On Occurrence and Informativeness Probabilities
IR Festival Glasgow 2005

Thomas Rölleke
Queen Mary University of London
Department of Computer Science

Outline
• Motivation: Basics and Questions 3
• $idf$ 4
• $P(r|d, q)$ and BIRM 5
• $idf$-based Formulation of the BIRM 6
• The Probability of Being Informative 7
• Poisson-based $idf$ 8
• Context-specific $idf$ 9
• Probability estimation and $idf$ in the relational model / SQL 10
• A GUI for making theory work 11
• Summary, Conclusions and Outlook 12, 13, 14
Motivation: Basics and Questions

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

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

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 Bürgi, 1552-1632, Swiss clock maker.
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:

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

Another piece of IR granite.

idf-based Formulation of the BIRM

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

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

Joins two pieces of IR.

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

To be found in SIGIR: 2005.
The Probability of Being Informative

\[ P(t \text{ occurs}|c) := \frac{n(t, c)}{N(c)} \quad \text{or alternative} \]

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

Occurrence-Informativeness Theorem:

\[ P(t \text{ informs}|c) = \frac{-\log P(t \text{ occurs}|c)}{M} \iff \]

\[ P(t \text{ occurs}|c) = \lim_{M \to \infty} (1 - P(t \text{ informs}|c))^M \]

Proof:

\[ e^{-\lambda} = \lim_{M \to \infty} \left(1 - \frac{\lambda}{M}\right)^M \]

---

Poisson-based idf (occurrence)

<table>
<thead>
<tr>
<th>Linear estimate</th>
<th>Poisson-based estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occurrence (Documents)</td>
<td>[ P_D(t</td>
</tr>
<tr>
<td></td>
<td>[ P_D(t</td>
</tr>
<tr>
<td>Within-document occurrence (Locations)</td>
<td>[ P_L(t</td>
</tr>
<tr>
<td></td>
<td>[ P_L(t</td>
</tr>
</tbody>
</table>

Notation and dualities in general matrix framework for IR (IP&M).
**Context-specific idf**

![Diagram showing context-specific idf calculation]

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

**Slide 10**

- \( df = \text{Project\{all\}}[1](\text{collection}); \)
- \( \text{INSERT INTO df} \)
  \( \text{SELECT term} \)
  \( \text{FROM collection}; \)
- \( \text{idf} = \text{Bayes\{max\_idf\}}[1](df); \)
- \( \text{INSERT INTO idf} \)
  \( \text{SELECT term} \)
  \( \text{FROM df} \)
  \( \text{ASSUMPTION MAX\_IDF}; \)
Summary

- Robertson:JDOC:2004: BIRM is explanation for idf
- idf-based formulation of BIRM
- $P(t \text{ informs})$: 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
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

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