Slide 1

On Information Retrieval Models and DB+IR

Thomas Roelleke

Queen Mary, University of London

Apriorie Ltd

Max Planck Institute Saarbruecken, Oct 2006

Slide 2

Outline

- Motivation and Background
- A general matrix framework for IR
  notation re-used
- Probabilistic retrieval models and idf
- Parallel derivation of probabilistic retrieval models
- Modelling retrieval with DB+IR technology

Slide 3

Motivations

Implement IR models in high-level abstraction (mathematical and probabilistic logical), to support the engineering of customised information management applications.

To achieve this, understand the depth of IR models; what is common ground? Which general concepts do we need to model IR?

Slide 4

Background

Rijsbergen:CJ:1986: $P(d \rightarrow q)$
Wong/Yao:TOIS:1995: Probabilistic framework to explain IR modes
Fuhr:SIGIR:1996: Probabilistic Datalog (IP&M 2000)
Fuhr/Roelleke:TOIS:1997: PRA
Croft/Lafferty:2003: Language Modelling Book
Lafferty/Zhai:2003: Intro in LM Book
Hiemstra:JDlib:2000: Probabilistic interpretation of tf-idf
Roelleke:SIGIR:2003: Probability of being informative
Robertson:JDOC:2005: Understanding IDF: On theoretical arguments
deVries/Roelleke:SIGIR:2005: Relevance feedback: ”gain” for idf
Roelleke/etal:TREC:2005: PSQL
Roelleke/etal:IP&M:2006: General matrix framework
Roelleke/Wang:SIGIR:2006: Parallel derivation of IR models

Slide 5

A general matrix framework for IR

Spaces: collection c, document d, query q

Content: collection with document and term dimension, document with location and term dimension

DT_c **matrix,** LT_d **matrix**

Structure: collection/document with parent and child dimension

PC_c **matrix,** PC_d **matrix**

Evaluation: query with document and assessor dimension

DA_q **matrix**

Roelleke/etal:IPM:2006, more slides in Barcelona seminar talk

Slide 6

Content: The DT_c matrix of collection c

	sailing	boats	east	coast	$n_T(d,c)$
doc1	1	1			2
doc2		1	1		2
doc3	1	1			2
doc4	1				1
doc5	1		1	1	3
$n_D(t,c)$	4	3	2	1	

Note: $n_D(\cdot, c) = D^T \cdot DT_c$**:** D^T **is transpose of** D**,** $D = (1, 1, ...)$**.**

Slide 7

Notation - Notation - Notation

Motivation: A consistent and dual notation:

$n_D(t, c)$	**Number of documents in which term t occurs in collection c**
$N_D(c)$	**Number of documents in c**

Replace document dim D by location dim L	
$n_L(t, c)$	**Number of locations at which term t occurs in collection c**
$N_L(c)$	**Number of locations in c**

Slide 8

Notation - Notation - Notation

Replace collection space c by document space d	
$n_L(t, d)$	**Number of locations at which term t occurs in collection d**
$N_L(d)$	**Number of locations in d**

Slide 9

More Matrices? Yes!

- **Structure matrices PC_c (structure of collection c) and PC_d (structure of each document d)**
- **Evaluation matrices DA_q (document assessment per query)**

$$DD = DT \times DT^T, \quad TT = DT^T \times DT$$

DD**: Number of shared terms: Document similarity: co-containment**

TT**: Number of shared documents: Term similarity: co-occurrence**

Eigenvectors: $\lambda\vec{x} = A\vec{x}$**.**

Try for $\vec{d'} = TT \cdot \vec{d}$**.**

Slide 10

$P(d,q)$ and the trick with the diagonal

Remember $RSV = \vec{d}^T \cdot G \cdot \vec{q}$**?**

What about $RSV = \vec{d}^T \cdot IDF \cdot \vec{q}$**?**

$IDF = diag(\textbf{\textit{idf}}(\cdot))$ **is a diagonal matrix of *idf* values.**

$$IDF = \begin{bmatrix} \textbf{\textit{idf}}(sailing) & 0 & 0 & 0 \\ 0 & \textbf{\textit{idf}}(boats) & 0 & 0 \\ 0 & 0 & \textbf{\textit{idf}}(east) & 0 \\ 0 & 0 & 0 & \textbf{\textit{idf}}(coast) \end{bmatrix}$$

This is a valuable link to probabilistic models:

$P(d,q) := \sum_t P(d|t)P(q|t)P(t)$**,** $P(t) \propto \textbf{\textit{idf}}(t)$**.**

Slide 11

Matrix framework: Conclusion

- **Motivated by Wong/Yao:TOIS:1995: Link of vector-space model $\vec{d} \cdot \vec{q}$ and $P(q|d) = \sum_t P(q|t)P(t|d)$. Interpretations of $P(d \rightarrow q)$ to describe IR models. Matrix/vector algebra to describe IR concepts.**
- **Content, structure and evaluation in the same framework; parallel interpretations of co-containment, co-occurrence, co-citation, co-assessment, ...**
- **Mathematical/formal foundation for IR concepts (not just models)**

Slide 12

Probabilistic retrieval models and idf

Hiemstra:JDLib:2000, Robertson:JDOC:2005

$$RSV(d,q) := O(r|d,q) \propto \sum_{t \in d \cap q} \log \frac{P(t|r)P(\bar{t}|\bar{r})}{P(t|\bar{r})P(\bar{t}|r)}$$

$$\log \frac{1}{P(t|\bar{r})} = -\log P(t|\bar{r}) = -\log P(t|c) = \boldsymbol{idf}(t,c)$$

Vries/Roelleke:2005:

$$RSV(d,q) = \sum_{t \in d \cap q} -\boldsymbol{idf}(t,r) + \boldsymbol{idf}(t,\bar{r})$$

***idf*(t,r) in relevant *reduces* basic *idf*(t,c).**

Slide 13

Probability of being informative

$$\textbf{\textit{idf}}(t, c) := -\log P(t \textbf{\textit{ occurs}}|c)$$

Motivation: In a probabilistic reasoning system, we need probabilities proportional to *idf*. Interpretation?

$$e^{-\textbf{\textit{idf}}(t,c)} = P(t \textbf{\textit{ occurs}}|c)$$

$$P(t \textbf{\textit{ occurs}}|c) = \lim_{N \to \infty} \left(1 - \frac{\textbf{\textit{idf}}(t,c)}{N}\right)^N$$

$$P(t \textbf{\textit{ informs}}|c) := \frac{\textbf{\textit{idf}}(t,c)}{N}$$

Roelleke:SIGIR:2003, IR-Theory-Workshop:Glasgow-IR-Festival:2005

Slide 14

A parallel derivation of probabilistic IR models

Are there IR "quarks" that explain IR models, since origin is $P(r|d,q)$?

$$\begin{aligned}
RSV_{BIR}(d,q) &= \sum_{t \in d \cap q} \log \frac{P(t|r)P(\bar{t}|\bar{r})}{P(t|\bar{r})P(\bar{t}|r)} \\
RSV_{LM}(d,q) &= \sum_{t \in q} \log(\delta P(t|d) + (1-\delta)P(t|c)) \\
RSV_{PM}(d,q) &= \sum_{t \in d \cap q} \log \left(\frac{\lambda(t,r)}{\lambda(t,\bar{r})}\right)^{n_L(t,d)}
\end{aligned}$$

Note: We use δ for LM, since we reserve λ for Poisson.

Slide 15

Event spaces and probabilities

BIR	Poisson	LM
Judgements on Documents	**Frequencies of Terms**	**Terms at Locations**
$P_{BIR}(t\|c) := \frac{n_D(J=1, c_t)}{N_D(c_t)}$	$\lambda(t,c) := \frac{n_L(T=t,c)}{N_D(c)}$ $P_{PM}(t\|c) = \frac{\lambda^{n(t)}}{n(t)!}e^{\lambda}$	$P_{LM}(t\|c) := \frac{n_L(T=t,c)}{N_L(c)}$

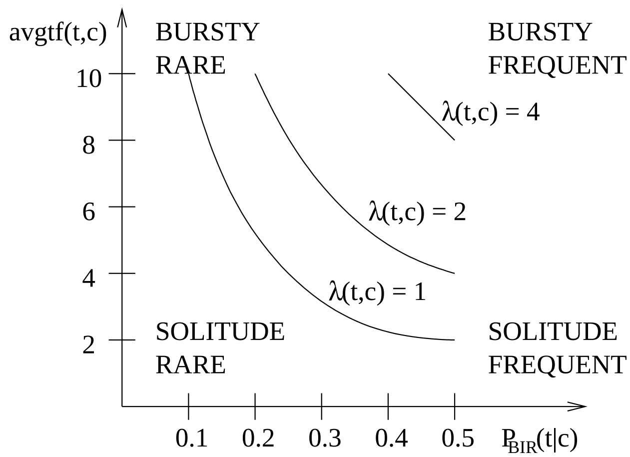
Slide 16

Poisson Bridge

$$
\begin{array}{ccccccccc}
P_{BIR}(t|c) & \cdot & ? & = & ? & \cdot & P_{LM}(t|c) \\
\frac{n_D(t,c)}{N_D(c)} & \cdot & ? & = & ? & \cdot & \frac{n_L(t,c)}{N_L(c)} \\
\frac{n_D(t,c)}{N_D(c)} & \cdot & \frac{n_L(t,c)}{n_D(t,c)} & = & \frac{N_L(c)}{N_D(c)} & \cdot & \frac{n_L(t,c)}{N_L(c)} \\
P_{BIR}(t|c) & \cdot & \textbf{\textit{avgtf}}(t,c) & = & \textbf{\textit{avgdl}}(c) & \cdot & P_{LM}(t|c) \\
 & \lambda(t,c) & & = & & \lambda(t,c) &
\end{array}
$$

Slide 17

Bursty and solitude terms



Slide 18

TF-IDF explanation

Take RSV_{PM} and Poisson bridge and obtain:

$$RSV_{PM}(d,q) = \sum_{t \in d \cap q} n_L(t,d) \cdot -\log \frac{P_{BIR}(t|\bar{r}) \cdot \boldsymbol{avgtf}(t,\bar{r})}{P_{BIR}(t|r) \cdot \boldsymbol{avgtf}(t,r)}$$

Compare to tf-idf:

$$RSV_{tfidf}(d,q) = \sum_{t} tf(t,d) \cdot -\log P_{BIR}(t|c)$$

Standard tf-idf ”drops” relevance, and assumes $\bar{r} = c$.

RSV_{PM} **shows how to incorporate relevance.**

Poisson bridge yields dual LM-based formulation.

Slide 19

Parallel derivation: Summary

- **Probability $P(r|d, q)$ origin of probabilistic models**
- **BIR, Poisson, and LM based on different event spaces**
- **Poisson bridge connects BIR and LM**
- **TF-IDF is close to Poisson model**
- **Poisson model and idf-based BIR formulation show effect of relevance**

Slide 20

DB+IR: Probability Aggregation

Probability aggregation in HySpirit/Apriorie PSQL:

```
CREATE VIEW retrieve AS
SELECT DISJOINT queryId, documentId
FROM weightedQuery, tf
WHERE weightedQuery.term = tf.term
TOP 10;
```

PRA basics in Fuhr/Roelleke:TOIS:1997, PSQL in Roelleke/etal:TREC:2005.

Slide 21

DB+IR: Probability Estimation

Probability estimation in HySpirit/Apriorie PSQL:

```
CREATE VIEW idf AS
SELECT term
FROM collection
ASSUMPTION MAX INFORMATIVE
EVIDENCE KEY ();
```

Slide 22

DB+IR Demo

```
<par>Tweety is a bird<par>
<par>Tweety is not a bird<par>
```

```
# POOL
doc_1 [
  par_1 [ 0.6/0.2 bird(tweety) ]
  par_2 [ NOT bird(tweety) ]
]
?- D[ bird(X) ]
?- D [ NOT bird(X) ]
?- bird(X)
```

Slide 23

Summary and Conclusions

- General matrix framework: notation and framework to describe IR concepts such as frequencies, ranking models, authorities, evaluation, etc
- Probabilistic models and idf: BIR and idf related Poisson model explains tf-idf, Poisson bridge leads to dual notation either based on BIR or LM parameters
- DB+IR: high-level, abstract implementation of IR concepts to realise customised IR applications at low-costs (Ralf: It was easy with Oracle ...)

Slide 24

What is going on?

- Dalvi/Suciu/etal: semantics in probabilistic databases
- MPI Saarbruecken: top-k
- deVries@cwi: efficient DB technology for IR; matrix framework
- Frommholz@duisburg: annotation logic POLAR
- Heng Zhi Wu, Hany Azzam: Efficient processing of PRA, query optimisation
- Jun Wang: Retrieval models, context-specific idf in structured document retrieval
- Frederik Forst: Summarisation logic POLIS (based on POOL, Kripke structured, description logic)
- Follow-up of SIGIR Sheffield 2004 DB+IR workshop?