TF-IDF Uncovered: A Study of Theories and Probabilities (and Physics)

ACM SIGIR 2008, Singapore

Thomas Roelleke and Jun Wang
Queen Mary Univerity of London (QMUL)

- Introduction
  - Motivation & Background
  - Independence and Disjointness: Math
  - Independence and Disjointness: Weather in Glasgow
- Independent Terms
  - $P(q|d)$: LM: Linear mixture and event space mix
  - $P(d|q)$: "Extreme" mixture explains TF-IDF
- Disjoint Terms
  - Document-Query Independence (DQI)
  - Integral $\text{TF-IDF}(t) = \int \text{DQI}(t,x)\ \mathrm{d}x$; x is term probability
- Summary & Outlook

1. Uncover TF-IDF: Why?
2. TF-IDF: Math
3. Integral $\int \frac{1}{x} \, \mathrm{d}x = \log x$
4. TF-IDF and BIR
5. TF-IDF and LM
6. TF-IDF and Poisson
7. Other approaches

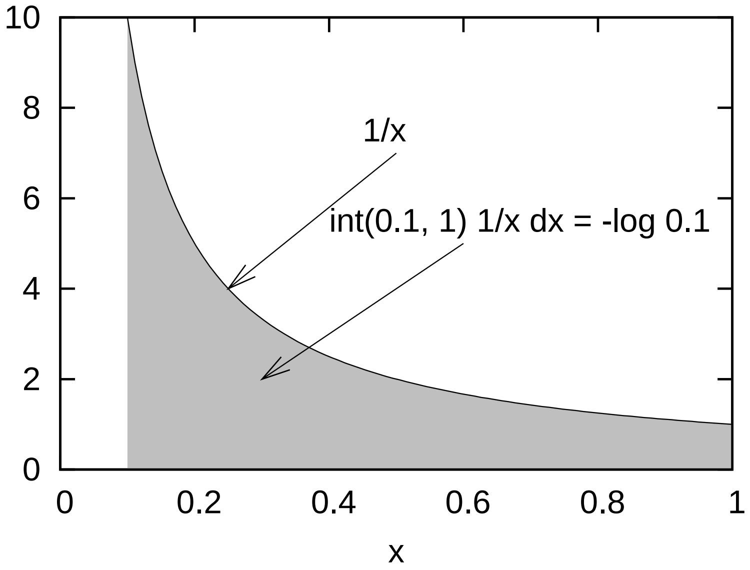
- TF-IDF is intuitive. "Probabilistic" interpretations "heavy"?
- LM has a probabilistic and "light" interpretation:
  1. Start: $P(q|d)$
  2. Assume independence: $P(q|d) = \prod_{t \in q} P(t|d)$
  3. Assume mixture: $P(t|d,c) = \delta \cdot P(t|d) + (1 - \delta) \cdot P(t|c)$
  4. Normalise
- Probabilistic and "light" interpretation of TF-IDF?
- Achieve a probabilistic relational framework for modelling *ALL* retrieval models ([Roelleke et al., 2008])
  - unifies IR models and
  - supports tuple rather than "just" document retrieval

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

$\mathrm{tf}(t,d)$	$\mathrm{tf}(t,q)$	$\mathrm{idf}(t,c)$
$\frac{n_L(t,d)}{n_L(t,d)+K}$	$n_L(t,q)$	$\log \frac{1}{P(t\|c)}$
$P(t\|d)$?	$P(t\|q)$?	$\frac{1}{P(t\|c)}$?
$P(d\|t)$?	$P(q\|t)$?	$P(t\|c)$?

Probabilistic interpretation of TF-IDF, tf(t,x), and idf(t,c)?
[Zaragoza et al., 2003], Bayesian extension of LM, integral over model parameters ...

$$\int \frac{1}{x} = \log x$$


10
8
6
4
2
0
0
0.2
0.4
0.6
0.8
1
x
1/x
int(0.1, 1) 1/x dx = -log 0.1

TF-IDF and BIR

[Robertson, 2004]: understanding IDF: on theoretical arguments

$$w_{\text{BIR-simplified}}(t, r, \bar{r}) := \frac{P_D(t|r)}{P_D(t|\bar{r})}$$

$$\log \frac{P_D(t|r)}{P_D(t|\bar{r})} = \log \frac{1}{P_D(t|c)} = \text{idf}(t, c)$$

[Croft and Harper, 1979]: constant $P(t|r)$

TF-IDF and LM

[Hiemstra, 2000]: probabilistic interpretation of TF-IDF

$$w_{\text{LM}}(t,d,c) := 1 + \frac{\delta}{1-\delta} \cdot \frac{P_L(t|d)}{P_D(t|c)}$$

Event space mix?
Should it be

$$\frac{P_L(t|d)}{P_L(t|c)}$$

TF-IDF and Poisson

[Roelleke and Wang, 2006]: parallel derivation, Poisson bridge

- Relationship between location-based and document-based probabilities $P_L(t|c)$ and $P_D(t|c)$
- 2-Poisson ([Robertson and Walker, 1994]) motivates $\mathrm{tf}_{\mathrm{BM25}} := \frac{n}{n+K}$

Other approaches

- Information-theoretic [Aizawa, 2003]
  $H(t) := \sum_t P(t) \cdot - \log P(t)$
- IDF is deviation from Poisson [Church and Gale, 1995]
- Probability of being informative [Roelleke, 2003]; Euler convergence $e^{-\lambda} = \lim_{N \to \infty} \left(1 - \frac{\lambda}{N}\right)^N$
- [Amati and van Rijsbergen, 2002]: risk times information gain: $\frac{1}{n+1} \cdot n \cdot \mathrm{idf}$

Independence: $$P(q|d) = \prod_{t \in q} P(t|d)^{n_L(t,q)}$$

Disjointness: $$P(q|d) = \sum_t P(q|t) \cdot P(t|d)$$

$P(q\|d)$	LM	?
$P(d\|q)$	?	TF-IDF?
	Independence	Disjointness

Retrieve the cities (documents) that imply the weather (query):

$$P(q|d) = P(\text{Weather}|\text{City})$$

A weather (query) instance: q = rainy, windy, rainy, sunny

Independent $P(\text{rainy, ...}|\text{glasgow}) = \prod_{t \in \{\text{rainy, ...}\}} P(t|\text{glasgow})^{n_L(t,q)}$

What if $P(\text{sunny}|\text{glasgow}) = 0$!?
$P(\text{sunny}|\text{glasgow}) = \delta \cdot P(\text{sunny}|\text{glasgow}) + (1-\delta) \cdot P(\text{sunny}|\text{uk})$

Disjoint $P(\text{rainy, ...}|\text{glasgow}) = \sum_t P(\text{rainy, ...}|t) \cdot P(t|\text{glasgow})$

1. $P(q|d)$: "Fix" of the event space mix in LM
2. $P(d|q)$: "Extreme" mixture explains TF-IDF
3. $O(r|d,q)$: ... in paper

$$P(q|d,c) = \prod_{t \in q} P(t|d,c)^{n_L(t,q)}$$

Linear mixture:

$$P(t|d,c) = \delta \cdot P_L(t|d) + (1-\delta) \cdot P_D(t|c)$$

Mix of Location-based and Document-based term probabilities!?

Result 1: "Fix" of the event space mix in LM.

$$P(d|q,c) = \prod_{t \in d} P(t|q,c)^{n_L(t,d)}$$

“Extreme” mixture:

$$P(t|q,c) = \begin{cases} 1 \cdot P(t|q) + 0 \cdot P(t|c), \text{ if } t \in q, \text{ then } \delta = 1 \\ 0 \cdot P(t|q) + 1 \cdot P(t|c), \text{ if } t \notin q, \text{ then } \delta = 0 \end{cases}$$

... after few steps ...

$$\sum_{t \in d \cap q} n_L(t,d) \cdot -\log P_D(t|c)$$

Result 2: “Extreme” mixture explains TF-IDF.

1. Decomposition of joint probability $P(d, q)$
2. Document-Query Independence (DQI)
3. TF-IDF is integral of DQI over term probability $P_D(t|c)$

$$P(d,q|c) = \sum_{t \in d \cap q} P(d|t) \cdot P(q|t) \cdot P(t|c)$$

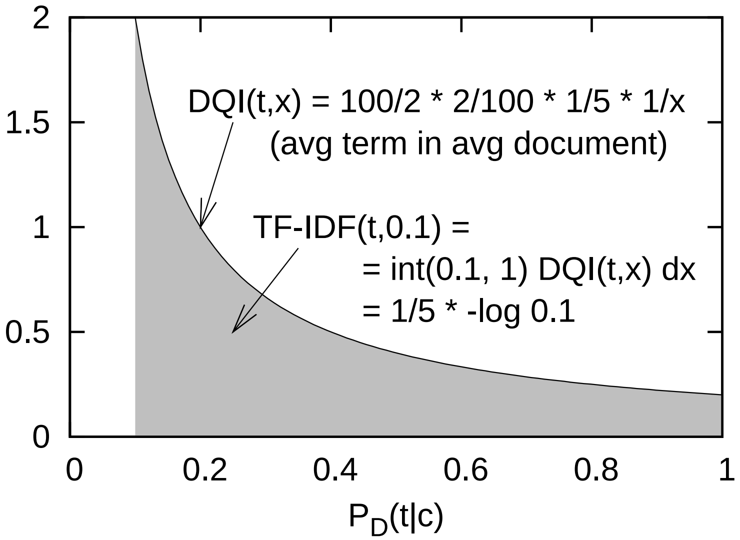
$$\frac{P(d,q|c)}{P(d|c) \cdot P(q|c)} = \sum_{t \in d \cap q} P(t|d) \cdot P(t|q) \cdot \frac{1}{P(t|c)}$$

Document-Query Independence (DQI)

$$\mathrm{DQI}(d,q|c) := \frac{P(d,q|c)}{P(d|c) \cdot P(q|c)} =$$

$$= \sum_t \frac{\mathrm{avgdl}(c)}{\mathrm{avgtf}(t,c)} \cdot P_L(t|d) \cdot P_L(t|q) \cdot \frac{1}{P_D(t|c)}$$

- > 1: the overlap of document and query is *greater* than if they were independent
- $= 1$: document and query are conditionally independent
- < 1: the overlap is *less* than if they were independent

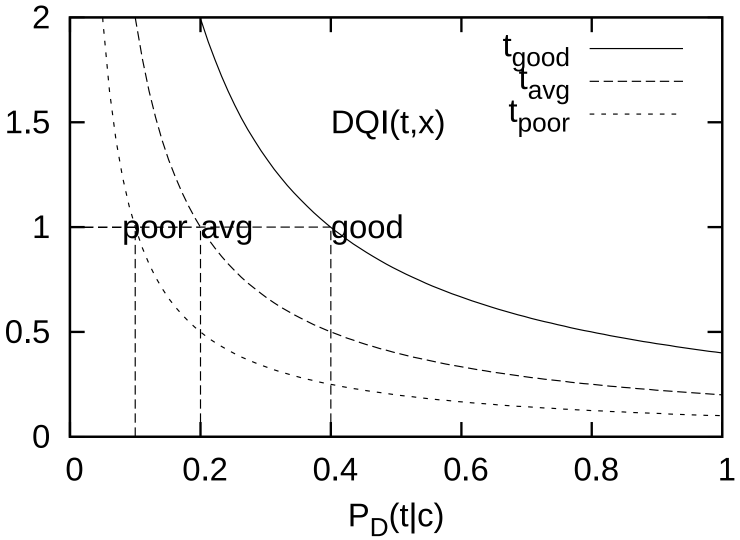

2
1.5
1
0.5
0
0
0.2
0.4
0.6
0.8
1
DQI(t,x) = 100/2 * 2/100 * 1/5 * 1/x
(avg term in avg document)
TF-IDF(t,0.1) =
= int(0.1, 1) DQI(t,x) dx
= 1/5 * -log 0.1
P_D(t|c)

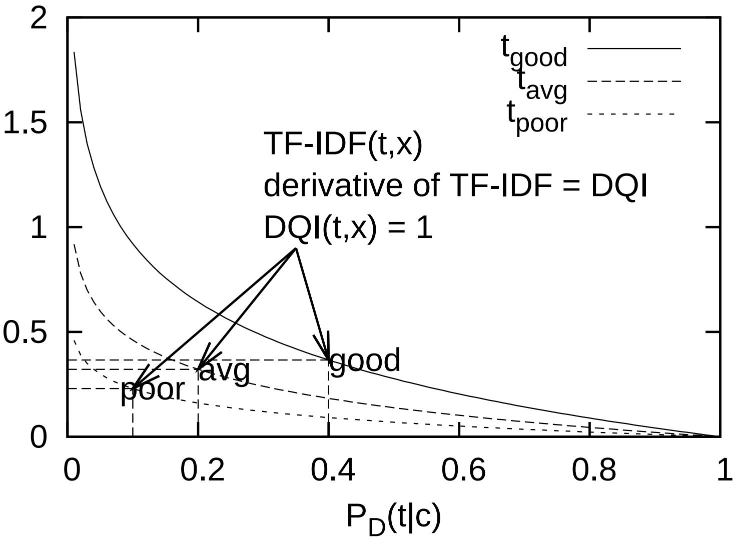
Start:

$$\int \frac{1}{x}\,\mathrm{d}x = \log x$$

Refinement: Definite integral: $\int_{x_0}^{1} \frac{1}{x}\,\mathrm{d}x = -\log x_0$

$$\begin{aligned}\int_{P_D(t|c)}^{1.0} \mathrm{DQI}(t,x)\,\mathrm{d}x &= \text{TF-IDF(t)}\\ \int_{P_D(t|c)}^{1.0} m \cdot P(t|d) \cdot P(t|q) \cdot \frac{1}{x}\,\mathrm{d}x &= m \cdot P(t|d) \cdot P(t|q) \cdot \mathrm{idf}(t,c)\end{aligned}$$


2
1.5
1
0.5
0
0
0.2
0.4
0.6
0.8
1
$P_D(t|c)$
DQI(t,x)
t_{good}
t_{avg}
t_{poor}
poor
avg
good


2
1.5
1
0.5
0
0
0.2
0.4
0.6
0.8
1
P_D(t|c)
t_good
t_avg
t_poor
TF-IDF(t,x)
derivative of TF-IDF = DQI
DQI(t,x) = 1
good
avg
poor

- Independent Terms
  1. $P(q|d)$: "Fix" for event space mix in LM
  2. $P(d|q)$: "Extreme" mixture explains TF-IDF
  3. $O(r|d,q)$: $r = q$
- Disjoint Terms
  1. Derivation of Document-Query Independence (DQI)
  2. TF-IDF is an integral of DQI over the collection-wide term probability $P(t|c)$

1. So? A contribution to explain and relate IR models.
2. DQI
    - independent terms?
    - entropy, dependence measures, ...?
3. $\text{DQI}(t) = 1$ for query term selection?
4. Is this study a basis for an analytical factor between TF-IDF and LM?

Thank you.

Aizawa, A. (2003).
An information-theoretic perspective of tf-idf measures.
Information Processing and Management, 39:45–65.

Amati, G. and van Rijsbergen, C. J. (2002).
Probabilistic models of information retrieval based on measuring the divergence from randomness.
ACM Transaction on Information Systems (TOIS), 20(4):357–389.

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

Croft, B. and Lafferty, J., editors (2003).
Language Modeling for Information Retrieval.
Kluwer.

Croft, W. and Harper, D. (1979).
Using probabilistic models of document retrieval without relevance information.
Journal of Documentation, 35:285–295.

Fang, H. and Zhai, C. (2006).
Semantic term matching in axiomatic approaches to information retrieval.
In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 115–122, New York, NY, USA. ACM.

Hiemstra, D. (2000).
A probabilistic justification for using tf.idf term weighting in information retrieval.
International Journal on Digital Libraries, 3(2):131–139.

Lafferty, J. and Zhai, C. (2003).

Probabilistic Relevance Models Based on Document and Query Generation, chapter 1.
In [Croft and Lafferty, 2003].

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

Robertson, S. E. and Walker, S. (1994).
Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval.
In Croft, W. B. and van Rijsbergen, C. J., editors, *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, London, et al. Springer-Verlag.

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

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

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.

Zaragoza, H., Hiemstra, D., and Tipping, M. (2003).
Bayesian extension to the language model for ad hoc information retrieval.
In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 4–9, New York, NY, USA. ACM Press.

Zobel, J. and Moffat, A. (1998).
Exploring the similarity space.
SIGIR Forum, 32(1):18–34.