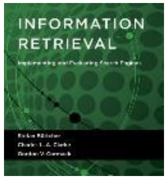
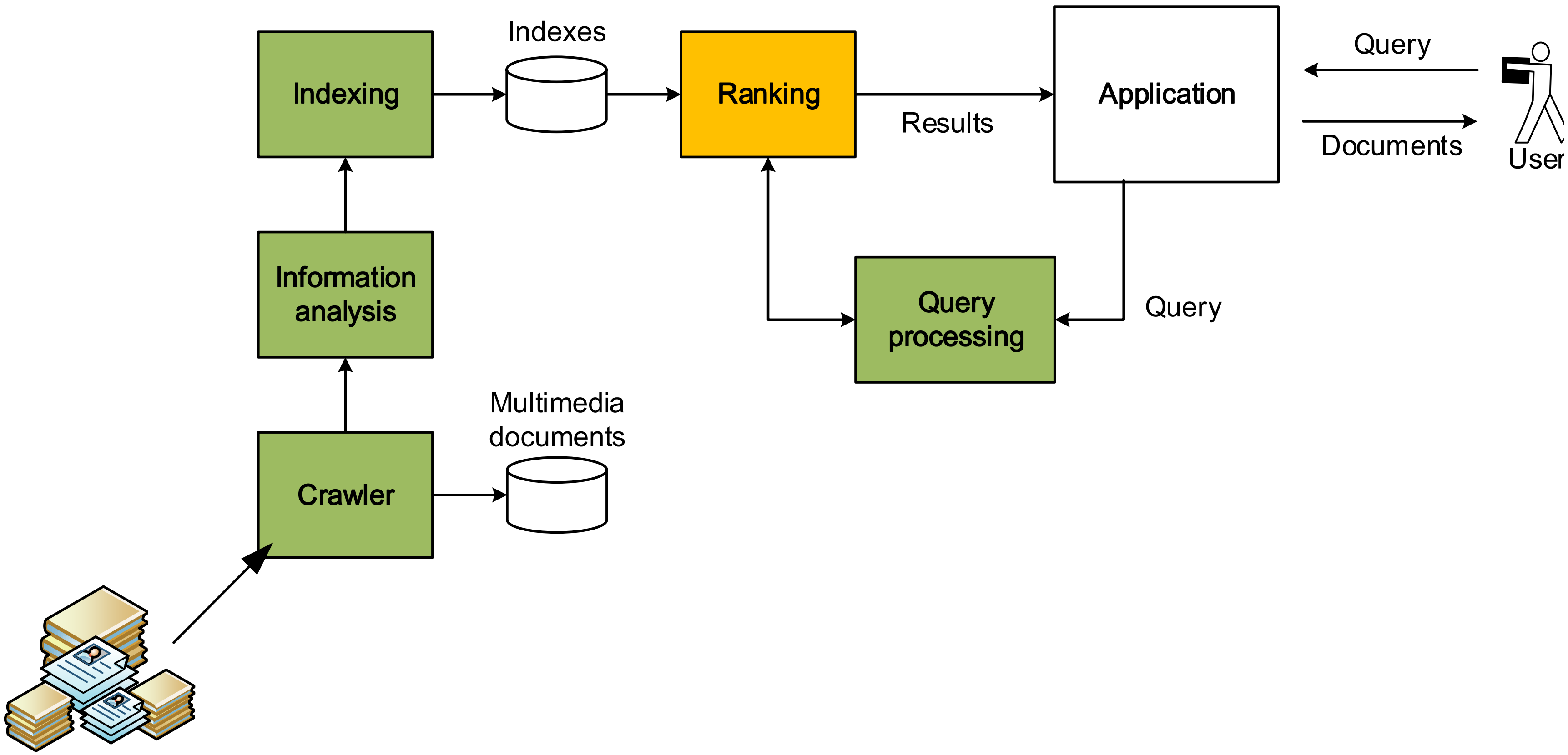
Probabilistic Retrieval Models

Relevance; Binary Independent Model; BM25

Web Search

Slides based on the books:  

Overview



Indexing
Indexes
Ranking
Application
Query
Results
Documents
User
Information analysis
Query processing
Query
Multimedia documents
Crawler

Retrieval models

- Geometric/linear spaces
  - Vector space model
- Probability ranking principle
- Language models approach to IR
  - An important emphasis in recent work
- Probabilistic retrieval model
  - Binary independence model
  - Okapi’s BM25

Part 1: Probabilistic Retrieval Models

Binary independence model

Probabilistic retrieval models

- PRP in action: Rank all documents by $p(r = 1|q, d)$
  - Theorem: Using the PRP is optimal, in that it minimizes the loss (Bayes risk) under 1/0 loss
  - Provable if all probabilities correct, etc. [e.g., Ripley 1996]

$$p(r|q,d) = \frac{p(d,q|r)p(r)}{p(d,q)}$$

- Using odds, we reach a more convenient ranking formulation:

$$O(R|q,d) = \frac{p(r=1|q,d)}{p(r=0|q,d)} = \frac{p(d|q,r=1)p(q|r=1)p(r=1)}{p(d|q,r=0)\underbrace{p(q|r=0)p(r=0)}_{\text{Constant part}}}$$

$$\boxed{O(R|q,d) \propto \frac{p(d|q,r=1)}{p(d|q,r=0)}}$$

Binary Independence Model

- Binary representation of words, i.e., documents are represented as binary incidence vectors of terms:

$$d = (d_0, d_1, \dots, d_n)$$

$d_i = 1$ iff term *i* is present in document *d*, and $d_i = 0$ otherwise.

- Queries: binary term incidence vectors

- Independence: terms occur in documents independently.
  - Different documents can be modeled as the same vector.

Binary Independence Model

- Will use odds and Bayes' Rule: $O(R|q,d) \propto \frac{p(d|q,r=1)}{p(d|q,r=0)}$

and the independence assumption:

$$O(R|q,d) \propto \frac{p(d|q,r=1)}{p(d|q,r=0)} = \frac{p(d_0,d_1,\dots,d_n|q,r=1)}{p(d_0,d_1,\dots,d_n|q,r=0)}$$

$$O(R|q,d) \propto \prod_{t=1}^{n} \frac{p(d_t|q_i,r=1)}{p(d_t|q_i,r=0)}$$

Binary Independence Model

$$O(R|q,d) \propto \frac{p(d|q,r=1)}{p(d|q,r=0)} = \prod_{t=1}^{n} \frac{p(d_t|q,r=1)}{p(d_t|q,r=0)}$$

- Since d_i is always 0 or 1:

$$O(R|q,d) \propto \prod_{t\in(q\cap d)} \frac{p(d_t=1|q_t,r=1)}{p(d_t=1|q_t,r=0)} \prod_{t\in(q\setminus d)} \frac{p(d_t=0|q_t,r=1)}{p(d_t=0|q_t,r=0)}$$

- Converting to log-odds and considering only the query terms:

$$O(R|q,d) \propto \sum_{t\in(q\cap d)} \log \frac{p(d_t=1|r=1)}{p(d_t=1|r=0)} \frac{p(d_t=0|r=0)}{p(d_t=0|r=1)}$$

Binary Independence Model

- In the end, all boils down to computing the Retrieval Status Value (log-odds):

$$RSV = \sum_{t \in (q \cap d)} w_t \,,$$

where $$w_t = \log \frac{p(d_t = 1|r = 1)}{p(d_t = 1|r = 0)} \frac{p(d_t = 0|r = 0)}{p(d_t = 0|r = 1)}$$

- Letting $p_t = p(d_t = 1|r = 1)$ and $\overline{p_t} = p(d_t = 1|r = 0)$, we get:

$$\boxed{w_t = \log \frac{p_t(1 - \overline{p_t})}{\overline{p_t}(1 - p_t)}}$$

Binary Independence Model

- Estimating w_t coefficients becomes the central problem in BIM.

Documents	Relevant	Non-relevant	Total
$d_t = 1$	$N_{t,r}$	$N_t - N_{t,r}$	N_t
$d_t = 0$	$N_r - N_{t,r}$	$(N - N_t) - \left(N_r - N_{t,r}\right)$	$N - N_t$
Total	N_r	$N - N_r$	N

$$p_t = p(d_t = 1|r = 1) = \frac{N_{t,r}}{N_r} \qquad \overline{p_t} = p(d_t = 1|r = 0) = \frac{N_t - N_{t,r}}{N - N_r}$$

$$w_t = \log \frac{p_t(1 - \overline{p_t})}{\overline{p_t}(1 - p_t)} = \log \frac{N_{t,r}\left(N - N_t - N_r + N_{t,r}\right)}{\left(N_r - N_{t,r}\right)\left(N_t - N_{t,r}\right)}$$

Estimation

$$w_t = \log \frac{N_{t,r}\left(N - N_t - N_r + N_{t,r}\right)}{\left(N_r - N_{t,r}\right)\left(N_t - N_{t,r}\right)} = \log \frac{N_{t,r}}{N_r - N_{t,r}} + \log \frac{N - N_t - N_r + N_{t,r}}{N_t - N_{t,r}}$$

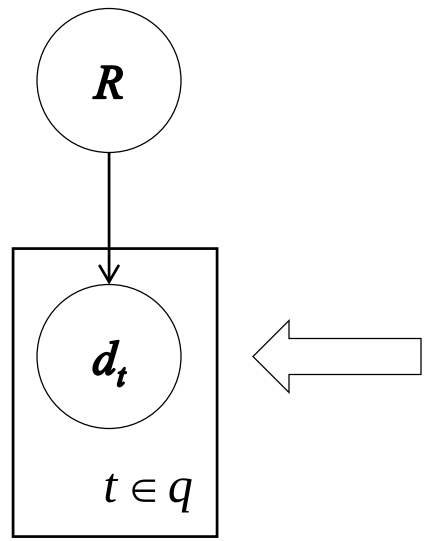
- On the second term, considering that N_r and $N_{t,r}$ are very small relative to N_t and N, we can approximate $N_{t,r} = N_r = 0$

$$w_t = \text{logit}(p_t) + \log \frac{N - N_t}{N_t}$$

- Moreover, if $2N_r \approx N_{t,r}$ and N_t is small relative to N, we get the IDF formula:

$$\boxed{w_t = \log \frac{N}{N_t} = IDF}$$

Graphical model for BIM



Binary variables

$$d_t = (tf_t \neq 0)$$

Experimental comparison

	TREC45				Gov2			
	1998		1999		2005		2006	
Method	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP
BIM	0.256	0.141	0.224	0.148	0.069	0.050	0.106	0.083
2-Poisson	0.402	0.177	0.406	0.207	0.418	0.171	0.538	0.207
BM25	0.424	0.178	0.440	0.205	0.471	0.243	0.534	0.277
LMD	0.450	0.193	0.428	0.226	0.484	0.244	0.580	0.293
BM25F					0.482	0.242	0.544	0.277
BM25+PRF	0.452	0.239	0.454	0.249	0.567	0.277	0.588	0.314
RRF	0.462	0.215	0.464	0.252	0.543	0.297	0.570	0.352
LR			0.446	0.266			0.588	0.309
RankSVM			0.420	0.234			0.556	0.268

Results under TREC45 have the same index. Results under Gov2 have the same index.
Results in different years have different queries.

Key limitations of the BIM

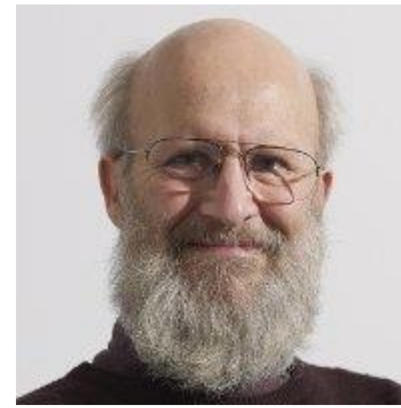
- BIM – like much of original IR – was designed for titles or abstracts, and not for modern full text search.
- We want to pay attention to **term frequency** and **document lengths**.
- Want some model of **how often terms occur in docs**.

Part 2: Probabilistic Retrieval Models

Okapi BM25

Okapi BM25

- BM25 “Best Match 25” (they had a bunch of tries!)
  - Developed in the context of the Okapi system.
  - Started to be increasingly adopted by other teams during the TREC competitions.
  - It works well!



- Goal: be sensitive to these quantities while not adding too many parameters
  - (Robertson and Walker 1994; Robertson and Zaragoza 2009; Spärck Jones et al. 2000)

Term frequency

- Probability Ranking Principle $O(R|q,d) \propto \frac{p(d|q,r=1)}{p(d|q,r=0)}$

- If we represent documents by its term presences (binary):

$$O(R|q,d) \propto \sum_{t \in q} \log \frac{p(d_t = 1|r)}{p(d_t = 1|\bar{r})} \frac{p(d_t = 0|\bar{r})}{p(d_t = 0|r)}$$

- If we represent documents by its term frequencies (pos-integer):

$$\boxed{O(R|q,d) \propto \sum_{t \in q} \log \frac{p(F_t = f_t|r)}{p(F_t = f_t|\bar{r})} \frac{p(F_t = 0|\bar{r})}{p(F_t = 0|r)}}$$

What are the best estimates of these probabilities?

Generative model for documents

- Words are drawn independently from the vocabulary using a multinomial distribution
- Distribution of term frequencies (*tf*) follows a Poisson distribution

... the **draft** is that each **team** is given a position in the **draft** ...

Poisson distribution

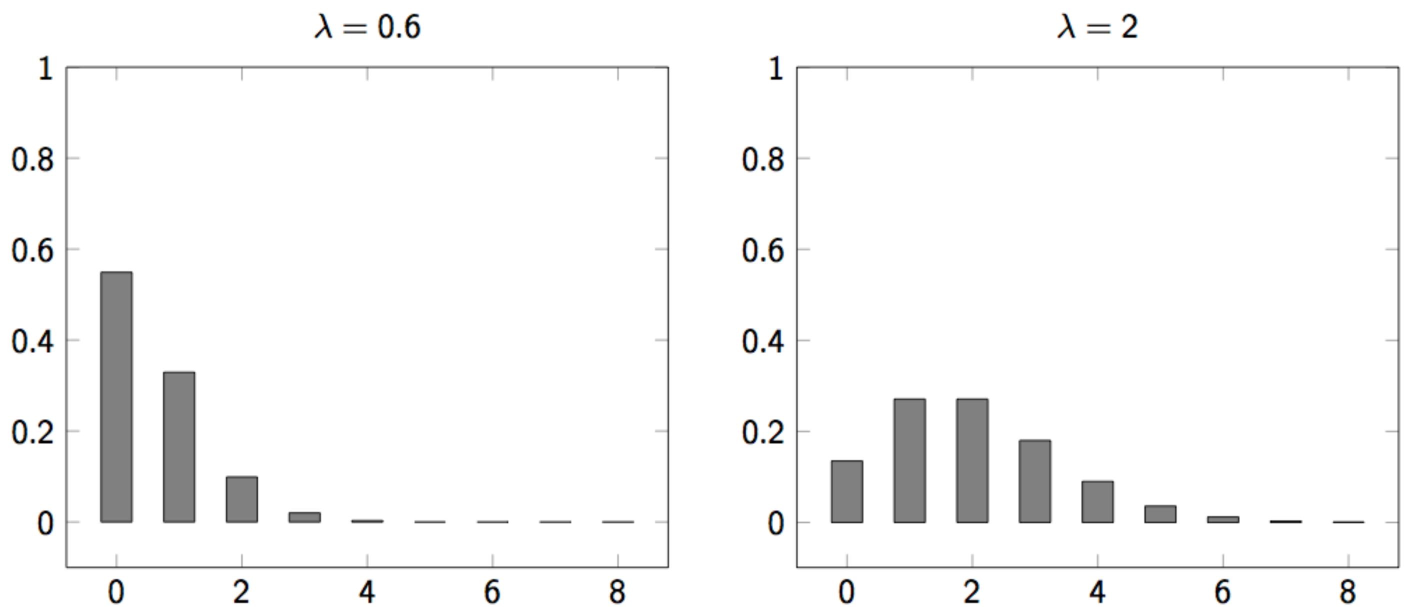
- The Poisson distribution models the probability of k, the number of events occurring in a fixed interval of time/space, with known average rate $\mu = (\mathrm{c}f\,/T)$, independent of the last event

$$g(k|\mu) = \frac{e^{-\mu} \cdot \mu^k}{k!}$$

- Examples
  - Number of cars arriving at the toll booth per minute
  - Number of typos on a page

Poisson model

- Assume that term frequencies in a document (tf_i) follow a Poisson distribution
  - "Fixed interval" implies fixed document length… assume roughly constant-sized document abstracts



(One) Poisson Model

- Is a reasonable fit for “general” words
- Is a poor fit for topic-specific words
  - get higher p(k) than predicted too often

Same frequency, different distribution.

		Documents containing *k* occurrences of word (λ = 53/650)												
Freq	**Word**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
53	**expected**	**599**	**49**	**2**										
52	*based*	600	48	2										
53	*conditions*	604	39	7										
55	*cathexis*	619	22	3	2	1	2	0	1					
51	*comic*	642	3	0	1	0	0	0	0	0	0	1	1	2

Harter, “A Probabilistic Approach to Automatic Keyword Indexing”, JASIST, 1975

The mismatch with the 1-Poisson model suggests fitting 2-Poisson distributions

Eliteness ("aboutness")

- Model term frequencies using *eliteness*
- What is eliteness?
  - Hidden variable for each document-term pair, denoted as E_i for term i
  - Represents *aboutness*: a term is elite in a document if, in some sense, the document is about the concept denoted by the term
  - Eliteness is binary
  - Term occurrences depend only on eliteness… but eliteness depends on relevance

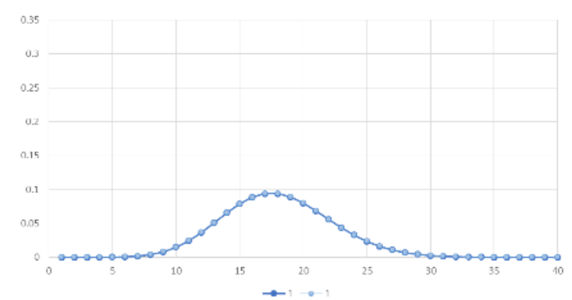
$$p(F_t = f_t | r)$$

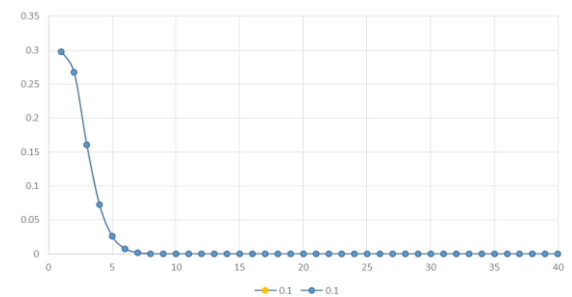
For an elite term, what is the probability of that term occurring # times on a relevant document?

Elite terms

Text from the Wikipedia page on the NFL draft showing **elite terms**

The **National Football League Draft** is an annual event in which the **National Football League** (**NFL**) **teams select eligible college football players**. It serves as the **league's** most common source of **player recruitment**. The basic design of the **draft** is that each **team** is given a **position** in the **draft order** in **reverse order** relative to its **record** …





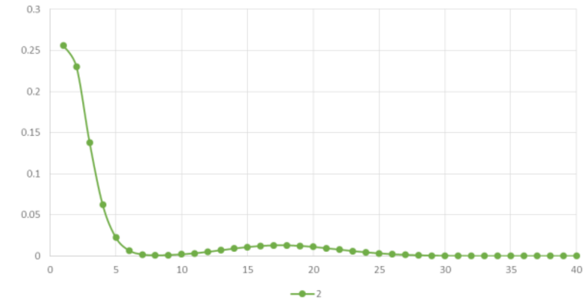
2-Poisson model

- In the "2-Poisson", the distribution is different depending on whether the term is elite or not

$$p(F_t = f_t|r) = p(e_t|r)g\left(f_t\middle|\mu_{e_t}\right) + p(\overline{e_t}|r)g\left(f_t\middle|\mu_{\overline{e_t}}\right)$$

$$p(F_t = f_t|r) = \pi\frac{e^{-\mu_{e_t}} \cdot {\mu_{e_t}}^{f_t}}{f_t!} + (1-\pi)\frac{e^{-\mu_{\overline{e_t}}} \cdot {\mu_{\overline{e_t}}}^{f_t}}{f_t!}$$

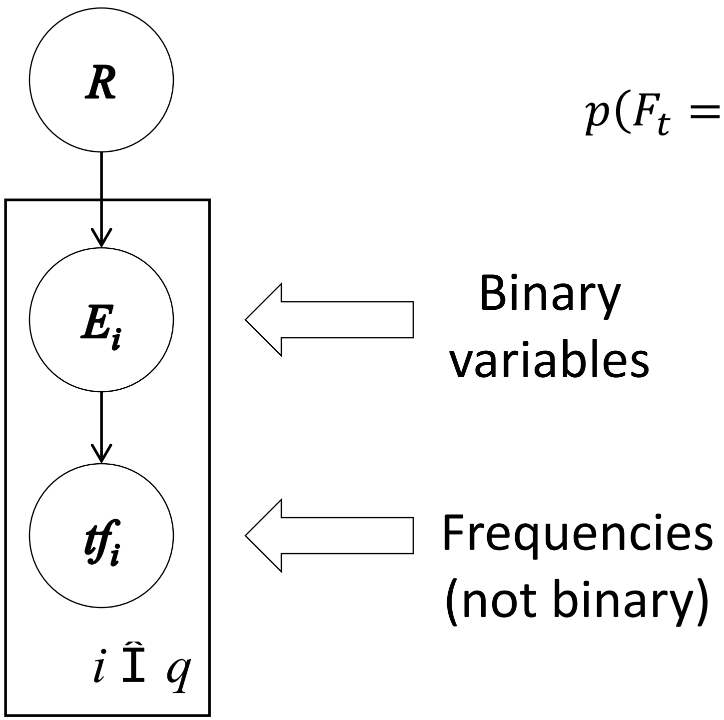
- where π is probability that document is elite for term but, unfortunately, we don't know $\mu_{e_t}, \mu_{\overline{e_t}}\ , \pi$



Graphical model with eliteness

$$p(F_t = f_t|r) = p(e_t|r)g\big(f_t\big|\mu_{e_t}\big) + \big(1 - p(e_t|r)\big)g\big(f_t\big|\mu_{\overline{e_t}}\big)$$

$$p(F_t = f_t|\bar{r}) = \pi\frac{e^{-\mu_{e_t}} \cdot {\mu_{e_t}}^{f_t}}{f_t!} + (1-\pi)\frac{e^{-\mu_{\overline{e_t}}} \cdot {\mu_{\overline{e_t}}}^{f_t}}{f_t!}$$



Retrieval Status Value

- Going back to the Probability Ranking Principle:

$$O(R|q,d) \propto \sum_{t \in q} \log \frac{p(F_t = f_t|r)}{p(F_t = f_t|\bar{r})} \frac{p(F_t = 0|\bar{r})}{p(F_t = 0|r)}$$

and considering the 2-Poisson model

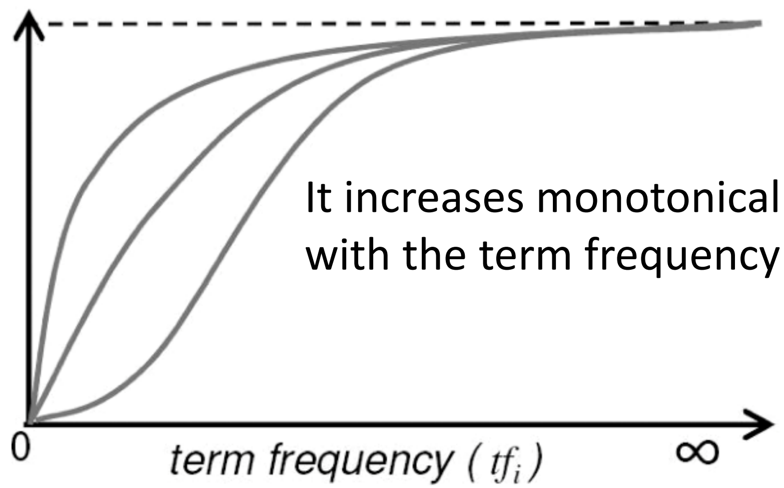
$$p(F_t = f_t|r) = p(e_t|r) g\left(f_t \middle| \mu_{e_t}\right) + \left(1 - p(e_t|r)\right) g\left(f_t \middle| \mu_{\overline{e_t}}\right)$$

$$p(F_t = f_t|\bar{r}) = p(e_t|\bar{r}) g\left(f_t \middle| \mu_{e_t}\right) + \left(1 - p(e_t|\bar{r})\right) g\left(f_t \middle| \mu_{\overline{e_t}}\right)$$

we realize that computing the parameters $\mu_{e_t}, \mu_{\overline{e_t}}, \pi$ for each term is too difficult.

Let’s get an insight: Graphing the RSV of several elite terms

Asymptotically approaches a maximum value as the term frequency increases



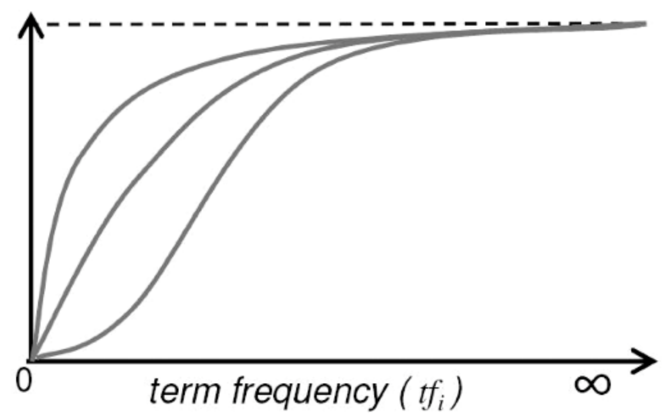
It increases monotonically with the term frequency

Its values are 0 when the term is absent.

Saturation property

- It can be demonstrated that when $tf \rightarrow \infty$ and $\mathrm{e}^{\mu_{\overline{e_t}} - \mu_{e_t}}$ is small, the RSV is approximated by:

$$\log \frac{p(e_t|r)\big(1 - p(e_t|\bar{r})\big)}{p(e_t|\bar{r})\big(1 - p(e_t|r)\big)}$$



- Note: the asymptotic saturation happens for the query terms on the document's high-frequency terms.

Approximating the saturation function

- Estimating parameters for the 2-Poisson model is not easy

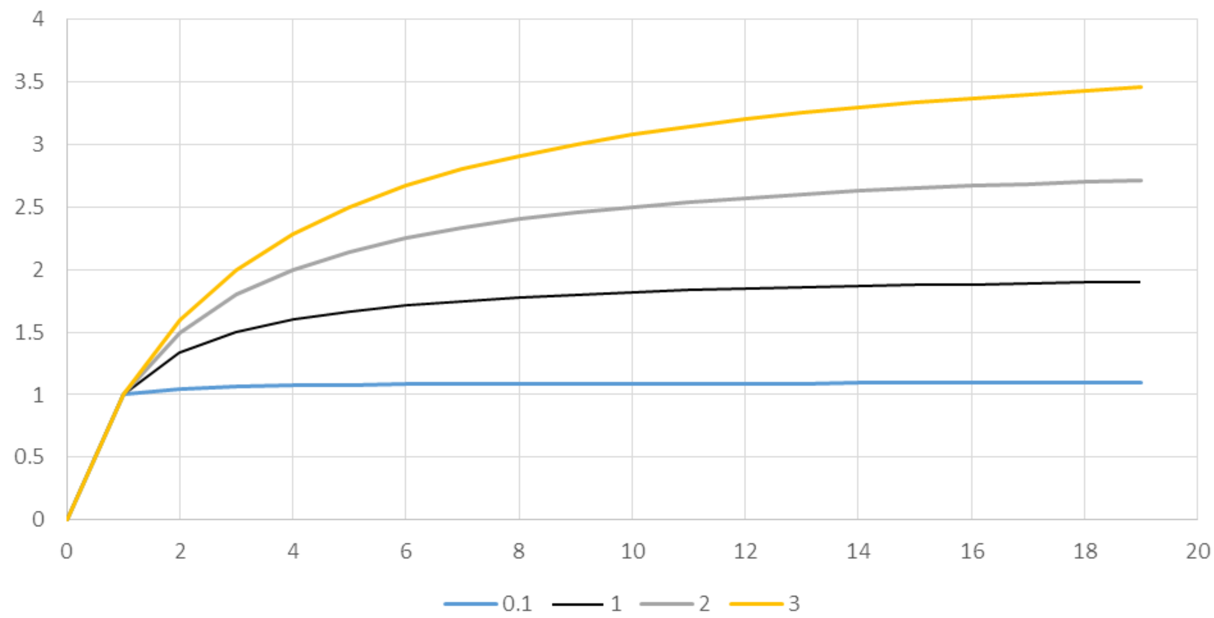
- <u>We are interested that the result averaged over all terms is correct, the individual curves are less important.</u>

- We can approximate the RSV with a simple parametric curve that has the same qualitative properties

$$\frac{(k_1 + 1) \cdot tf}{k_1 + tf}$$

Saturation function

$$\frac{(k_1 + 1) \cdot tf}{k_1 + tf}$$



- For high values of k_1, increments in tf_i continue to contribute significantly to the score
- Contributions tail off quickly for low values of k_1

Approximating the 2-Poisson: BM15

- Based on the previous observations, a simple approximation to the *RSV* with the two-Poisson model term weight is:

$$\sum q_t \cdot \frac{f_{t,d}(k_1 + 1)}{k_1 + f_{t,d}} \cdot w_t$$

where $w_t = IDF$ and $f_{t,d}$ is the frequency of term t.

Experimental comparison

	TREC45				Gov2			
	1998		1999		2005		2006	
Method	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP
Binary	0.256	0.141	0.224	0.148	0.069	0.050	0.106	0.083
2-Poisson	0.402	0.177	0.406	0.207	0.418	0.171	0.538	0.207
BM25	0.424	0.178	0.440	0.205	0.471	0.243	0.534	0.277
LMD	0.450	0.193	0.428	0.226	0.484	0.244	0.580	0.293
BM25F					0.482	0.242	0.544	0.277
BM25+PRF	0.452	0.239	0.454	0.249	0.567	0.277	0.588	0.314
RRF	0.462	0.215	0.464	0.252	0.543	0.297	0.570	0.352
LR			0.446	0.266			0.588	0.309
RankSVM			0.420	0.234			0.556	0.268

Results under TREC45 have the same index. Results under Gov2 have the same index.
Results in different years have different queries.

Document length normalization

- The Poisson Distribution assumed documents of same length.
- Why might documents be longer?
  - Verbosity: suggests observed tf_i too high
  - Larger scope: suggests observed tf_i may be right
- A real document collection probably has both effects.
- The term frequency should be normalized according to the document lengths

Normalizing by doc-length: BM11

- The term frequency can be represented as a normalized value with respect to the average document length versus the document length

$$f'_{t,d} = f_{t,d} \cdot \left(\frac{l_{avg}}{l_d}\right)$$

- Plugging into the BM15 formula, we get the BM11 retrieval model:

$$RSV = \sum q_t \cdot \frac{f'_{t,d}(k_1+1)}{k_1 + f'_{t,d}} \cdot w_t = \sum q_t \cdot \frac{f_{t,d} \cdot \left(\frac{l_{avg}}{l_d}\right)(k_1+1)}{k_1 + f_{t,d} \cdot \left(\frac{l_{avg}}{l_d}\right)} \cdot w_t$$

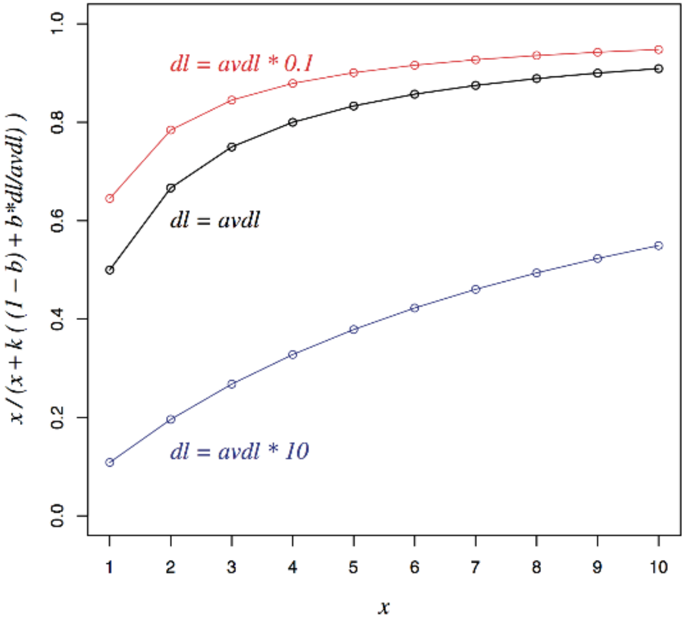
$$RSV = \sum q_t \cdot \frac{f_{t,d}(k_1+1)}{k_1 \cdot \left(\frac{l_d}{l_{avg}}\right) + f_{t,d}} \cdot w_t$$

Document length normalization

- Length normalization component

$$(1-b)+b\left(\frac{l_d}{l_{avg}}\right)$$

  - $b = 1$ full document length normalization
  - $b = 0$ no document length normalization
  - $avdl$: average document length over collection



Okapi BM25

$$RSV = \sum q_t \cdot \frac{f_{t,d}(k_1 + 1)}{k_1\left((1-b) + b\left(\frac{l_d}{l_{avg}}\right)\right) + f_{t,d}} \cdot IDF_t$$

- k_1 controls term frequency scaling **-> the saturation effect**
  - $k_1 = 0$ is binary model;
  - $k_1 = 1$ is raw term frequency.

- *b* controls document length normalization
  - *b = 0* is no length normalization;
  - *b = 1* is fully scaled by document length.

- Typically, $k_1 \in [1.2, 2.0]$ and $b{\sim}0.75$

Experimental comparison

	TREC45				Gov2			
	1998		1999		2005		2006	
Method	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP
Binary	0.256	0.141	0.224	0.148	0.069	0.050	0.106	0.083
2-Poisson	0.402	0.177	0.406	0.207	0.418	0.171	0.538	0.207
BM25	0.424	0.178	0.440	0.205	0.471	0.243	0.534	0.277
LMD	0.450	0.193	0.428	0.226	0.484	0.244	0.580	0.293
BM25F					0.482	0.242	0.544	0.277
BM25+PRF	0.452	0.239	0.454	0.249	0.567	0.277	0.588	0.314
RRF	0.462	0.215	0.464	0.252	0.543	0.297	0.570	0.352
LR			0.446	0.266			0.588	0.309
RankSVM			0.420	0.234			0.556	0.268

Results under TREC45 have the same index. Results under Gov2 have the same index.
Results in different years have different queries.

Summary and readings

- Probability Ranking Principle
- Binary Independence Model
- Modelling term frequency
  - 2-Poisson Model
  - 2-Poisson with document length normalization
- Sections 8.1 to 8.5 and 8.8 of: