Universidade Federal do Rio de Janeiro (UFRJ)
Departamento de Ciência da Computação (DCC)

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO - UFRJ -

Recuperação da Informação (MAB605)

Modelo Probabilístico

Profa. Giseli Rabello Lopes

Roteiro

- Introdução
- Modelo probabilístico
- BM25
- Referências

Introdução

- O modelo probabilístico propõe uma solução para o problema de RI utilizando um *framework* probabilístico
- Dada uma consulta do usuário, existe um **conjunto de resposta ideal** para essa consulta
- Dada uma descrição desse conjunto de resposta ideal, podemos recuperar os documentos relavantes
- Consulta é vista como uma especificação das **propriedades** desse conjunto de resposta ideal
  - Porém, quais são essas propriedades?

Introdução

- Um conjunto de documentos é recuperado de alguma forma
- O usuário inspeciona esses documentos procurando por aqueles que são relevantes (na verdade, apenas top 10-20 precisam ser inspecionados)
- O sistema de RI usa essa informação para refinar a descrição do conjunto de resposta ideal
- Pela repetição desse processo, é esperado que a descrição do conjunto de resposta ideal seja melhorado

Modelo probabilístico

- O modelo probabilístico
  - Tenta estimar a probabilidade de um documento ser relevante para uma dada consulta do usuário
  - Assume que essa probabilidade depende apenas das representações da consulta e do documento
  - O conjunto de resposta ideal, referido como R, deve maximizar a probabilidade de relevância
- Mas,
  - Como computar essas probabilidades?
  - Qual é o espaço amostral?

Modelo probabilístico

- Utiliza teoria das probabilidades como fundamentação para prover o raciocínio na presença de incerteza
  - Baseia-se na estimativa da probabilidade de um termo aparecer em um documento relevante para classificação
- Alguns modelos probabilísticos:
  - ***Binary Independence Model***
  - *Probability Ranking Principle*

Modelo probabilístico

- *Binary Independence Model* (BIM)
  - Proposto por Robertson e Sparck Jones em 1976
  - Modelo probabilístico original e ainda mais influente
    - *Binary*: documentos representados como vetores de termos com valores binários (incidência)
    - *Independence*: incidência dos termos nos documentos considerada de maneira independente
  - Utiliza teoria das probabilidades e regra de *Bayes*

O ranqueamento

- Seja,
  - R o conjunto de documentos relevantes para a consulta q
  - $\overline{R}$ o conjunto de documentos não relevantes para a consulta q
  - $P(R|\vec{d}_j)$ a probabilidade de que d_j seja relevante para a consulta q
  - $P(\overline{R}|\vec{d}_j)$ a probabilidade de que d_j não seja relevante para a consulta q
- A similaridade $sim(d_j, q)$ pode ser definida como

$$sim(d_j, q) = \frac{P(R|\vec{d}_j)}{P(\overline{R}|\vec{d}_j)}$$

Por *Odds* (chance do documento d_j ser relevante à consulta q)

Teorema de Bayes

- Permite calcular a seguinte probabilidade:

$$\boxed{P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}}$$

  - $P(A)$ e $P(B)$ são as probabilidades a priori de A e B
  - $P(B \mid A)$ e $P(A \mid B)$ são as probabilidades a posteriori de B condicional a A e de A condicional a B, respectivamente

- A ideia principal é que a probabilidade de um evento A *dado um evento* B depende não apenas do relacionamento entre os eventos A e B, mas também da probabilidade marginal (ou "probabilidade simples") da ocorrência de cada evento

Fonte: https://pt.wikipedia.org/wiki/Teorema_de_Bayes

Exemplo – Teorema de Bayes

- A probabilidade de alguém ter câncer de mama sabendo-se que a mamografia deu resultado positivo
  - Sabendo-se que:
    - A probabilidade condicional de pessoas com câncer que tiveram seu exame com resultado positivo é 95%
    - A probabilidade marginal deste tipo de câncer é 1%
    - A probabilidade marginal de mamografias resultado positivo para o câncer é 5%

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

$$P(canc.|mam.p.) = \frac{P(mam.p.|canc.)P(canc.)}{P(mam.p.)}$$

$$= 0{,}95 * 0{,}01 / 0{,}05 = 0{,}19 = 19\%$$

Fonte: https://pt.wikipedia.org/wiki/Teorema_de_Bayes

O ranqueamento

- Utilizando a regra de Bayes,

$$sim(d_j, q) = \frac{P(\vec{d_j}|R, q) \times P(R, q)}{P(\vec{d_j}|\overline{R}, q) \times P(\overline{R}, q)} \sim \frac{P(\vec{d_j}|R, q)}{P(\vec{d_j}|\overline{R}, q)}$$

- onde
  - $P(\vec{d_j}|R, q)$: probabilidade de aleatoriamente selecionar o documento d_j do conjunto R
  - $P(R, q)$: probabilidade de que um documento selecionado aleatoriamente a partir de toda coleção seja relevante para a consulta q
  - $P(\vec{d_j}|\overline{R}, q)$ e $P(\overline{R}, q)$: análogos e complementares

O ranqueamento

- Assumindo que os pesos $w_{i,j}$ são todos binários e independência entre os termos de indexação:

$$sim(d_j, q) \sim \frac{(\prod_{k_i|w_{i,j}=1} P(k_i|R,q)) \times (\prod_{k_i|w_{i,j}=0} P(\overline{k}_i|R,q))}{(\prod_{k_i|w_{i,j}=1} P(k_i|\overline{R},q)) \times (\prod_{k_i|w_{i,j}=0} P(\overline{k}_i|\overline{R},q))}$$

- onde
  - $P(k_i|R, q)$: probabilidade que o termo k_i esteja presente em um documento aleatoriamente selecionado a partir do conjunto R
  - $P(\overline{k_i}|R, q)$: probabilidade que o termo k_i não esteja presente em um documento aleatoriamente selecionado a partir do conjunto R
  - probabilidades com $\overline{R}$: análogas às já descritas

O ranqueamento

- Para simplificar a notação, são adotadas as seguintes convenções

$$p_{iR} = P(k_i|R,q)$$
$$q_{iR} = P(k_i|\overline{R},q)$$

- Como

$$P(k_i|R,q) + P(\overline{k}_i|R,q) = 1$$
$$P(k_i|\overline{R},q) + P(\overline{k}_i|\overline{R},q) = 1$$

- Escrevemos:

$$sim(d_j,q) \sim \frac{(\prod_{k_i|w_{i,j}=1} p_{iR}) \times (\prod_{k_i|w_{i,j}=0}(1-p_{iR}))}{(\prod_{k_i|w_{i,j}=1} q_{iR}) \times (\prod_{k_i|w_{i,j}=0}(1-q_{iR}))}$$

O ranqueamento

- Tomando os logaritmos, temos

$$
\begin{aligned}
sim(d_j, q) \quad \sim \quad & \log \prod_{k_i | w_{i,j}=1} p_{iR} + \log \prod_{k_i | w_{i,j}=0} (1 - p_{iR}) \\
& - \log \prod_{k_i | w_{i,j}=1} q_{iR} - \log \prod_{k_i | w_{i,j}=0} (1 - q_{iR})
\end{aligned}
$$

O raqueamento

- Adicionando termos que se cancelam mutuamente, obtemos

$$
\begin{aligned}
sim(d_j, q) \quad \sim \quad & \log \prod_{k_i|w_{i,j}=1} p_{iR} + \log \prod_{k_i|w_{i,j}=0} (1 - p_{ir}) \\
& \boxed{- \log \prod_{k_i|w_{i,j}=1} (1 - p_{ir}) + \log \prod_{k_i|w_{i,j}=1} (1 - p_{ir})} \\
& - \log \prod_{k_i|w_{i,j}=1} q_{iR} - \log \prod_{k_i|w_{i,j}=0} (1 - q_{iR}) \\
& \boxed{+ \log \prod_{k_i|w_{i,j}=1} (1 - q_{iR}) - \log \prod_{k_i|w_{i,j}=1} (1 - q_{iR})}
\end{aligned}
$$

O ranqueamento

- Usando operações sobre logaritmos, obtemos

$$
\begin{aligned}
sim(d_j, q) \;\sim\; & \log \prod_{k_i | w_{i,j}=1} \frac{p_{iR}}{(1 - p_{iR})} + \log \prod_{k_i} (1 - p_{iR}) \\
& + \log \prod_{k_i | w_{i,j}=1} \frac{(1 - q_{iR})}{q_{iR}} - \log \prod_{k_i} (1 - q_{iR})
\end{aligned}
$$

- Note que dois dos fatores na fórmula acima são uma função de todos os termos de indexação e não dependem do documento d_j. Eles são constantes para uma dada consulta e podem ser descartados para propósitos de ranqueamento

O ranqueamento

- Além disso, assumindo que

$$\forall\ k_i \notin q,\ \ p_{iR} = q_{iR}$$

- e convertendo os logaritmos de produtórios em somatórios de logaritmos, finalmente obtemos

$$\boxed{sim(d_j, q) \ \sim\ \sum_{k_i \in q \wedge k_i \in d_j} \log\left(\frac{p_{iR}}{1-p_{iR}}\right) + \log\left(\frac{1-q_{iR}}{q_{iR}}\right)}$$

- que é a expressão-chave para a computação do *ranking* no modelo probabilístico

Tabela de contingência das incidências de termos

- Seja,
  - N o número de documentos na coleção
  - n_i o número de documentos que contêm o termo k_i
  - R o número total de documentos relevantes para a consulta q
  - r_i o número de documentos relevantes que contêm o termo k_i
- Baseado nessas variáveis, podemos construir a seguinte tabela de contingência

	Relevantes	**Não relevantes**	**Total**
Documentos que contêm k_i	$\boxed{r_i}$	$n_i - r_i$	$\boxed{n_i}$
Documentos que não contêm k_i	$R - r_i$	$N - n_i - (R - r_i)$	$N - n_i$
Todos os documentos	$\boxed{R}$	$N - R$	$\boxed{N}$

Fórmula de ranqueamento

- Se a informação na tabela de contingência estivesse disponível para uma dada consulta, poderíamos escrever

$$p_{iR} = \frac{r_i}{R}$$

$$q_{iR} = \frac{n_i - r_i}{N - R}$$

- Então, a equação para computação do *ranking* no modelo probabilístico poderia ser reescrita como

$$sim(d_j, q) \sim \sum_{k_i[q,d_j]} \log \left(\frac{r_i}{R - r_i} \times \frac{N - n_i - R + r_i}{n_i - r_i} \right)$$

  - onde $k_i[q,d_j]$ é uma notação resumida para $k_i \in q \wedge k_i \in d_j$

Fórmula de ranqueamento

- Na fórmula prévia, também dependemos de estimar quais são os documentos relevantes para a consulta
- Para lidar com valores pequenos de r_i, adicionamos 0.5 a cada um dos termos da fórmula anterior, que muda $sim(d_j\ , q)$ para

$$\sum_{k_i[q,d_j]} \log \left(\frac{r_i + 0.5}{R - r_i + 0.5} \times \frac{N - n_i - R + r_i + 0.5}{n_i - r_i + 0.5} \right)$$

- Essa fórmula é considerada como a equação de *ranking* clássica para o modelo probabilístico e é conhecida como a equação de Robertson-Sparck Jones

Exemplo [Grossman & Frieder, 2004]

- Dados:
  - Q: “gold silver truck”
  - D_1: “Shipment of gold damaged in a fire”
  - D_2: “Delivery of silver arrived in a silver truck” (R)
  - D_3: “Shipment of gold arrived in a truck” (R)

	gold	silver	truck
N			
n_i			
R			
r_i			

Exemplo [Grossman & Frieder, 2004]

- Dados:
  - Q: “gold silver truck”
  - D_1: “Shipment of gold damaged in a fire”
  - D_2: “Delivery of silver arrived in a silver truck” (R)
  - D_3: “Shipment of gold arrived in a truck” (R)

	gold	silver	truck
N	3	3	3
n_i			
R			
r_i			

Exemplo [Grossman & Frieder, 2004]

- Dados:
  - Q: “gold silver truck”
  - D_1: “Shipment of gold damaged in a fire”
  - D_2: “Delivery of silver arrived in a silver truck” (R)
  - D_3: “Shipment of gold arrived in a truck” (R)

	gold	silver	truck
N	3	3	3
n_i	2		
R			
r_i			

Exemplo [Grossman & Frieder, 2004]

- Dados:
  - Q: “gold silver truck”
  - D_1: “Shipment of gold damaged in a fire”
  - D_2: “Delivery of silver arrived in a silver truck” (R)
  - D_3: “Shipment of gold arrived in a truck” (R)

	gold	silver	truck
N	3	3	3
n_i	2	1	
R			
r_i			

Exemplo [Grossman & Frieder, 2004]

- Dados:
  - Q: “gold silver truck”
  - D_1: “Shipment of gold damaged in a fire”
  - D_2: “Delivery of silver arrived in a silver truck” (R)
  - D_3: “Shipment of gold arrived in a truck” (R)

	gold	silver	truck
N	3	3	3
n_i	2	1	2
R			
r_i			

Exemplo [Grossman & Frieder, 2004]

- Dados:
  - Q: “gold silver truck”
  - D_1: “Shipment of gold damaged in a fire”
  - D_2: “Delivery of silver arrived in a silver truck” (R)
  - D_3: “Shipment of gold arrived in a truck” (R)

	gold	silver	truck
N	3	3	3
n_i	2	1	2
R	2	2	2
r_i			

Exemplo [Grossman & Frieder, 2004]

- Dados:
  - Q: “gold silver truck”
  - D_1: “Shipment of gold damaged in a fire”
  - D_2: “Delivery of silver arrived in a silver truck” (R)
  - D_3: “Shipment of gold arrived in a truck” (R)

	gold	silver	truck
N	3	3	3
n_i	2	1	2
R	2	2	2
r_i	1		

Exemplo [Grossman & Frieder, 2004]

- Dados:
  - Q: “gold silver truck”
  - D_1: “Shipment of gold damaged in a fire”
  - D_2: “Delivery of <u>silver</u> arrived in a <u>silver</u> truck” (R)
  - D_3: “Shipment of gold arrived in a truck” (R)

	gold	silver	truck
N	3	3	3
n_i	2	1	2
R	2	2	2
r_i	1	1	

Exemplo [Grossman & Frieder, 2004]

- Dados:
  - Q: "gold silver truck"
  - D_1: "Shipment of gold damaged in a fire"
  - D_2: "Delivery of silver arrived in a silver <u>truck</u>" (R)
  - D_3: "Shipment of gold arrived in a <u>truck</u>" (R)

	gold	silver	truck
N	3	3	3
n_i	2	1	2
R	2	2	2
r_i	1	1	2

Exemplo [Grossman & Frieder, 2004]

- Dados:
  - Q: “gold silver truck”
  - D_1: “Shipment of gold damaged in a fire”
  - D_2: “Delivery of silver arrived in a silver truck” (R)
  - D_3: “Shipment of gold arrived in a truck” (R)

	gold	silver	truck
N	3	3	3
n_i	2	1	2
R	2	2	2
r_i	1	1	2

Exemplo [Grossman & Frieder, 2004]

$$w_{k_i} = \log\left(\frac{r_i + 0.5}{R - r_i + 0.5} \times \frac{N - n_i - R + r_i + 0.5}{n_i - r_i + 0.5}\right)$$

	gold	silver	truck
N	3	3	3
n_i	2	1	2
R	2	2	2
r_i	1	1	2

Exemplo [Grossman & Frieder, 2004]

$$w_{k_i} = \log\left(\frac{r_i + 0.5}{R - r_i + 0.5} \times \frac{N - n_i - R + r_i + 0.5}{n_i - r_i + 0.5}\right)$$

	gold	silver	truck
N	3	3	3
n_i	2	1	2
R	2	2	2
r_i	1	1	2

$$w_{gold} = \log\left(\frac{1+0.5}{2-1+0.5} \times \frac{3-2-2+1+0.5}{2-1+0.5}\right) = \log(1 \times 0{,}333) = -0.477$$

$$w_{silver} = \log\left(\frac{1+0.5}{2-1+0.5} \times \frac{3-1-2+1+0.5}{1-1+0.5}\right) = \log(1 \times 3) = 0.477$$

$$w_{truck} = \log\left(\frac{2+0.5}{2-2+0.5} \times \frac{3-2-2+2+0.5}{2-2+0.5}\right) = \log(5 \times 3) = 1.176$$

*Valores apresentados usando logaritmo na base 10.

Exemplo [Grossman & Frieder, 2004]

- Dados:
  - Q: “gold silver truck”
  - D_1: “Shipment of gold damaged in a fire”
  - D_2: “Delivery of silver arrived in a silver truck” (R)
  - D_3: “Shipment of gold arrived in a truck” (R)

$$\begin{cases} w_{gold} = -0.477 \\ w_{silver} = 0.477 \\ w_{truck} = 1.176 \end{cases}$$

$$\boxed{sim(d_j, q) = \sum_{k_i [q, d_j]} w_{k_i}}$$

$$sim(D_1, Q) = w_{gold} = -0.477$$

$$sim(D_2, Q) = w_{silver} + w_{truck} = 1.653$$

$$sim(D_3, Q) = w_{gold} + w_{truck} = 0.699$$

Fórmula de ranqueamento

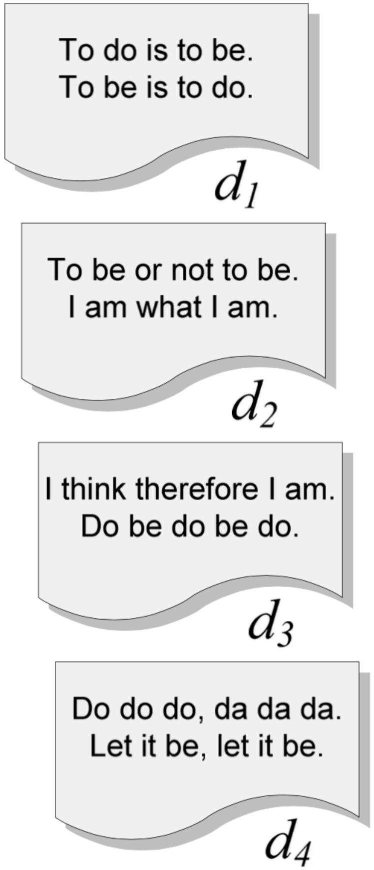
- A equação prévia não pode ser computada sem estimar r_i e R
- Uma possibilidade é assumir $R = r_i = 0$, como uma forma de inicialização à equação de ranqueamento, o que leva a

$$\boxed{sim(d_j, q) \sim \sum_{k_i[q,d_j]} \log \left(\frac{N - n_i + 0.5}{n_i + 0.5} \right)}$$

- Essa equação provê uma computação de *ranking* como *idf*
- Na ausência de informação de relevância, essa é a equação para ranqueamento do modelo probabilístico

Exemplo de ranqueamento

- Rank dos documentos computado pela equação de *ranking* probabilístico prévia (slide anterior) para a consulta "to do"



doc	rank computation	rank
d_1	$\log \frac{4-2+0.5}{2+0.5} + \log \frac{4-3+0.5}{3+0.5}$	- 1.222
d_2	$\log \frac{4-2+0.5}{2+0.5}$	0
d_3	$\log \frac{4-3+0.5}{3+0.5}$	- 1.222
d_4	$\log \frac{4-3+0.5}{3+0.5}$	- 1.222

*Valores apresentados usando logaritmo na base 2.

Exemplo de ranqueamento

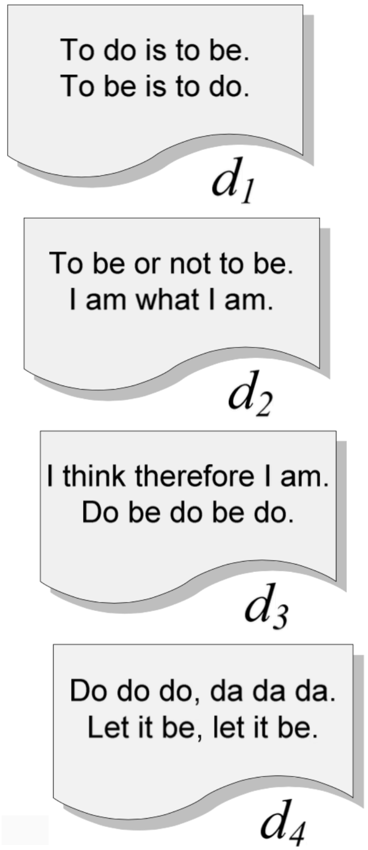
- A computação do *ranking* levou a pesos negativos devido ao termo "do"
- Na verdade, a equação do *ranking* probabilístico produz termos negativos sempre que $n_i > N/2$
- Um artifício possível para conter o efeito dos pesos negativos é mudar a equação anterior para:

$$sim(d_j, q) \sim \sum_{k_i[q,d_j]} \log \left(\frac{N + 0.5}{n_i + 0.5} \right)$$

- Fazendo isso, um termo que ocorre em todos os documentos ($n_i = N$) produz um peso igual a zero

Exemplo de ranqueamento

- Usando essa última formulação, reconstruímos a computação do *ranking* para a consulta “to do” e obtemos



doc	rank computation	rank
d_1	$\log \frac{4+0.5}{2+0.5} + \log \frac{4+0.5}{3+0.5}$	1.210
d_2	$\log \frac{4+0.5}{2+0.5}$	0.847
d_3	$\log \frac{4+0.5}{3+0.5}$	0.362
d_4	$\log \frac{4+0.5}{3+0.5}$	0.362

*Valores apresentados usando logaritmo na base 2.

Estimando r_i e R

- Nossos exemplos anteriores consideraram $r_i = R = 0$
- Uma alternativa é estimar r_i e R efetuando uma busca inicial (utilizando a equação anterior):
  - Selecionar os top 10-20 documentos ranqueados
  - Inspecioná-los para reunir novas estimativas para r_i e R
  - Remover os 10-20 documentos usados da coleção
  - Reprocessar a consulta com as estimativas obtidas para r_i e R
- Infelizmente, procedimentos como estes requerem intervenção humana para inicialmente selecionar os documentos relevantes

Melhorando o ranqueamento inicial

- Considerando a equação

$$sim(d_j, q) \sim \sum_{k_i \in q \wedge k_i \in d_j} \log\left(\frac{p_{iR}}{1 - p_{iR}}\right) + \log\left(\frac{1 - q_{iR}}{q_{iR}}\right)$$

- Como obter as probabilidades p_{iR} e q_{iR} ?
- Estimadas com base nas assunções:
  - $p_{iR} = 0.5$
  - $q_{iR} = (n_i / N)$ onde n_i é o número de documentos que contêm k_i
  - Usa esta estimativa inicial para recuperar um *ranking* inicial
  - Aperfeiçoa este *ranking* inicial

Melhorando o ranqueamento inicial

- Substituindo p_{iR} and q_{iR} na equação anterior, obtemos:

$$sim(d_j, q) \sim \sum_{k_i \in q \wedge k_i \in d_j} \log\left(\frac{N - n_i}{n_i}\right)$$

- que é a mesma equação usada quando nenhuma informação de relevância é provida, sem o fator de correção de 0.5
- Dada a estimativa inicial, podemos prover um *ranking* probabilístico inicial
- Depois disso, podemos tentar melhorar o *ranking* inicial, como segue

Melhorando o ranqueamento inicial

- Podemos tentar melhorar o *ranking* inicial, como segue
- Seja
  - D: conjunto de documentos recuperados inicialmente
  - D_i: subconjunto de documentos recuperados que contêm k_i
- Reavaliar as estimativas:
  - $p_{iR} = D_i \,/\, D$
  - $q_{iR} = (n_i - D_i) \,/\, (N - D)$
- Esse processo pode ser repetido recursivamente

Melhorando o ranqueamento inicial

$$sim(d_j, q) \sim \sum_{k_i \in q \wedge k_i \in d_j} \log\left(\frac{N - n_i}{n_i}\right)$$

- Para evitar problemas com $D = 1$ e $D_i = 0$:

$$p_{iR} = \frac{D_i + 0.5}{D + 1}; \quad q_{iR} = \frac{n_i - D_i + 0.5}{N - D + 1}$$

- Outra alternativa,

$$p_{iR} = \frac{D_i + \frac{n_i}{N}}{D + 1}; \quad q_{iR} = \frac{n_i - D_i + \frac{n_i}{N}}{N - D + 1}$$

Vantagens e desvantagens

- Vantagens:
  - Documentos ranqueados em ordem decrescente de acordo com sua probabilidade de relevância
    (na prática, contudo, isso não funciona tão bem, porque a relevância de um documento é afetada por variáveis externas ao sistema)
- Desvantagens:
  - Necessidade de estimativa inicial para p_{iR}
  - Método não leva em consideração os fatores tf
  - A falta de normalização pelo tamanho dos documentos
  - Adoção de independência dos termos

Comparação entre os modelos clássicos

- Modelo booleano não provê casamento parcial e é considerado o modelo clássico mais “fraco”
- Existem controvérsias sobre o modelo probabilístico superar o modelo vetorial
  - Croft & Harper [1979] sugerem que o modelo probabilístico provê uma melhor performance na recuperação
  - Entretanto, Salton & Buckley [1998] mostraram que o modelo vetorial supera-o em coleções gerais
  - Este também parece ser o pensamento dominante entre os pesquisadores e profissionais de RI

BM25 – *Best Match* 25

- BM25 foi criado como resultado de uma série de experimentos em variações do modelo probabilístico
- Uma boa ponderação de termos é baseada em três princípios
  - Frequência inversa de documento (IDF)
  - Frequência do termo (TF)
  - Normalização pelo tamanho do documento
- O modelo probabilístico cobre apenas o primeiro desses princípios
- Este raciocínio levou a uma série de experimentos com o sistema Okapi, o que levou à fórmula de ranqueamento BM25

BM25 – Fórmula de ranqueamento

- A motivação foi combinar os fatores de frequência de termos BM11 e BM15 (outras variações) como segue

$$\boxed{\mathcal{B}_{i,j} = \frac{(K_1 + 1) f_{i,j}}{K_1 \left[(1 - b) + b \frac{len(d_j)}{avg\_doclen} \right] + f_{i,j}}}$$

- Onde:
  - b é uma constante como valores no intervalo [0, 1]
    - Se $b = 0$, isso reduz o fator ao fator de frequência de termo BM15
    - Se $b = 1$, isso reduz ao fator de frequência de termo BM11
    - Para valor de b entre 0 e 1, a equação provê uma combinação de BM11 com BM15
  - $len(d_j)$ é o tamanho do documento (ex. contando o número de termos do doc.); avg_doclen é o tamanho médio dos documentos da coleção

BM25 – Fórmula de ranqueamento

- A equação de *ranking* para o modelo BM25 pode então ser escrita como

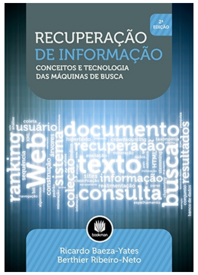
$$sim_{BM25}(d_j, q) \sim \sum_{k_i[q,d_j]} \mathcal{B}_{i,j} \times \log \left(\frac{N - n_i + 0.5}{n_i + 0.5} \right)$$

- Onde K_1 e b são constantes empíricas
  - $K_1 = 1$ funciona bem com coleções reais
  - b deve ser mantida próximo de 1 para enfatizar o efeito da normalização pelo tamanho do documento na fórmula BM11
    - Por ex., $b = 0.75$ é uma suposição razoável
  - Os valores das constantes podem ser ajustados para coleções particulares através da experimentação adequada

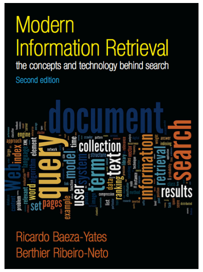
BM25

- Ao contrário do modelo probabilístico, a fórmula BM25 pode ser computada sem informação de relevância
- Há um consenso de que BM25 supera o modelo vetorial clássico para coleções gerais
- Assim, tem sido utilizado como base para a avaliação de novas funções de *ranking*, em substituição ao modelo vetorial clássico

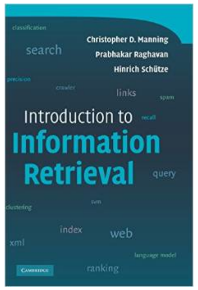
Referências



- Baeza-Yates, R.; Ribeiro-Neto, B. Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca. 2 ed. Bookman, 2013.



- Baeza-Yates, R.; Ribeiro-Neto, B. Modern Information Retrieval. Wokingham, UK: Addison-Wesley, 2 ed., 2011.



- Manning, C. D.; Raghavan, P.; Schütze, H. Introduction to Information Retrieval. Cambridge University Press, 2008.

  Online edition 2009: http://nlp.stanford.edu/IR-book/

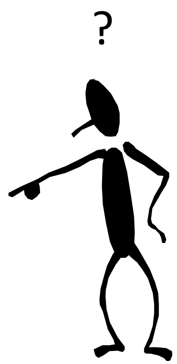
Referências

- Croft, W.; Harper, D. Using probabilistic models of retrieval without relevance information. Journal of Documentation, 35(4):285-295, 1979.
- Grossman, D. A.; Frieder, O. Information retrieval: algorithms and heuristics. 2nd ed. Dordrecht: Springer, c2004. 332p.
- Salton, G.; Buckley, C. Term-weighting approaches in automatic retrieval. Information Processing & Management, 24(5):513-523, 1988.

Universidade Federal do Rio de Janeiro (UFRJ)
Departamento de Ciência da Computação (DCC)

Recuperação da Informação (MAB605)

Dúvidas?



Profa. Giseli Rabello Lopes
giseli@dcc.ufrj.br
CCMN - DCC - Sala E-2012

Material adaptado e traduzido de slides do capítulo 2 do livro [Baeza-Yates & Ribeiro-Neto, 2013]