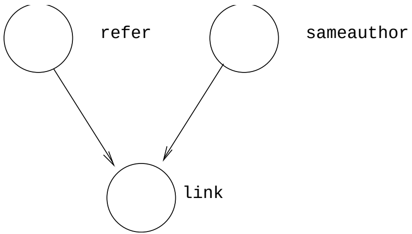
```
sameauthor(D1,D2) :- author(D1,X) & author(D2,X).
0.5 link(D1,D2) :- refer(D1,D2).
0.2 link(D1,D2) :- sameauthor(D1,D2).

?? link(D1,D2) :- refer(D1,D2) & sameauthor(D1,D2).
```

$P(l|r)$, $P(l|s) \rightarrow P(l|r \wedge s)$?

Rules for independent events
Modeling probabilistic inference networks

```
0.7 link(D1,D2) :- refer(D1,D2) & sameauthor(D1,D2).
0.5 link(D1,D2) :- refer(D1,D2) & not(sameauthor(D1,D2)).
0.2 link(D1,D2) :- sameauthor(D1,D2) & not(refer(D1,D2)).
```

Probabilistic inference networks,
rules define link matrix



Vague Predicates

- The Logical View on Vague Predicates
- Vague Predicates in IR and Databases
- Probabilistic Modeling of Vague Predicates

Vague Predicates
Motivating Example

"lcd tv 46inch"

Showing 1 - 16 of 3,851 Results

Samsung LN46E550 46-Inch 1080p 60Hz LCD HDTV by Samsung

~~$879.99~~ Click for product details
Order in the next **5 hours** and get it by **Wednesday, Jan 16**.
More Buying Choices
$463.80 used & new (14 offers)

(43)
Eligible for **FREE** Super Saver Shipping.
Electronics: See all 3,536 items

Samsung LN46D550 46-Inch 1080p 60Hz LCD HDTV (Black) by Samsung

~~$899.99~~ **$599.27**
Only 15 left in stock - order soon.
More Buying Choices
$599.27 new (4 offers)
$490.00 used (10 offers)

(161)
Electronics: See all 3,536 items

Cheetah Mounts APTMM2B Flush Tilt Dual Hook (1.3" from wall) Flat Screen
Cheetah

~~$40.99~~ **$27.99**
Order in the next **7 hours** and get it by **Wednesday, Jan 16**.
More Buying Choices
$27.99 new (9 offers)

(2,125)
#1 Best Seller in TV Accessories
Eligible for **FREE** Super Saver Shipping.
Electronics: See all 3,536 items

Vague Predicates
Motivating Example

"lcd tv 45inch"

Showing 1 - 16 of 2,617 Results

RCA 32LB45RQ 32-Inch Full 1080p 60Hz LCD HDTV by RCA

$228.38 used (4 offers)

(138)

Electronics See all 1,914 items

RCA 42LB45RQ 42-Inch 1080p 60Hz LCD HDTV (Black) by RCA

$476.99

Only 1 left in stock - order soon.

More Buying Choices

$476.99 new (2 offers)

$333.67 used (3 offers)

(138)

See newer version of this item

Electronics See all 1,914 items

RCA 22LB45RQD 22-Inch Full 1080p LCD/DVD Combo HDTV by RCA

~~$229.99~~ **$219.99**

Only 1 left in stock - order soon.

More Buying Choices

$188.99 new (3 offers)

$125.00 used (19 offers)

(80)

Electronics See all 1,914 items

Propositional vs. Predicate Logic

- Current IR systems are based on proposition logic (query term present/absent in document)

Propositional vs. Predicate Logic

- Current IR systems are based on proposition logic (query term present/absent in document)
- Similarity of values not considered

Propositional vs. Predicate Logic

- Current IR systems are based on proposition logic (query term present/absent in document)
- Similarity of values not considered
- but multimedia IR deals with similarity already

Propositional vs. Predicate Logic

- Current IR systems are based on proposition logic (query term present/absent in document)
- Similarity of values not considered
- but multimedia IR deals with similarity already




Propositional vs. Predicate Logic

- Current IR systems are based on proposition logic (query term present/absent in document)
- Similarity of values not considered
- but multimedia IR deals with similarity already
- $\rightsquigarrow$ transition from propositional to predicate logic necessary

Propositional vs. Predicate Logic

- Current IR systems are based on proposition logic (query term present/absent in document)
- Similarity of values not considered
- but multimedia IR deals with similarity already
- $\rightsquigarrow$ transition from propositional to predicate logic necessary
- $\Rightarrow$ Probabilistic databases / Datalog are already based on predicate logic!

Vague Predicates in Probabilistic Datalog

[Fuhr & Roelleke 97] [Fuhr 00]

- Example: Shopping 45 inch LCD TV
- vague predicates as builtin predicates: $X \approx Y$
- `query(D):- Category(D,tv) & type(D,lcd) & size(D,X) & ≈(X,45)`

$X \approx Y$

β	X	Y
0.7	42	45
0.8	43	45
0.9	44	45
1.0	45	45
0.9	46	45
0.8	47	45
...	...	...

Data types and vague predicates in IR

Data type: domain + (vague) predicates

- Language (multilingual documents) / (language-specific stemming)
- Person names / “his name sounds like Jones”
- Dates / “about a month ago”
- Amounts / “orders exceeding 1 Mio $”
- Technical measurements / “at room temperature”
- Chemical formulas

Vague Criteria in Fact Databases

"I am looking for a 45-inch LCD TV with

- wide viewing angle
- high contrast
- low price
- high user rating"

→ vague criteria are very frequent in end-user querying of fact databases

→ but no appropriate support in SQL

Vague Criteria in Fact Databases

"I am looking for a 45-inch LCD TV with

- wide viewing angle
- high contrast
- low price
- high user rating"

→ vague criteria are very frequent in end-user querying of fact databases

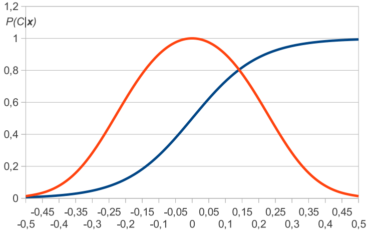
→ but no appropriate support in SQL

vague conditions → similar to fuzzy predicates

Probabilistic Modeling of Vague Predicates

[Fuhr 90]

- learn vague predicates from feedback data
- construct feature vector $\vec{x}(q_i, d_i)$ from query value q_i and document value d_i *(e.g. relative difference)*
- apply logistic regression



Expressiveness

- Retrieval Rules, Joins, Aggregations and Restructuring
- Expressiveness in XML Retrieval

Expressiveness
Formulating Retrieval Rules

```
about(D,T) :- docTerm(D,T).
```

Expressiveness
Formulating Retrieval Rules

```
about(D,T) :- docTerm(D,T).
```

consider document linking / anchor text

```
about(D,T) :- link(D1,D),about(D1,T).
```

Expressiveness
Formulating Retrieval Rules

```
about(D,T) :- docTerm(D,T).
```

consider document linking / anchor text

```
about(D,T) :- link(D1,D),about(D1,T).
```

consider term hierarchy

```
about(D,T) :- subconcept(T,T1) & about(D,T1).
```

Expressiveness
Formulating Retrieval Rules

```
about(D,T) :- docTerm(D,T).
```

consider document linking / anchor text

```
about(D,T) :- link(D1,D),about(D1,T).
```

consider term hierarchy

```
about(D,T) :- subconcept(T,T1) & about(D,T1).
```

field-specific term weighting

```
0.9 docTerm(D,T) :- occurs(D,T,title).
0.5 docTerm(D,T) :- occurs(D,T,body).
```

Expressiveness
Joins

IR authors:

```
irauthor(N):- about(D,ir) & author(D,N).
```

Expressiveness
Joins

IR authors:

```
irauthor(N):- about(D,ir) & author(D,N).
```

Smith's IR papers cited by Miller

```
?- author(D,smith) & about(D,ir) &
    author(D1,miller) & cites(D,D1).
```

Expressiveness
Aggregation (1)

Who are the major IR authors?

index

β	DNO	TERM
0.9	1	ir
0.8	1	db
0.6	2	ir
0.8	3	ir
0.7	3	ai

author

DNO	NAME
1	smith
2	miller
3	smith

irauthor

0.98	smith
0.6	miller

```
irauthor(A):- index(D,ir) & author(D,A).
```

Expressiveness
Aggregation (1)

Who are the major IR authors?

index

β	DNO	TERM
0.9	1	ir
0.8	1	db
0.6	2	ir
0.8	3	ir
0.7	3	ai

author

DNO	NAME
1	smith
2	miller
3	smith

irauthor

0.98	smith
0.6	miller

```
irauthor(A):- index(D,ir) & author(D,A).
```

Aggregation through projection!

Expressiveness
Aggregation (2)

Who are the major IR authors?

index

β	DNO	TERM
0.9	1	ir
0.8	1	db
0.6	2	ir
0.8	3	ir
0.7	3	ai

author

DNO	NAME
1	smith
2	miller
3	smith

irauths

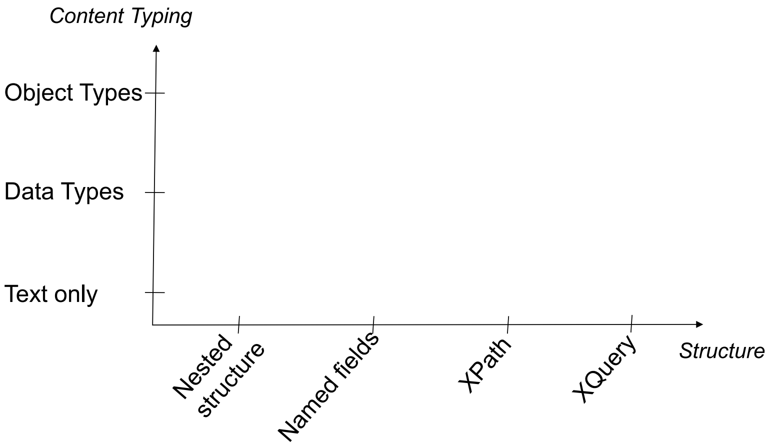
1.7	smith
0.6	miller

Aggregation through summing:

```
irauth(D,A):- index(D,ir) & author(D,A).
irauths SUM(Name) :- irdbauth(Doc,Name) | (Name)
```

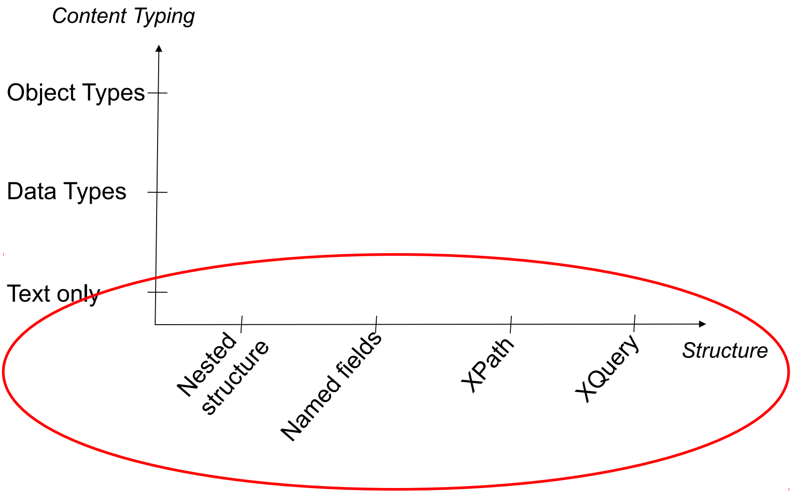
Expressiveness in XML Retrieval

[Fuhr & Lalmas 07]



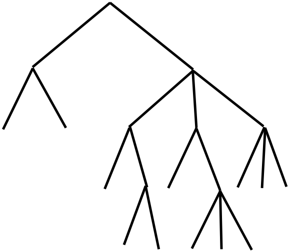
Expressiveness in XML Retrieval

[Fuhr & Lalmas 07]



XML structure: 1. Nested Structure

- XML document as hierarchical structure
- Retrieval of elements (subtrees)
- Typical query language does not allow for specification of structural constraints
- Relevance-oriented selection of answer elements: return the most specific relevant elements



XML structure: 2. Named Fields

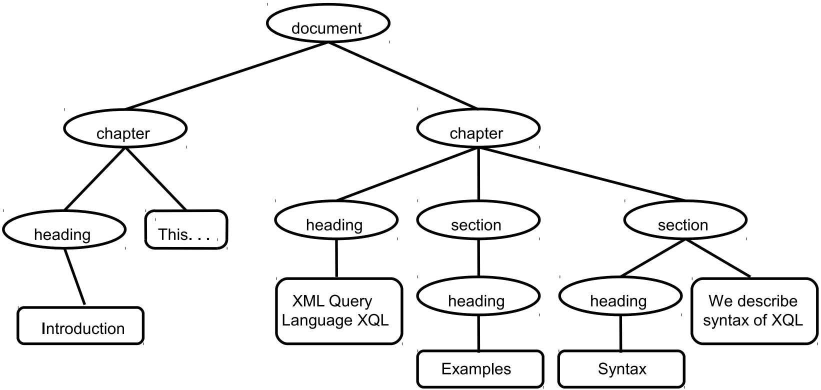
- Reference to elements through field names only
- Context of elements is ignored
  (e.g. author of article vs. author of referenced paper)
- Post-Coordination may lead to false hits
  (e.g. author name – author affiliation)

Example: Dublin Core

```
<oai_dc:dc xmlns:dc=
"http://purl.org/dc/elements/1.1/">
<dc:title>Generic Algebras
...  </dc:title>
<dc:creator>A. Smith (ESI),
B. Miller (CMU)</dc:creator>
<dc:subject>Orthogonal group,
Symplectic group</dc:subject>
<dc:date>2001-02-27</dc:date>
<dc:format>application/postscript</dc:
<dc:identifier>ftp://ftp.esi.ac.at/pub
<dc:source>ESI preprints
</dc:source>
<dc:language>en</dc:language>
</oai_dc:dc>
```

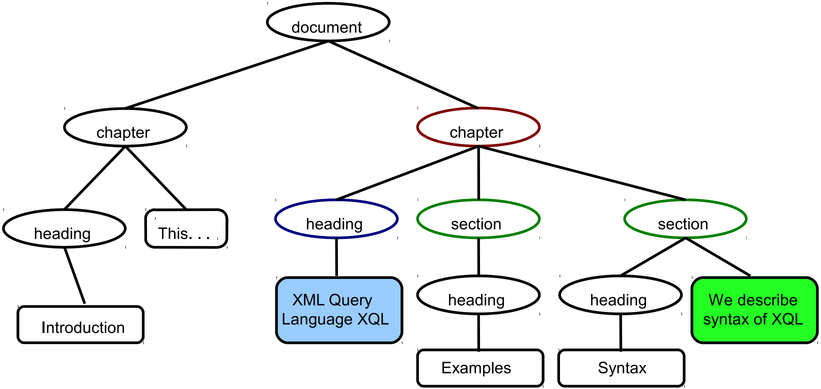
XML structure: 3. XPath

```
/document/chapter[about(./heading, XML) AND
about(./section//*,syntax)]
```



XML structure: 3. XPath

```
/document/chapter[about(./heading, XML) AND
about(./section//*,syntax)]
```



XML structure: 3. XPath (cont'd)

- Full expressiveness for navigation through document tree (+links)
  - Parent/child, ancestor/descendant
  - Following/preceding, following-sibling, preceding-sibling
  - Attribute, namespace
- Selection of arbitrary elements/subtrees
  (but anser can be only a single element of the originating document)

XML structure: 4. XQuery

Higher expressiveness, especially for database-like applications:

- Joins (trees $\rightarrow$ graphs)
- Aggregations
- Constructors for restructuring results

XML structure: 4. XQuery

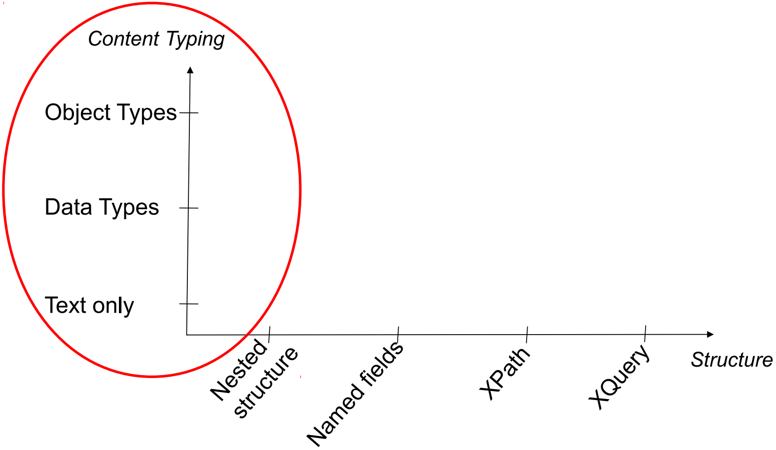
Higher expressiveness, especially for database-like applications:

- Joins (trees $\rightarrow$ graphs)
- Aggregations
- Constructors for restructuring results

Example: List each publisher and the average price of its books

```
FOR $p IN distinct(document("bib.xml")//publisher)
LET $a := avg(document("bib.xml")//book[publisher =
$p]/price)
RETURN
 <publisher>
 <name> $p/text() </name>
 <avgprice> $a </avgprice>
 </publisher>
```

XML content typing



XML content typing: 1. Text

```
  <book>
  <author>John Smith</author>
  <title>XML Retrieval</title>
  <chapter> <heading>Introduction</heading>
    This text explains all about XML and IR.
  </chapter>
  <chapter>
    <heading> XML Query Language XQL
    </heading>
    <section>
        <heading>Examples</heading>
    </section>
    <section>
        <heading>Syntax</heading>
        Now we describe the XQL syntax.
    </section>
  </chapter>
</book>
```

Example query

```
//chapter[about(.,
 XML query language]
```

XML content typing: 2. Data Types

- Data type: domain + (vague) predicates (see above)
- Close relationship to XML Schema, but
    - XMLS supports syntactic type checking only
    - No support for vague predicates

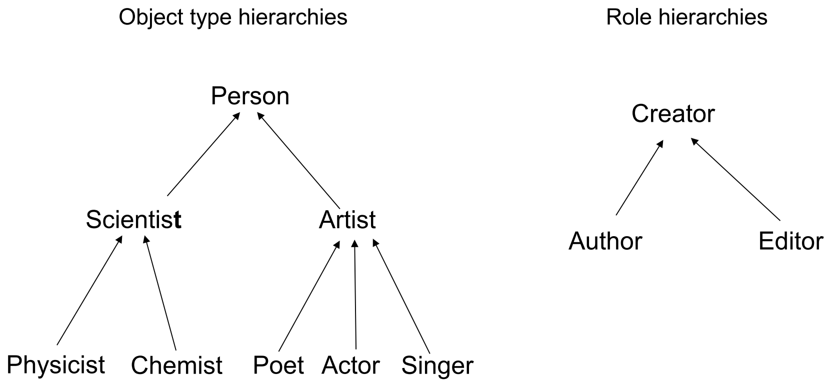
XML content typing: 3. Object Types
Based on Tagging / Named Entity Recognition

- Object types: Persons, Locations. Dates, ...

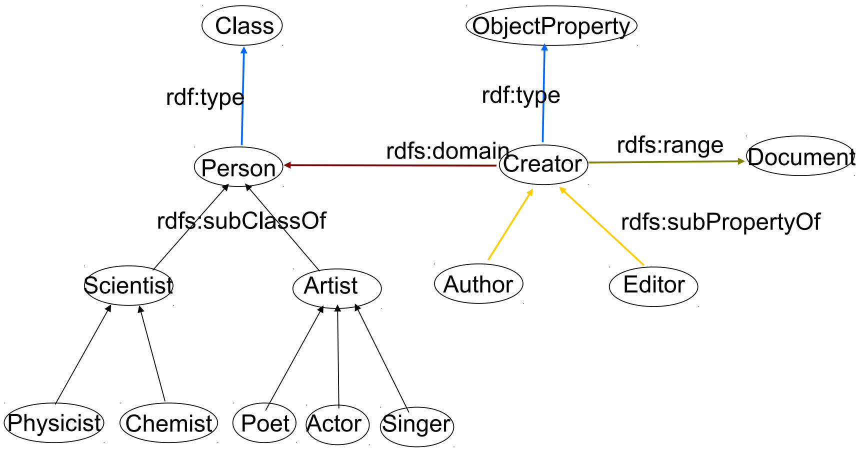
  Pablo Picasso (October 25, 1881 - April 8, 1973) was a Spanish painter and sculptor..... In Paris, Picasso entertained a distinguished coterie of friends in the Montmartre and Montparnasse quarters, including André Breton, Guillaume Apollinaire, and writer Gertrude Stein.

  To which other artists did Picasso have close relationships? Did he ever visit the USA?
- Named entity recognition methods allow for automatic markup of object types
- Object types support increased precision

XML content typing
Tag semantics modelled as hierarchies



XML content typing
Tag semantics modelled in OWL



Further Concepts

Further Concepts
4-valued (probabilistic) logics

Supported concepts

- conflicting knowledge
- open + closed world assumptions

Further Concepts
4-valued (probabilistic) logics

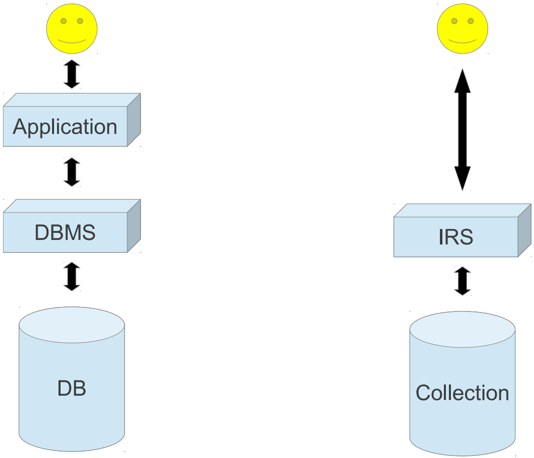
Supported concepts

- conflicting knowledge
- open + closed world assumptions

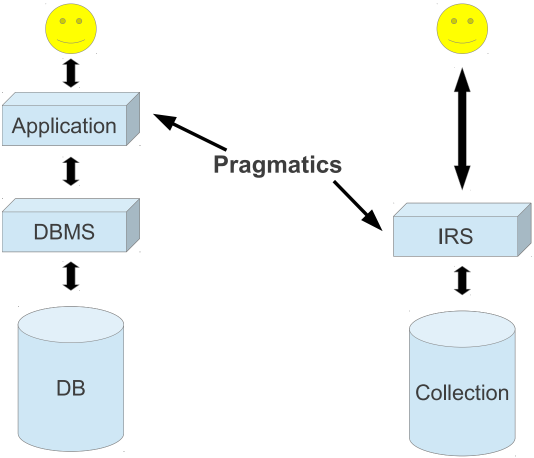
Applications

- 4-valued probabilistic Datalog [Fuhr & Roelleke 98]
- POOL: Probabilistic Object-Oriented Logic [Lalmas et al. 02]
- POLAR: Retrieval with Annotations [Frommholz & Fuhr 06]
- POLIS: Information summarization [Forst et al. 07]

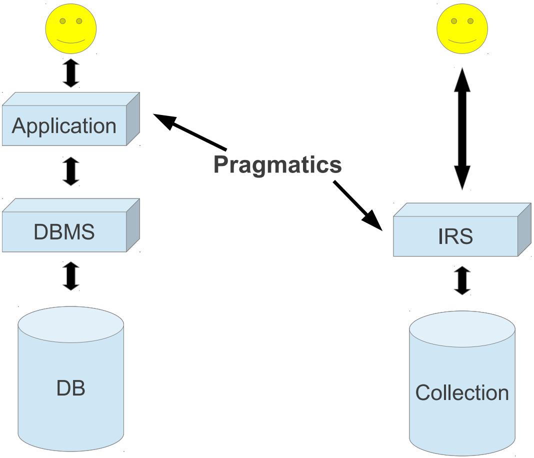
IR Systems vs. DBMS



IR Systems vs. DBMS

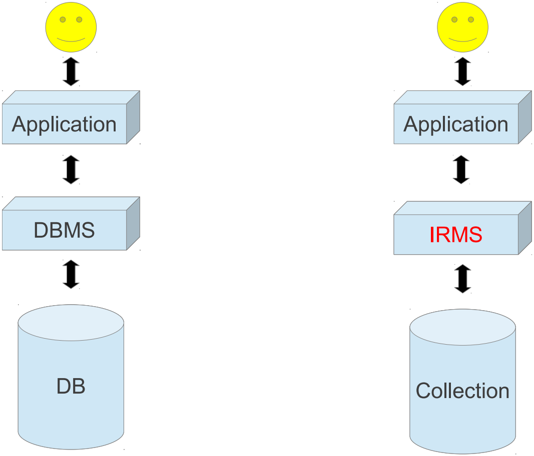


IR Systems vs. DBMS

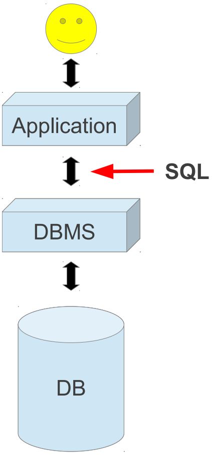


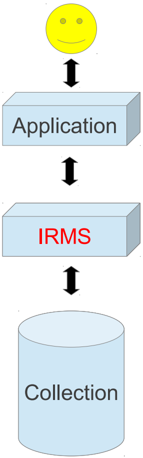
Separation between IRS and IR application?

Towards an IRMS

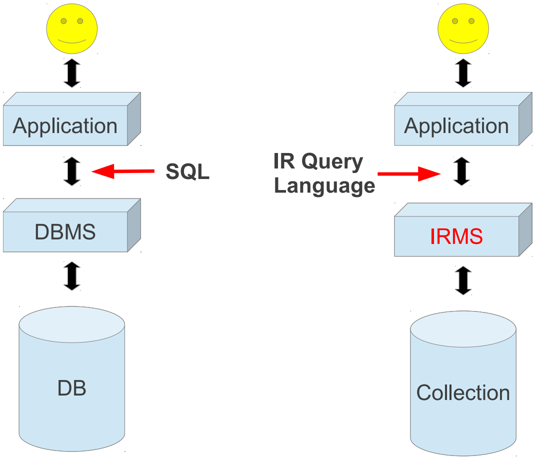


Towards an IRMS





Towards an IRMS



Conclusion

Conclusion

Inference

- Probabilistic relational model supports integration of IR+DB
- Probabilistic Datalog as powerful inference mechanism
- Allows for formulating retrieval strategies as logical rules

Conclusion

Inference

- Probabilistic relational model supports integration of IR+DB
- Probabilistic Datalog as powerful inference mechanism
- Allows for formulating retrieval strategies as logical rules

Vague predicates

- Natural extension of IR methods to attribute values
- Vague predicates can be learned from feedback data
- Transition from propositional to predicate logic

Conclusion

Inference

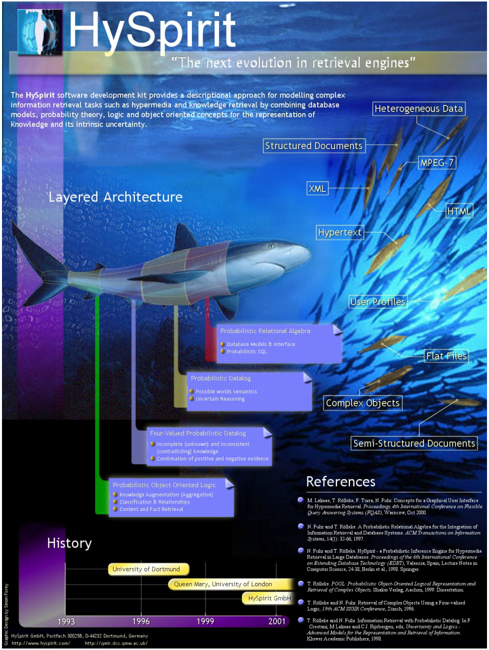
- Probabilistic relational model supports integration of IR+DB
- Probabilistic Datalog as powerful inference mechanism
- Allows for formulating retrieval strategies as logical rules

Vague predicates

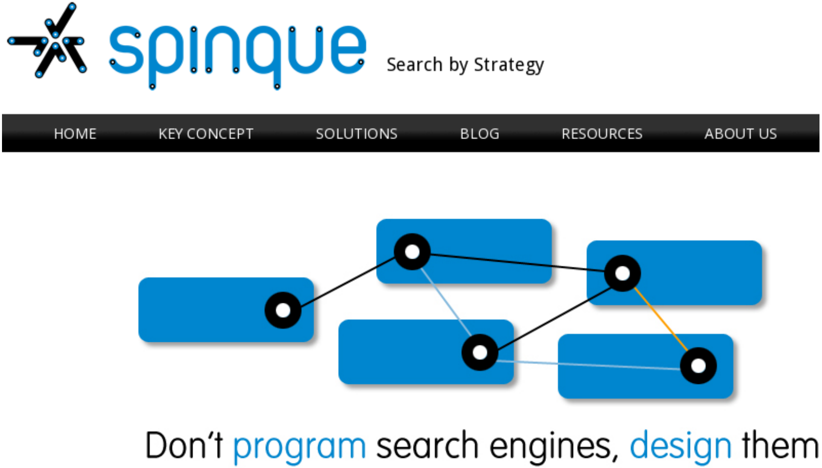
- Natural extension of IR methods to attribute values
- Vague predicates can be learned from feedback data
- Transition from propositional to predicate logic

Expressive query language

- Joins
- Aggregations
- (Re)structuring of results



http://www.eecs.qmul.ac.uk/~thor/



http://www.spinque.com/

References I

Dalvi, N. N.; Suciu, D.
(2007).
Efficient query evaluation on probabilistic databases.
VLDB J. 16(4), pages 523–544.

Forst, J. F.; Tombros, A.; Roelleke, T.
(2007).
POLIS: A Probabilistic Logic for Document Summarisation.
In: *Proceedings of the 1st International Conference on Theory of Information Retrieval (ICTIR 07) - Studies in Theory of Information Retrieval*, pages 201–212.

Frommholz, I.; Fuhr, N.
(2006).
Probabilistic, Object-oriented Logics for Annotation-based Retrieval in Digital Libraries.
In: Nelson, M.; Marshall, C.; Marchionini, G. (eds.): *Opening Information Horizons – Proc. of the 6th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2006)*, pages 55–64. ACM, New York.

References II

Fuhr, N.; Lalmas, M.
(2007).
Advances in XML retrieval: the INEX initiative.
In: *IWRIDL '06: Proceedings of the 2006 international workshop on Research issues in digital libraries*, pages 1–6. ACM, New York, NY, USA.

Fuhr, N.; Rölleke, T.
(1997).
A Probabilistic Relational Algebra for the Integration of Information Retrieval and Database Systems.
ACM Transactions on Information Systems 14(1), pages 32–66.

Fuhr, N.; Rölleke, T.
(1998).
HySpirit – a Probabilistic Inference Engine for Hypermedia Retrieval in Large Databases.
In: *Proceedings of the 6th International Conference on Extending Database Technology (EDBT)*, pages 24–38. Springer, Heidelberg et al.

References III

Fuhr, N.
(1990).
A Probabilistic Framework for Vague Queries and Imprecise Information in Databases.
In: *Proceedings of the 16th International Conference on Very Large Databases*, pages 696–707. Morgan Kaufman, Los Altos, California.

Fuhr, N.
(2000).
Probabilistic Datalog: Implementing Logical Information Retrieval for Advanced Applications.
Journal of the American Society for Information Science 51(2), pages 95–110.

Lalmas, M.; Roelleke, T.; Fuhr, N.
(2002).
Intelligent Hypermedia Retrieval.
In: Szczepaniak, P. S.; Segovia, F.; Zadeh, L. A. (eds.): *Intelligent Exploration of the Web*, pages 324–344. Springer, Heidelberg et al.

References IV

Pearl, J.
(1988).
Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.
Morgan Kaufman, San Mateo, California.

Rölleke, T.; Wu, H.; Wang, J.; Azzam, H.
(2007).
Modelling retrieval models in a probabilistic relational algebra with a new operator: the relational Bayes.
The International Journal on Very Large Data Bases (VLDB) 17(1), pages 5–37.

Suciu, D.; Olteanu, D.; Ré, C.; Koch, C.
(2011).
Probabilistic Databases.
Morgan & Claypool Publishers.