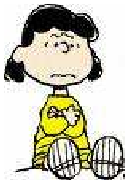
(me)

Hi! I'm Lillian Lee from Cornell University. Welcome to my poster!

IDF Revisited: A Simpler, Better Derivation



(you)

Ugh! ***Who needs yet another theoretical justification of the IDF?***

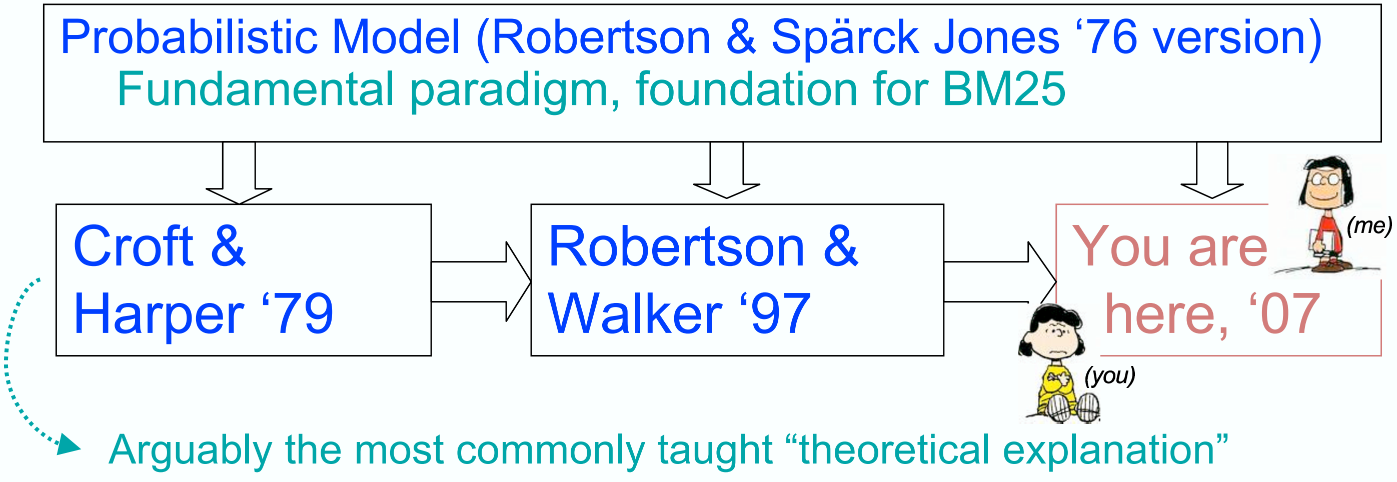
Reminder: The **inverse document frequency (IDF)**, a term-importance measure taking some variant of the form

$$\frac{N = \text{corpus size}}{n_i = \text{no. of docs containing the term } t_i}$$

is **used in (probably) all IR systems** (Harman ’05).

(me)

I do agree that there’s been much prior work on theoretically justifying IDF’s practical effectiveness...



Probabilistic Model (Robertson & Spärck Jones ‘76 version)
Fundamental paradigm, foundation for BM25

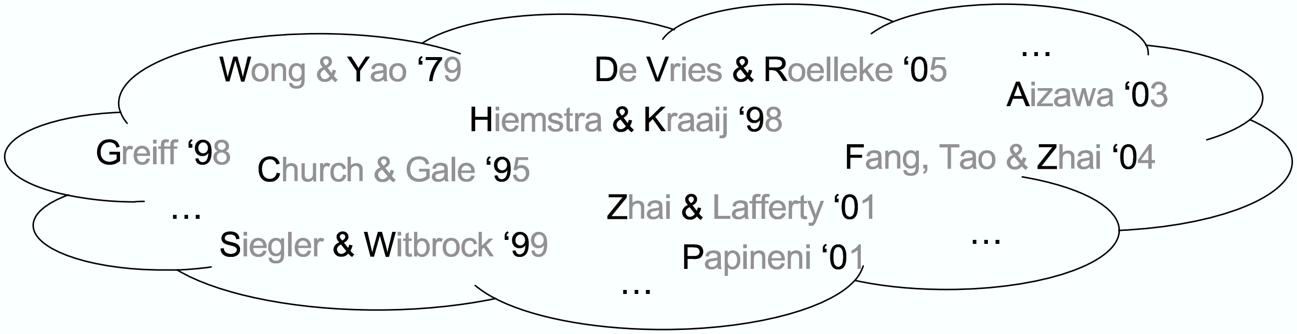
Croft & Harper ‘79

Robertson & Walker ‘97

You are here, ‘07

Arguably the most commonly taught “theoretical explanation”

Other foundations: VSM, LM, information theory, etc.



Wong & Yao ‘79

De Vries & Roelleke ‘05

…

Aizawa ‘03

Hiemstra & Kraaij ‘98

Greiff ‘98

Church & Gale ‘95

Fang, Tao & Zhai ‘04

Zhai & Lafferty ‘01

…

Siegler & Witbrock ‘99

…

Papineni ‘01

…

Robertson & Spärck Jones term weighting

The weight for query term t_i should be based in part on:

$$p_i \overset{\text{def}}{=} Pr(t_i \text{ occurs} \,|\, \textit{Relevant} = \text{“yes”})$$

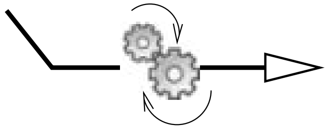
The full RSJ term-weight equation is omitted for clarity.

Challenge: estimating p_i without relevance info or feedback

(the “classic” ad hoc retrieval setting)

Croft & Harper (CH) assumption: all the query terms have the same occurrence probability within relevant docs:

$$\widehat{p_i} = k \quad \text{for some constant } k\,.$$



IDF

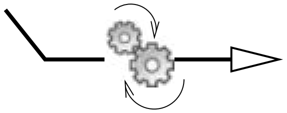
(algebra)

(you)

If the query is "Amsterdam NL", "NL" will appear in fewer relevant documents than "Amsterdam". Surely there's a more plausible assumption.

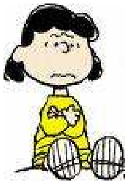
Robertson & Walker (RW) assumption: For some $k \in [0.5, 1]$,

$$\widehat{p_i} = \frac{k}{k + (1-k)\frac{N-n_i}{N}} \, .$$



IDF

(algebra)



(you)

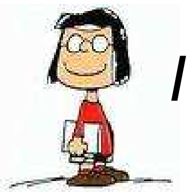
What's that supposed to mean?

(me)

I don't really know of an intuitive explanation for that equation. But ...

- RW's $\widehat{p_i}$ approximates linearity in n_i for $n_i \in [0, N]$, and thus fixes a technical problem with CH.
- Robertson & Walker assert that approximation is necessary: “the straight-line model is actually rather intractable, and does not lead to a simple weighting formula.”

OK, but surely there's a more intuitive assumption for us to use?



I'm glad you asked!

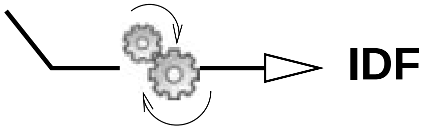
Intuition: A query term should be at least as likely to occur in a relevant doc. as it is to appear in any doc.

overall occurrence prob

$$\widehat{p_i} = \frac{n_i + L}{N + L}$$

"lift" for relevant docs

keeps estimate below 1

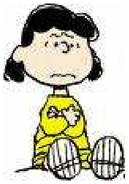


IDF

(algebra)

Our new estimate is:

- simple,
- intuitive, and
- linear in n_i: approximation turns out to be unnecessary


(you)

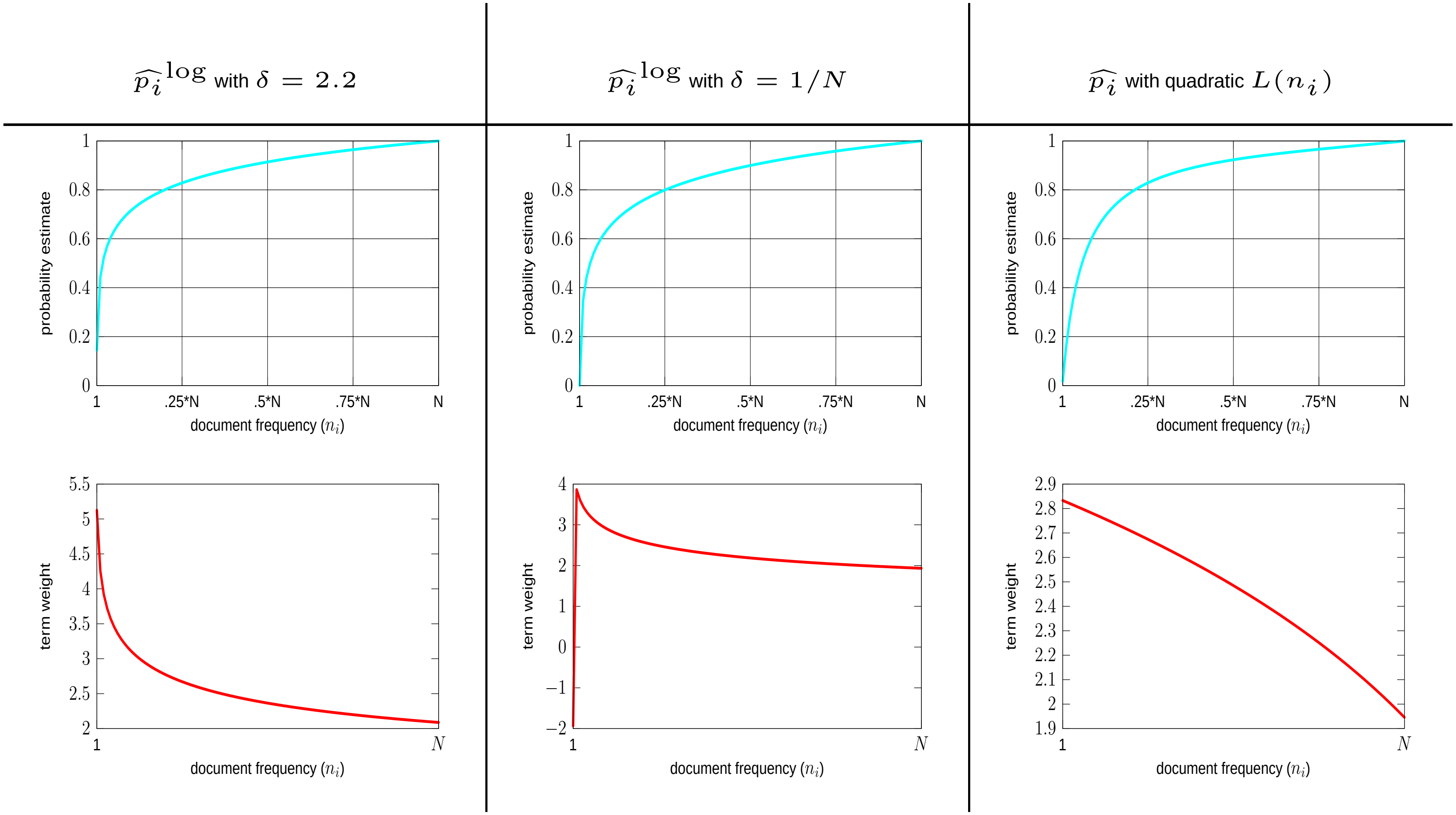
But even supposing I buy all that, is there any practical use to this work?

An extension of this idea might lead to new term-weighting components.

Idea: rewrite L as $L(n_i)$, a function of document frequency.

- Greiff's ('98) empirical study found p_i to be roughly logarithmic in n_i on some corpora.
- This behavior can be captured by our suggested extension via a *non-monotonic* $L(n_i)$.

Note: different "lift" functions can yield similar-looking p_is but very different term-weight components.



$\widehat{p_i}^{\log}$ with $\delta = 2.2$

$\widehat{p_i}^{\log}$ with $\delta = 1/N$

$\widehat{p_i}$ with quadratic $L(n_i)$

In summary, our new derivation:

(1) seems as simple yet more plausible than "RSJ + RW" or the commonly-taught "RSJ+ CH";

(2) solves Robertson & Walker's "intractable" problem; and

(3) could lead to new term-weighting schemes.

(you)

Thanks! I'll go see some other posters now ...

(me)

Sure! Thanks for stopping by!