Slide 1

On Occurrence and Informativeness Probabilities

IR Festival Glasgow 2005

Thomas Rölleke

Queen Mary University of London

Department of Computer Science

Slide 2

Outline

Slide 3

Motivation: Basics and Questions

- Theoretical explanation for *idf*?
- *idf* as a probabilistic estimate?
- Occurrence probability: n/N or other?
- *EFFECTIVE* DB+IR?
  - *idf* $\rightarrow$ relational model / SQL?
  - Scalability?

Slide 4

idf

$$P(t|c) := \frac{n(t,c)}{N(c)}$$
$$\textit{idf}(t,c) := -\log P(t|c)$$

A piece of IR granite.

Variations? Alternative distribution for $P(t|c)$ (DFR site).

Historical note

The first publication on the natural log was in 1614, paper by John Napier, 1550-1618, Scottish mathematician and astrologer.

Inventor of log: Joost Buergi, 1552-1632, swiss clock maker.

Slide 5

The Probability of Relevance and the BIRM

$P(r|d,q)$: Foundation for the BIRM and language modelling.

BIRM: After a number of steps, "tricks" and assumptions:

$$\sum_t \log \frac{P(t|r) \cdot P(\bar{t}|\bar{r})}{P(\bar{t}|r) \cdot P(t|\bar{r})}$$

Another piece of IR granite.

Slide 6

idf-*based Formulation of the BIRM*

Robertson:2004: *idf* is estimate for BIRM term weight if no relevance information is available.

$$\log P(t|r) - \log P(t|\bar{r}) = -\mathit{idf}(t,r) + \mathit{idf}(t,\bar{r})$$

Joins two pieces of IR.

t occurs in all relevant docs $\iff$ $\mathit{idf}(t,r) = 0$.

To be found in SIGIR:2005.

Slide 7

The Probability of Being Informative

$$P(t \textit{ occurs}|c) := \frac{n(t,c)}{N(c)} \text{ or alternative}$$

$$P(t \textit{ informs}|c) := \text{inverse to occurrence}$$

Explains $P(t \textit{ informs})$.

Occurrence-Informativeness Theorem:

$$P(t \textit{ informs}|c) = \frac{-\log P(t \textit{ occurs}|c)}{M} \iff$$

$$P(t \textit{ occurs}|c) = \lim_{M\to\infty} (1 - P(t \textit{ informs}|c))^M$$

Proof: $e^{-\lambda} = \lim_{M\to\infty} \left(1 - \frac{\lambda}{M}\right)^M$

Slide 8

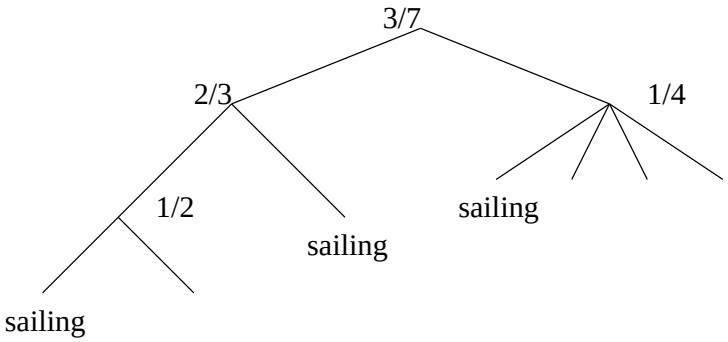
Poisson-based idf *(occurrence)* *Lift it.*

	Linear estimate	Poisson-based estimate
Occurrence (Documents)	$P_D(t\|c) := \frac{n_D(t,c)}{N_D(c)}$	$P_D(t\|c) := \frac{n_D(t,c)}{K_D(c) + n_D(t,c)}$
Within-document occurrence (Locations)	$P_L(t\|d) := \frac{n_L(t,d)}{n_L(t_{max},d)}$	$P_L(t\|d) := \frac{n_L(t,d)}{K_L(d) + n_L(t,d)}$

Notation and dualities in general matrix framework for IR (IP&M).

Slide 9

Context-specific idf



Slide 10

idf *in the Probabilistic Relational Algebra and SQL of HySpirit/Apriorie*

df = Project{all}[$1](collection);	INSERT INTO df SELECT term FROM collection;
idf = Bayes{max_idf}[](df);	INSERT INTO idf SELECT term FROM df ASSUMPTION MAX_IDF;

Slide 11

A GUI for Making Theory Work

Slide 12

Summary

- Robertson:JDOC:2004: BIRM is explanation for *idf*
- *idf*-based formulation of BIRM
- $P(t$ *informs*$)$: Semantics based on semantics of $\log$
- Poisson-based *idf*: Improves retrieval quality for long queries
- Context-specific *idf*: Solution for structured document retrieval
- HySpirit/Apriorie: Frequency-based and *idf*-based probability estimation integral part of Probabilistic Relational Model / SQL

Slide 13

Conclusions

- The *idf*-granite is hard (http://www.soi.city.ac.uk/ ser/idf.html, see relationship of idf and language modelling, Hiemstra, Nie).
- Lifting the occurrence probability appears to be a good idea (DFR, P_{risk} Amati/Rijsbergen)
- Recent experience shows: For increasing the impact of IR research, we need to
  - make IR theory applicable *AND* available to IR externals
  - integrate IR with other systems / research areas (e.g. bio-informatics, law enforcement), not vice versa

Slide 14

Outlook

- Occurrence-informativeness theorem (noise versus informativeness, Belew:2000 book)
- Structured IR: context-specific *idf*
- Efficiency/Scalability: special, probabilistic, relational indexing structures and relaxed fix-point semantics for ultimate scalability
- Knowledge-based reasoning: log-based negation
- Non-linear (chaotic) behaviour of retrieval functions