Harmony Assumptions: Extending Probability Theory for Information Retrieval (IR) and for Databases (DB) and for Knowledge Management (KM) and for Machine Learning (ML) and for Artificial Intelligence (AI)


Thomas Roelleke
Queen Mary University of London
1. Outline: 17 slides
2. Introduction
3. TF-IDF
4. TF Quantifications
5. Harmony Assumptions
6. Experimental Study: IR and Social Networks
7. Impact
8. Summary
9. Background
Probability Theory: Independence Assumption

\[ P(\text{sailing, boats, sailing}) = P(\text{sailing})^2 \cdot P(\text{boats}) \]

Applied in AI, DB and IR
and “Big Data” and “Data Science” and ...

TF-IDF

- the best known ranking formulae?
- known in IR, DB and AI and other disciplines?
- TF-IDF and probability theory?

\[ \log(P(\text{sailing, boats, sailing})) = 2 \cdot \log(P(\text{sailing})) + \ldots \]

- TF-IDF and LM (language modelling)?
Probability Theory: Independence Assumption

\[ P(\text{sailing, boats, sailing}) = P(\text{sailing})^2 \cdot P(\text{boats}) \]

Applied in AI, DB and IR
and “Big Data” and “Data Science” and ...

TF-IDF

- the best known ranking formulae?
- known in IR, DB and AI and other disciplines?
- TF-IDF and probability theory?

\[ \log (P(\text{sailing, boats, sailing})) = 2 \cdot \log (P(\text{sailing})) + \ldots \]

- TF-IDF and LM (language modelling)?
Research on foundations required for ...

Abstraction: DB+IR+KM+ML: probabilistic logical programming

# Probabilistic facts and rules are great, BUT ...
# one needs more expressiveness.

# For example:
# \[ P(t|d) = \frac{tf_d}{doclen} \]
\texttt{p.t.d SUM(T,D) :- term_doc(T,D)|(D);}

extended probability theory → DB+IR+KM+ML on the road
- a search for the missing science of consciousness

Preface: dad and daughter enter a cave:
- “Dad, that boulder at the entrance, if it comes down, we are locked in.”
- “Well, it stood there the last 10,000 years, so it won’t fall down just now.”
- “Dad, will it fall down one day?”
- “Yes.”
- “So it is more likely to fall down with every day it did not fall down?”

Taxi: on average, 1/6 taxis are free
busy busy ... after 7 busy taxis, keep waiting or give up?
TF-IDF

$$RSV_{TF-IDF}(d, q) := \sum_t TF(t, d) \cdot TF(t, q) \cdot IDF(t)$$

- How can someone spend 10 years looking at the equation?
- Maybe because of what Norbert Fuhr said:

  *We know why TF-IDF works; we have no idea why LM (language modelling) works.*

$$RSV_{LM}(d, q) \propto \frac{P(q|d)}{P(q)} \quad RSV_{TF-IDF}(d, q) \propto \frac{P(d|q)}{P(d)}$$
Harmony Assumptions: Extending Probability Theory

**TF-IDF**

\[
\text{RSV}_{\text{TF-IDF}}(d, q) := \sum_t \text{TF}(t, d) \cdot \text{TF}(t, q) \cdot \text{IDF}(t)
\]

- How can someone spend 10 years looking at the equation?
- Maybe because of what Norbert Fuhr said:
  
  *We know why TF-IDF works; we have no idea why LM (language modelling) works.*

\[
\text{RSV}_{\text{LM}}(d, q) \propto \frac{P(q|d)}{P(q)} \quad \text{RSV}_{\text{TF-IDF}}(d, q) \propto \frac{P(d|q)}{P(d)}
\]
% A document:
d1[sailing boats are sailing with other sailing boats in greece ...]

\[ w_{\text{TF-IDF}}(\text{sailing}, d1) = \text{TF}(\text{sailing}, d1) \cdot \text{IDF}(\text{sailing}) = 3 \cdot \log \frac{1000}{10} = 3 \cdot 2 = 6 \]

\[ w_{\text{TF-IDF}}(\text{boats}, d1) = \text{TF}(\text{boats}, d1) \cdot \text{IDF}(\text{boats}) = 2 \cdot \log \frac{1000}{1} = 2 \cdot 3 = 6 \]

**NOTE:**  
\[ w_{\text{TF-IDF}}(\text{sailing}, d1) = w_{\text{TF-IDF}}(\text{boats}, d1) \]

- Both terms have the same impact on the score of d1!
- The rare term should have MORE impact than the frequent one!
TF quantifications

Theoretical Justifications?!?!

\[ TF(t, d) := \begin{cases} 
& tf_d & \text{total TF: independence!} \\
& 1 + \log(tf_d) & \text{log TF: dependence?} \\
& \log(tf_d + 1) & \text{another log TF} \\
& \frac{tf_d}{(tf_d + K_d)} & \text{BM25 TF: dependence?} 
\end{cases} \]

\( K_d \): pivoted document length: \( K_d > 1 \) for long documents ...

- Experimental results:
  - log-TF much better than total TF (ltc, [Lewis, 1998])
  - BM25-TF better than log-TF

- Theoretical results?
  
  \textit{Why? Wieso - Weshalb - Warum?}
Harmony Assumptions: Extending Probability Theory

**TF Quantifications**

**BM25-TF**

![Graph showing TF-BM25 equation](image)

\[ TF_{BM25}(t, d) := \frac{tf_d}{tf_d + K_d} \]
Remember Naive TF-IDF? Now, try BM25-TF-IDF:

\[
\begin{align*}
\text{BM25-TF-IDF}(\text{sailing}, d1) &= \frac{3}{3 + 1} \cdot \log \frac{1000}{10} = \frac{3}{4} \cdot 2 = 1.5 \\
\text{BM25-TF-IDF}(\text{boats}, d1) &= \frac{2}{2 + 1} \cdot \log \frac{1000}{1} = \frac{2}{3} \cdot 3 = 2
\end{align*}
\]

**IMPORTANT:**

\[
\text{BM25-TF-IDF}(\text{sailing}, d1) < \text{BM25-TF-IDF}(\text{boats}, d1)
\]
Series-based explanations of the TF quantifications:

\[ TF_{\text{total}} \quad tf_d = 1 + 1 + \ldots + 1 \]

\[ TF_{\log} \quad 1 + \log (tf_d) \approx 1 + \frac{1}{2} + \ldots + \frac{1}{tf_d} \]

\[ TF_{\text{BM25}} \quad \frac{tf_d}{tf_d + 1} = \frac{1}{2} \cdot \left[ 1 + \frac{1}{1+2} + \ldots + \frac{1}{1+2+\ldots+tf_d} \right] \]
FORGET Information Retrieval

...  
BACK TO Probability Theory
Harmony Assumptions: Extending Probability Theory

Harmony Assumptions

\[
P^{(\text{sailing, \ldots})} = \frac{1}{\Omega} \cdot P^{(\text{sailing})^k} = \frac{1}{\Omega} \cdot P^{(\text{sailing})^{1+1+\ldots+1}}
\]

\[
P_\alpha^{(\text{sailing, \ldots})} = \frac{1}{\Omega} \cdot P^{(\text{sailing})^{1+\frac{1}{2\alpha}+\ldots+\frac{1}{k\alpha}}}
\]

- independent: \( \alpha = 0 \)
- square-root-harmonic: \( \alpha = 0.5 \)
- naturally harmonic: \( \alpha = 1 \)
- square-harmonic: \( \alpha = 2 \)
- ...

\( \Omega: \) Later
<table>
<thead>
<tr>
<th>assumption name</th>
<th>assumption function $af(n)$</th>
<th>description / comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>zero harmony</td>
<td>$1 + \frac{1}{2^0} + \ldots + \frac{1}{n^0}$</td>
<td>independence: $1+1+1+\ldots+1$</td>
</tr>
<tr>
<td>natural harmony</td>
<td>$1 + \frac{1}{2} + \ldots + \frac{1}{n}$</td>
<td>harmonic sum</td>
</tr>
<tr>
<td>alpha-harmony</td>
<td>$1 + \frac{1}{2^\alpha} + \ldots + \frac{1}{n^\alpha}$</td>
<td>generalised harmonic sum</td>
</tr>
<tr>
<td>sqrt harmony</td>
<td>$1 + \frac{1}{2^{1/2}} + \ldots + \frac{1}{n^{1/2}}$</td>
<td>$\alpha = 1/2; \text{ divergent}$</td>
</tr>
<tr>
<td>square harmony</td>
<td>$1 + \frac{1}{2^2} + \ldots + \frac{1}{n^2}$</td>
<td>$\alpha = 2; \text{ convergent: } \frac{\pi^2}{6} \approx 1.645$</td>
</tr>
<tr>
<td>Gaussian harmony</td>
<td>$2 \cdot \frac{n}{n+1} = 1 + \frac{1}{1+2} + \ldots + \frac{1}{1+\ldots+n}$</td>
<td>explains the BM25-TF $\frac{tf_d}{tf_d+pivdl}$</td>
</tr>
</tbody>
</table>
independent: $\alpha = 0$

$0.5 \cdot 0.5 = 0.25$

sqrt-harmonic: $\alpha = 1/2$

$0.5 \cdot 0.5^{1/\sqrt{2}} \approx 0.306$

naturally harmonic: $\alpha = 1$

$0.5 \cdot 0.5^{1/2} \approx 0.353$

The area of each circle corresponds to the single event probability: $p = 0.5$.
The overlap becomes larger for growing $\alpha$ (harmony).
Africa in TREC-3

742, 611 = 734, 078 + 8, 533

<table>
<thead>
<tr>
<th>k</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{\text{obs}}$ documents</td>
<td>0.9885</td>
<td>0.0062</td>
<td>0.0019</td>
<td>0.0011</td>
<td>0.0007</td>
<td>0.0005</td>
<td>0.0004</td>
<td>0.0002</td>
<td>0.0002</td>
</tr>
<tr>
<td>$P_{\text{binomial}}$</td>
<td>0.9738</td>
<td>0.0258</td>
<td>0.0003</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$P_{\alpha\text{-harmonic}, \alpha=0.41}$</td>
<td>0.9787</td>
<td>0.018</td>
<td>0.0023</td>
<td>0.0005</td>
<td>0.0002</td>
<td>0.0001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Binomial assumes independence:

- $P_{\text{binomial}}(1) > P_{\text{obs}}(1)!$
- $P_{\text{binomial}}(2) < P_{\text{obs}}(2)!$
- $P_{\text{binomial}}(3) = 0!$
Distribution of alpha’s: for many terms, $0.3 \leq \alpha \leq 0.8$. Sqrt-harmony appears to be a good default assumption.
Extended Probability Theory

applicable in DB+IR+KM+ML + other disciplines where probabilities and ranking are involved.

DB+IR+KM+ML: A new generation

1. \( w_{BM25}(\text{Term}, \text{Doc}) = -tf_d(\text{Term}, \text{Doc}) \cdot BM25 \cdot \text{piv}_{dl}(\text{Doc}) \);

2. \# w_{BM25}: a probabilistic variant of the BM25–TF weight.

4. \# What to add for modelling ranking algorithms (TF-IDF, BM25, LM, DFR)?

6. \# What makes engineers happy???

[Frommholz and Roelleke, 2016]: DB Spektrum
The Independence Assumption: easy and scales, BUT ...!!!

Many disciplines rely on probability theory.

Between Disjointness and Subsumption, there is more than Independence.

For example:

- Natural Harmony: \( \log_2(k + 1) \)
- Gaussian Harmony: \( 2 \cdot \frac{k}{k + 1} \)

BM25-TF: \[
2 \cdot \frac{tf_d}{tf_d + 1} = 1 + \frac{1}{1+2} + \cdots + \frac{1}{1+2+\cdots+tf_d}
\]
Other theories to model dependencies?

Questions?
Harmony Assumptions: Extending Probability Theory

---

Background

[Fagin and Halpern, 1994]: Reasoning about Knowledge and Probabilities
[Church and Gale, 1995a, Church and Gale, 1995b]: IDF ...
[Fuhr and Roelleke, 1997]: PRA (bibdb: Fuhr/Roelleke:94! 3 years!)
[Lewis, 1998]: Naive Bayes at Forty: The Independence Assumption in Information Retrieval
[Roelleke, 2003]: The Probability of Being Informative ... idf/maxidf
[Robertson, 2004]: On theoretical arguments for IDF
[Robertson, 2005]: Event spaces
[Roelleke and Wang, 2006, Roelleke and Wang, 2008]: ...
[Roelleke et al., 2008]: The Relational Bayes: ...
[Roelleke et al., 2013]: Modelling Ranking Algorithms in PDataatalog
[Roelleke, 2013]: IR Models: Foundations & Relationships
[Roelleke et al., 2015]: Harmony Assumptions in IR and Social Networks
[Frommholz and Roelleke, 2016]: Scalable DB+IR Tech: ProbDataatalog with HySpirit

red thread between IR Theory and abstraction for DB+IR
Background

Inverse document frequency (idf): A measure of deviation from Poisson.
In *Proceedings of the Third Workshop on Very Large Corpora*, pages 121–130.

Poisson mixture.

Reasoning about knowledge and probability.

Scalable DB+IR technology: Processing probabilistic datalog with hyspirit.

A probabilistic relational algebra for the integration of information retrieval and database systems.

Naive (Bayes) at forty: The independence assumption in information retrieval.

Understanding inverse document frequency: On theoretical arguments for idf.

On event spaces and probabilistic models in information retrieval.
Background

A frequency-based and a Poisson-based probability of being informative.


In *Proceedings of the 7th International Workshop on Ranking in Databases (DBRank)*. ACM.

Harmony assumptions in information retrieval and social networks.

A parallel derivation of probabilistic information retrieval models.
In *ACM SIGIR*, pages 107–114, Seattle, USA.

TF-IDF uncovered: A study of theories and probabilities.

Modelling retrieval models in a probabilistic relational algebra with a new operator: The relational Bayes.