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)

Lernen. Wissen. Daten. Analysen. LWDA Potsdam, September 2016

Thomas Roelleke
Queen Mary University of London

Probability Theory: Independence Assumption

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

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?

$$\log\left(P(\text{sailing}, \text{boats}, \text{sailing})\right) = 2 \cdot \log\left(P(\text{sailing})\right) + \ldots$$

- TF-IDF and LM (language modelling)?

Probability Theory: Independence Assumption

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

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?

$$\log\left(P(\text{sailing}, \text{boats}, \text{sailing})\right) = 2 \cdot \log\left(P(\text{sailing})\right) + \ldots$$

- 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) = tf_d / doclen
p_t_d SUM(T,D) :- term_doc(T,D)|(D);
```

extended probability theory $\rightarrow$ 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

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

- 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.

$$\mathrm{RSV}_{\mathrm{LM}}(d,q) \overset{!!!}{\propto} \frac{P(q|d)}{P(q)} \qquad \mathrm{RSV}_{\text{TF-IDF}}(d,q) \overset{???}{\propto} \frac{P(d|q)}{P(d)}$$

TF-IDF

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

- 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.

$$\mathrm{RSV}_{\mathrm{LM}}(d,q) \overset{!!!}{\propto} \frac{P(q|d)}{P(q)} \qquad \mathrm{RSV}_{\text{TF-IDF}}(d,q) \overset{???}{\propto} \frac{P(d|q)}{P(d)}$$

% A document:
d1[sailing boats are sailing with other sailing boats in greece ...]

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

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

NOTE:

$$w_{\text{TF-IDF}}(\text{sailing}, \text{d1}) = w_{\text{TF-IDF}}(\text{boats}, \text{d1})$$

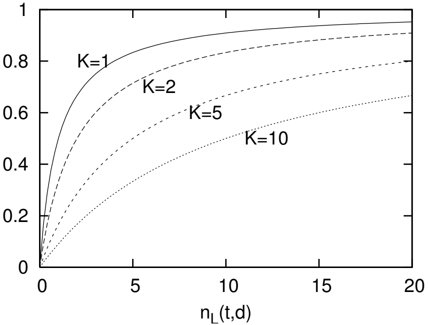
- Both terms have the same impact on the score of d1!
- The rare term should have MORE impact than the frequent one!

$$\mathrm{TF}(t, d) := \begin{cases} \mathrm{tf}_d & \text{total TF: independence!} \\ 1 + \log(\mathrm{tf}_d) & \text{log TF: dependence?} \\ \log(\mathrm{tf}_d + 1) & \text{another log TF} \\ \mathrm{tf}_d/(\mathrm{tf}_d + K_d) & \text{BM25 TF: dependence?} \end{cases}$$

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?

Why? Wieso - Weshalb - Warum?



$$TF_{\mathrm{BM25}}(t, d) := \frac{\mathrm{tf}_d}{\mathrm{tf}_d + K_d}$$

Remember Naive TF-IDF? Now, try BM25-TF-IDF:

$$w_{\text{BM25-TF-IDF}}(\text{sailing}, \text{d1}) = \frac{3}{3+1} \cdot \log \frac{1000}{10} = \frac{3}{4} \cdot 2 = 1.5$$

$$w_{\text{BM25-TF-IDF}}(\text{boats}, \text{d1}) = \frac{2}{2+1} \cdot \log \frac{1000}{1} = \frac{2}{3} \cdot 3 = 2$$

IMPORTANT:

$$w_{\text{BM25-TF-IDF}}(\text{sailing}, \text{d1}) < w_{\text{BM25-TF-IDF}}(\text{boats}, \text{d1})$$

Series-based explanations of the TF quantifications:

$$\mathrm{TF}_{\mathrm{total}} \qquad \mathrm{tf}_d = 1 + 1 + ... + 1$$

$$\mathrm{TF}_{\log} \qquad 1 + \log(\mathrm{tf}_d) \approx 1 + \frac{1}{2} + \ldots + \frac{1}{\mathrm{tf}_d}$$

$$\mathrm{TF}_{\mathrm{BM25}} \qquad \frac{\mathrm{tf}_d}{\mathrm{tf}_d + 1} = \frac{1}{2} \cdot \left[1 + \frac{1}{1+2} + \ldots + \frac{1}{1+2+...+\mathrm{tf}_d}\right]$$

FORGET Information Retrieval

...

BACK TO Probability Theory

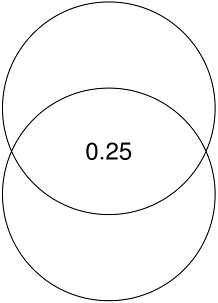
$$P(\overbrace{\text{sailing}, \dots}^{k}) = \frac{1}{\Omega} \cdot P(\text{sailing})^k = \frac{1}{\Omega} \cdot P(\text{sailing})^{1+1+\dots+1}$$

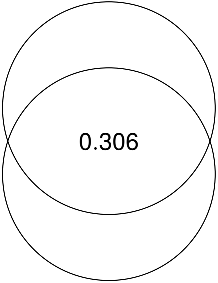
$$P_\alpha(\overbrace{\text{sailing}, \dots}^{k}) = \frac{1}{\Omega} \cdot P(\text{sailing})^{1+\frac{1}{2^\alpha}+\dots+\frac{1}{k^\alpha}}$$

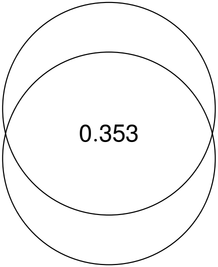
- independent: $\alpha = 0$
- square-root-harmonic: $\alpha = 0.5$
- naturally harmonic: $\alpha = 1$
- square-harmonic: $\alpha = 2$
- ...

Ω: Later

assumption name	assumption function $\mathrm{af}(n)$	description / comment
zero harmony	$1 + \frac{1}{2^0} + \ldots + \frac{1}{n^0}$	independence: 1+1+1+...+1
natural harmony	$1 + \frac{1}{2} + \ldots + \frac{1}{n}$	harmonic sum
alpha-harmony	$1 + \frac{1}{2^\alpha} + \ldots + \frac{1}{n^\alpha}$	generalised harmonic sum
sqrt harmony	$1 + \frac{1}{2^{1/2}} + \ldots + \frac{1}{n^{1/2}}$	$\alpha = 1/2$; divergent
square harmony	$1 + \frac{1}{2^2} + \ldots + \frac{1}{n^2}$	$\alpha = 2$; convergent: $\frac{\pi^2}{6} \approx 1.645$
Gaussian harmony	$2 \cdot \frac{n}{n+1} = 1 + \frac{1}{1+2} + \ldots + \frac{1}{1+\ldots+n}$	explains the BM25-TF $\frac{\mathrm{tf}_d}{\mathrm{tf}_d+\mathrm{pivdl}}$







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

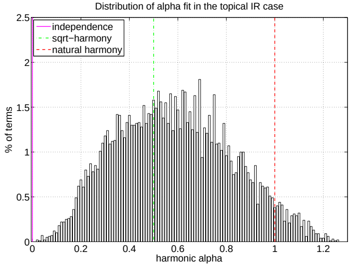
The area of each circle corresponds to the single event probability: $p = 0.5$.
The overlap becomes larger for growing α (harmony).

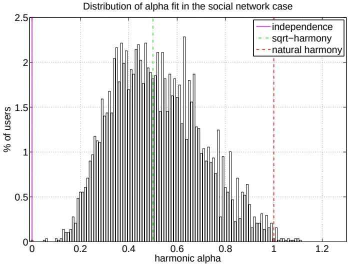
Africa in TREC-3

$$742,611 = 734,078 + 8,533$$

k	0	1	2	3	4	5	6	7	8
P_{obs}	0.9885	0.0062	0.0019	0.0011	0.0007	0.0005	0.0004	0.0002	0.0002
documents	734,078	4,584	1,462	809	550	345	271	182	137
P_{binomial}	0.9738	0.0258	0.0003	0	0	0	0	0	0
$P_{\text{alpha-harmonic},\alpha=0.41}$	0.9787	0.018	0.0023	0.0005	0.0002	0.0001	0	0	0

- 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

```
w_BM25(Term,Doc) :- tf_d(Term,Doc) BM25 & piv_dl(Doc);
# w_BM25: a probabilistic variant of the BM25-TF weight.

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

# 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: $2 \cdot k/(k+1)$
- BM25-TF: $2 \cdot \frac{\mathrm{tf}_d}{\mathrm{tf}_d+1} = 1 + \frac{1}{1+2} + \ldots + \frac{1}{1+2+\ldots+\mathrm{tf}_d}$

Harmony Assumptions: A link between
TF-IDF and Probability Theory

Other theories to model dependencies?

Questions?

[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 Algortihms in PDatalog
[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: ProbDatalog with HySpirit

---

red thread between IR Theory and abstraction for DB+IR

Church, K. and Gale, W. (1995a).
Inverse document frequency (idf): A measure of deviation from Poisson.
In *Proceedings of the Third Workshop on Very Large Corpora*, pages 121–130.

Church, K. and Gale, W. (1995b).
Poisson mixture.
Natural Language Engineering, 1(2):163–190.

Fagin, R. and Halpern, J. (1994).
Reasoning about knowledge and probability.
Journal of the ACM, 41(2):340–367.

Frommholz, I. and Roelleke, T. (2016).
Scalable DB+IR technology: Processing probabilistic datalog with hyspirit.
Datenbank-Spektrum, 16(1):39–48.

Fuhr, N. and Roelleke, T. (1997).
A probabilistic relational algebra for the integration of information retrieval and database systems.
ACM Transactions on Information Systems, 14(1):32–66.

Lewis, D. D. (1998).
Naive (Bayes) at forty: The independence assumption in information retrieval.
In *ECML '98: Proceedings of the 10th European Conference on Machine Learning*, pages 4–15, London, UK. Springer-Verlag.

Robertson, S. (2004).
Understanding inverse document frequency: On theoretical arguments for idf.
Journal of Documentation, 60:503–520.

Robertson, S. (2005).
On event spaces and probabilistic models in information retrieval.

Information Retrieval Journal, 8(2):319–329.

Roelleke, T. (2003).
A frequency-based and a Poisson-based probability of being informative.
In *ACM SIGIR*, pages 227–234, Toronto, Canada.

Roelleke, T. (2013).
Information Retrieval Models: Foundations and Relationships.
Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers.

Roelleke, T., Bonzanini, M., and Martinez-Alvarez, M. (2013).
On the Modelling of Ranking Algorithms in Probabilistic Datalog.
In *Proceedings of the 7th International Workshop on Ranking in Databases (DBRank)*. ACM.

Roelleke, T., Kaltenbrunner, A., and Baeza-Yates, R. A. (2015).
Harmony assumptions in information retrieval and social networks.
Comput. J., 58(11):2982–2999.

Roelleke, T. and Wang, J. (2006).
A parallel derivation of probabilistic information retrieval models.
In *ACM SIGIR*, pages 107–114, Seattle, USA.

Roelleke, T. and Wang, J. (2008).
TF-IDF uncovered: A study of theories and probabilities.
In *ACM SIGIR*, pages 435–442, Singapore.

Roelleke, T., Wu, H., Wang, J., and Azzam, H. (2008).
Modelling retrieval models in a probabilistic relational algebra with a new operator: The relational Bayes.
VLDB Journal, 17(1):5–37.