A03 Modelagem em RI

Prof. Dr. George H. G. Fonseca

CDD004 Técnicas de Recuperação da Informação
Pós Graduação em Ciência dos Dados
Universidade Federal de Ouro Preto

Janeiro de 2022

1. Introdução

Introdução

- Modelagem em RI é um processo complexo cujo objetivo é produzir uma **função de ranqueamento**, ou seja, uma função que atribui escores a documentos com relação a uma consulta.

- Esse processo pode ser dividido em duas tarefas principais:
  i. a criação de um arcabouço para representar documentos e consultas;
  ii. a definição de uma função de ranqueamento que calcula o grau de similaridade de cada documento com relação a uma consulta dada.

2. Modelagem e Ranqueamento

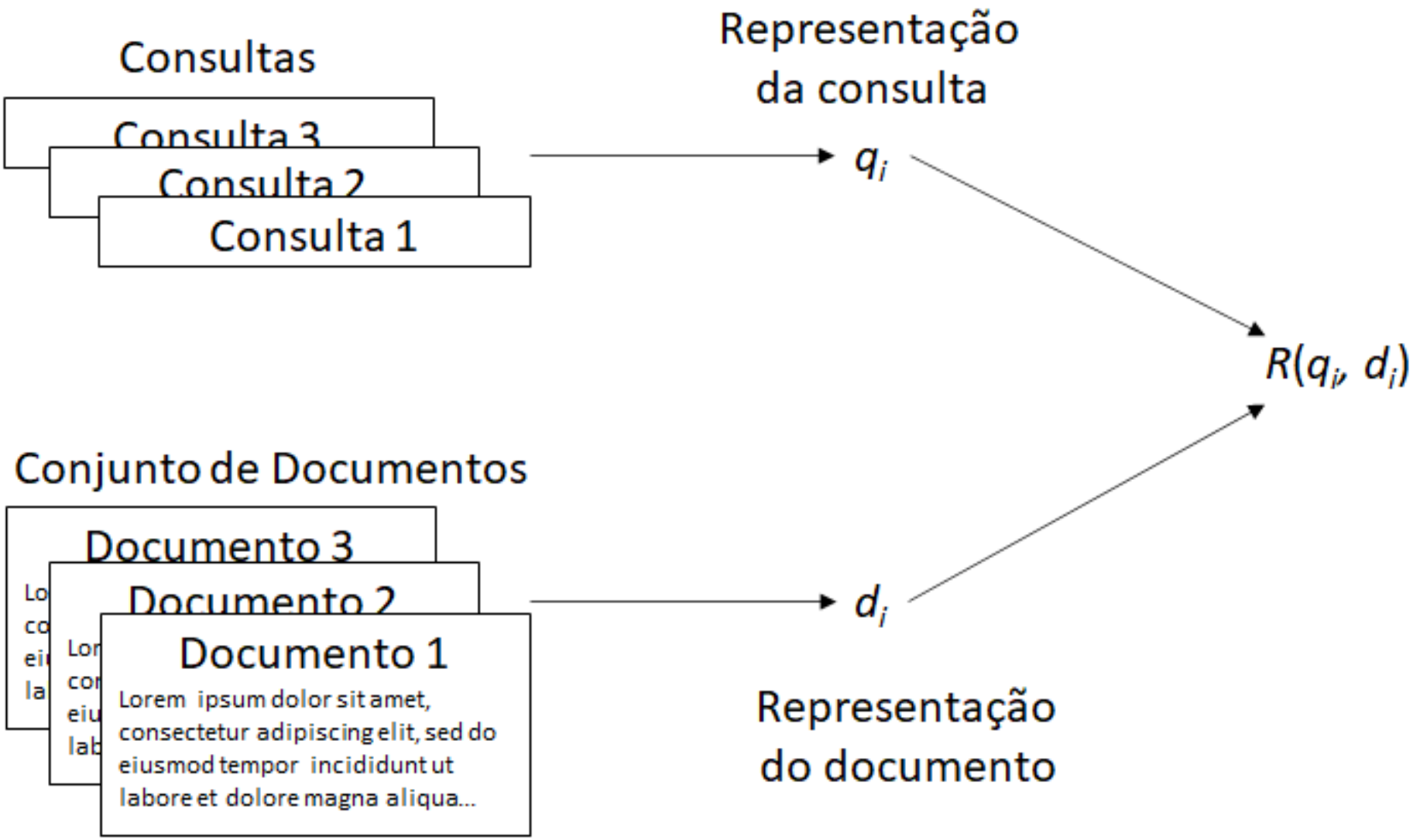
Modelagem e Ranqueamento

- Um desafio a qualquer sistema de RI é prever quais documentos os usuários irão considerar relevantes como resultado a uma consulta.
- Além disso há um grau de incerteza ou imprecisão devido ao fato de que dois usuários podem discordar sobre o que é e o que não é relevante.
- Para lidar com esse problema, sistemas de RI implementam um algoritmo preditivo que almeja aproximar-se da opinião de uma grande fração dos usuários com relação à relevância dos resultados de um consulta.
- Esse algoritmo de predição é essencialmente a **função de ranqueamento** utilizada para estabelecer uma ordenação dos documentos recuperados.

Caracterização de um Modelo de RI

- Um **modelo** de recuperação de informação é uma quádrupla $[\boldsymbol{D}, \boldsymbol{Q}, \mathcal{F}, R(q_i, d_j)]$ onde:
  1. $\boldsymbol{D}$ é um conjunto composto por visões lógicas (ou representações) dos **documentos** da coleção.
  2. $\boldsymbol{Q}$ é um conjunto composto por visões lógicas (ou representações) das necessidades de informação dos usuários. Essas representações são chamadas **consultas**.
  3. $\mathcal{F}$ é um arcabouço para modelar as representações dos documentos, das consultas e de seus relacionamentos, como conjuntos e relações Booleanas, vetores e operações da álgebra linear ou espaços e distribuições de probabilidade.
  4. $R(q_i, d_j)$ é uma função de ranqueamento que associa um número real à representação de uma consulta $q_i \in \boldsymbol{Q}$ e à representação de um documento $d_j \in \boldsymbol{D}$. Esse ranking define um ordenamento entre os documentos em relação à consulta q_i.

Caracterização de um Modelo de RI



3. Modelo Booleano

Modelo Booleano

- Considere $V = \{k_1, k_2, ..., k_t\}$ como o vocabulário da coleção.
- Se três termos de indexação k_a, k_b e k_c ocorrem em um mesmo documento d_j, dizemos que o padrão $[k_a, k_b, k_c]$ de **coocorrência** de termos foi observado.
- Para um vocabulário V de tamanho t, o número total de padrões de coocorrência de termos nos documentos da coleção é 2^t.

Modelo Booleano

- Como exemplo, o padrão $(1, 0, ..., 0)$ indica a presença do termo k_1 e nenhum outro.
- Já o padrão $(1, 1, ..., 1)$ indica a presença de todos os termos.
- Cada um desses padrões de coocorrências de termos é chamado de **componente conjuntivo de termo**.

Modelo Booleano

- A um documento d_j, associa-se um componente conjuntivo de termo $c(q)$ que descreve quais termos ocorrem na consulta e quais não ocorrem.
- O componente conjuntivo de termo $c(d_j)$ fornece uma representação do documento d_j no sistema e o componente conjuntivo de termo $c(q)$ fornece uma representação da consulta q no sistema.

Modelo Booleano

- No modelo Booleano, uma consulta q é uma expressão Booleana sobre os termos da indexação, como, por exemplo: $[q = k_a \wedge (k_b \vee \neg k_c)]$.
- Dada uma consulta, um componente conjuntivo de termo que satisfaz suas condições é chamado de **componente conjuntivo de consulta** $c(q)$.
- Ao se compilar todos os componentes conjuntivos da consulta, pode-se reescrevê-la como uma disjunção desses componentes.
- Essa é conhecida como **forma normal disjuntiva** da consulta, referenciada como q_{DNF}.

Modelo Booleano

- Como exemplo, considere novamente a consulta $[q = k_a \wedge (k_b \vee \neg k_c)]$ e suponha que o vocabulário da coleção seja dado por $V = \{k_a, k_b, k_c\}$.

- A consulta q pode ser reescrita em forma normal disjuntiva como:

$$q_{DNF} = (1, 1, 1) \vee (1, 1, 0) \vee (1, 0, 0) \tag{1}$$

Modelo Booleano

- Essa abordagem funciona mesmo que o vocabulário da coleção contenha termos que não estão na consulta.
- Suponha que o vocabulário do exemplo anterior seja dado por $V = \{k_a, k_b, k_c, k_d\}$.
- Então, um documento d_j que contenha apenas os termos k_a, k_b e k_c é representado pelo componente conjuntivo de termo $(1, 1, 1, 0)$ e a consulta $[q = k_a \wedge (k_b \vee \neg k_c)]$ é representada na forma normal disjuntiva como:

$$\begin{aligned} q_{DNF} = &(1,1,1,0) \vee (1,1,1,1) \vee \\ &(1,1,0,0) \vee (1,1,0,1) \vee \\ &(1,0,0,0) \vee (1,0,0,1) \end{aligned} \tag{2}$$

Modelo Booleano

- Ou seja, o termo k_d é considerado ao mesmo tempo como ausente e presente em cada conjunto conjuntivo, uma vez que ele não foi especificado na consulta.
- Se o documento satisfaz as condições envolvendo os termos da consulta, então existe um componente conjuntivo da consulta que casa com o componente conjuntivo do documento, que nessa situação pode ou não conter k_d.
- Por simplicidade, podemos restringir a representação dos componentes conjuntivos aos termos que ocorrem explicitamente na consulta, conforme será feito adiante.

Modelo Booleano

- Considere $c(q)$ como qualquer dos componentes conjuntivos da consulta q.
- Dado um documento d_j, sendo $c(d_j)$ seu componente conjuntivo de documento correspondente, então a **similaridade** entre o documento d_j e a consulta q é dada por

$$sim(d_j, q) = \begin{cases} 1 & \text{se } \exists c(q) \mid c(q) = c(d_j) \\ 0 & \text{caso contrário} \end{cases} \tag{3}$$

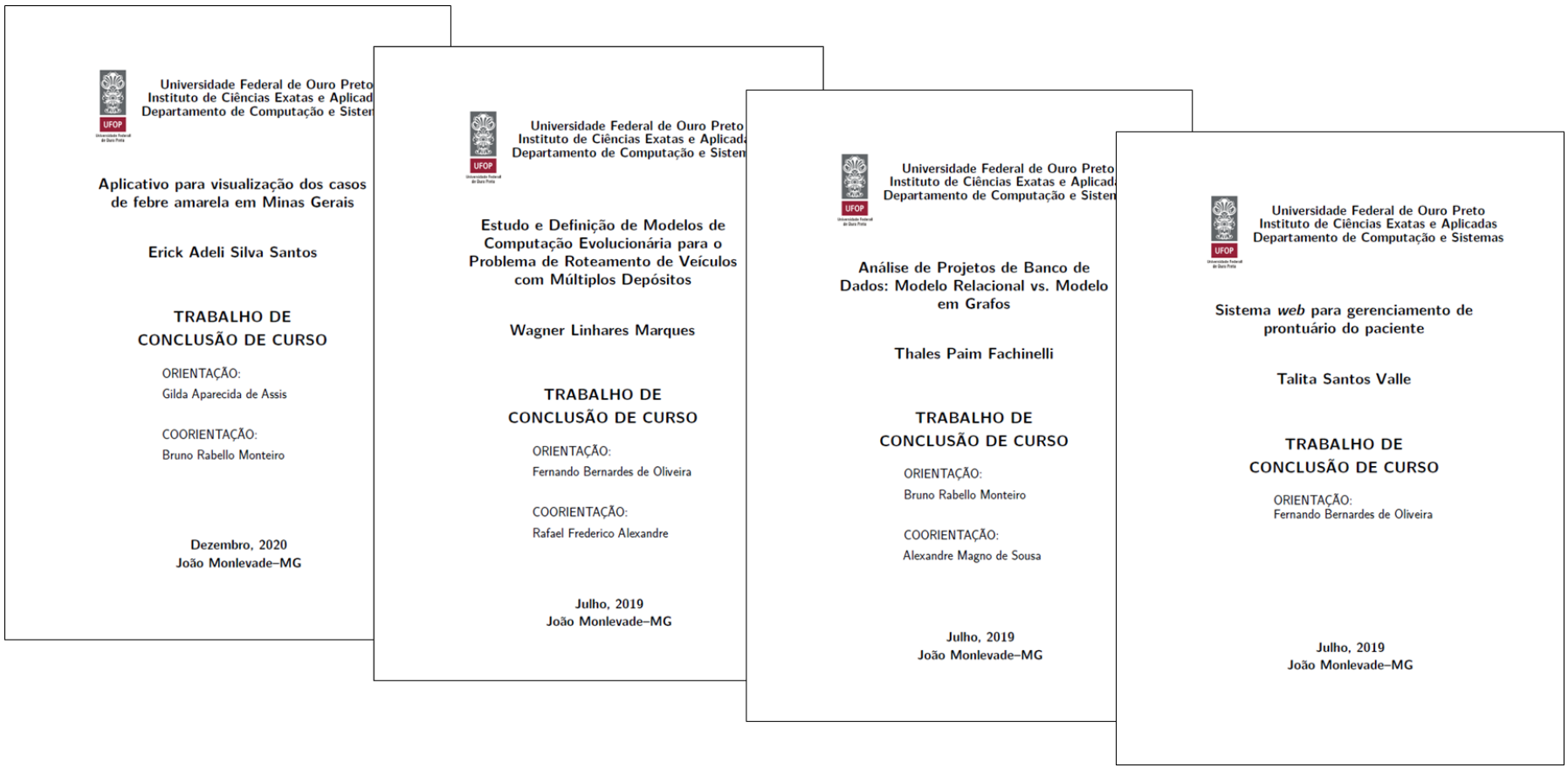
Modelo Booleano

- Note que, no modelo Booleano, **não há satisfação parcial** das condições da consulta.
- Se $sim(d_j, q) = 1$, então o modelo prevê que o documento d_j seja relevante à consulta q; caso contrário a predição é de que d_j não seja relevante.
- Esse critério binário de decisão, sem nenhuma noção de grau, impede uma boa qualidade na recuperação.
- Além disso, com frequência não é possível traduzir uma necessidade de informação em uma expressão Booleana.

4. Ponderação de Termos

Ponderação de Termos

- Dado um conjunto de termos de indexação para um documento, pode-se notar que nem todos os termos são igualmente importantes para descrever o conteúdo dos documentos.



Ponderação de Termos

- A fim de caracterizar a importância dos termos, um **peso** $w_{i,j} > 0$ é associado a cada termo de indexação k_i de um documento d_j na coleção.
- Para um termo de indexação que não aparece em k_i, $w_{i,j} = 0$.
- Para contabilizar a importância dos termos de indexação, deve-se computar os pesos que refletem a importância do termo na coleção e em cada documento em particular.
- Esses pesos dependem das frequências de ocorrência dos termos nos documentos.

Ponderação de Termos

- Seja $f_{i,j}$ a **frequência de ocorrência** do termo de indexação k_i no documento d_j, ou seja, o número de vezes que k_i aparece em d_j.

- A **frequência total** F_i do termo k_i na coleção é a soma das frequências de ocorrência do termo em todos os documentos:

$$F_i = \sum_{j=1}^{N} f_{i,j} \tag{4}$$

onde N é o número de documentos na coleção. A **frequência de documento** para um termo k_i é o número de documentos nos quais k_i ocorre e é indicado simplesmente como n_i. Observe que $n_i \leq F_i$.

Correlação entre Termos

- Até então, os pesos dos termos de indexação tem sido considerados como **mutualmente independentes**.
- Isso significa que saber o peso $w_{i,j}$ associado ao par (k_i, d_j) não traz informação alguma a respeito do peso $w_{i+1,j}$ associado ao par k_{i+1,d_j}.
- Entretanto, usualmente as ocorrências dos termos de indexação nos documentos são **correlacionadas**.
  - Considere, por exemplo, que os termos "dados" e "ciência" sejam utilizados para indexar um documento sobre Ciência dos Dados.
  - Com frequência, nesse documento, a ocorrência de um desses termos atrai a ocorrência do outro. Assim, eles são correlacionados e seus pesos deveriam refletir tal correlação.

5. Ponderação TF-IDF

Ponderação TF-IDF

- A Frequência do Termo (TF, *Term Frequency*) e a Frequência Inversa de Documento (IDF, *Inverse Document Frequency*) são os fundamentos do esquema de ponderação mais popular em RI, denominado TF-IDF.

Ponderação da Frequência de Termos

- O valor ou peso de um termo k_i que ocorre em um documento d_j é simplesmente proporcional à frequência do termo $f_{i,j}$.
- Isto é, quanto mais frequentemente um termo k_i ocorrer no texto do documento d_j maior será a **frequência do termo** $TF_{i,j}$.
- Essa hipótese baseia-se na observação que termos com alta frequência são importantes para descrever os tópicos-chave de um documento, a qual leva à seguinte formulação da ponderação TF:

$$tf_{i,j} = f_{i,j} \tag{5}$$

Ponderação da Frequência de Termos

- Uma variante da ponderação TF também frequentemente usada é

$$tf(i,j) = \begin{cases} 1 + \log_2 f_{i,j} & \text{se } f_{ij} > 0 \\ 0 & \text{caso contrário} \end{cases} \quad (6)$$

Ponderação da Frequência de Termos

- Com exemplo, temos as frequências de cada termo em cada documento da coleção referência ($f_{i,j}$) e os valores logarítmicos de frequência de termos (TF).

#	Termo	$f_{i,1}$	$f_{i,2}$	$f_{i,3}$	$f_{i,4}$	$TF_{i,1}$	$TF_{i,2}$	$TF_{i,3}$	$TF_{i,4}$
1	se	2	0	0	0	2,00	0,00	0,00	0,00
2	eu	4	2	1	3	3,00	2,00	1,00	2,58
3	for	2	0	0	0	2,00	0,00	0,00	0,00
4	vou	2	0	0	3	2,00	0,00	0,00	2,58
5	não	2	1	1	0	2,00	1,00	1,00	0,00
6	sou	0	3	2	0	0,00	2,58	2,00	0,00
7	o	0	2	0	0	0,00	2,00	0,00	0,00
8	que	0	2	1	0	0,00	2,00	1,00	0,00
9	pensam	0	1	1	0	0,00	1,00	1,00	0,00
10	louco	0	0	2	0	0,00	0,00	2,00	0,00
11	pra	0	0	0	1	0,00	0,00	0,00	1,00
12	casa	0	0	0	1	0,00	0,00	0,00	1,00
13	agora	0	0	0	1	0,00	0,00	0,00	1,00

Ponderação da Frequência Inversa de Documentos

- O IDF se baseia nas noções de exaustividade e de especificidade dos termos do vocabulário.
- Exaustividade é uma propriedade das descrições dos documentos e especificidade é uma propriedade dos termos de indexação.
- A **exaustividade** da descrição de um documento é interpretada como a abrangência que ela provê para os tópicos principais do documento.
- A **especificidade** de um termo de indexação é interpretada como quão bem um termo descreve o tópico de um documento.

Ponderação da Frequência Inversa de Documentos

- Se adicionarmos novos termos do vocabulário a um documento, a exaustividade da descrição do documento aumenta.
- Além disso, a probabilidade de que esse documento satisfaça a uma consulta também aumenta.
- De fato, quanto mais termos de indexação são atribuídos a um documento, mais exaustiva fica sua descrição. Sua probabilidade de recuperação em resposta a uma consulta aleatória também aumenta.

Ponderação da Frequência Inversa de Documentos

- Entretanto, se muitos termos estiverem presentes em um documento, ele será retornado mesmo para consultas para os quais não é relevante.
- Isso sugere que o número médio de termos de indexação por documento deve ser otimizado de modo que a probabilidade de relevância de um documento recuperado seja maximizada.
- Esse número ótimo de termos de indexação define a **exaustividade ótima** para as descrições desses documentos.

Ponderação da Frequência Inversa de Documentos

- A especificidade é uma propriedade da semântica do termo, que pode ser mais ou menos específico dependendo do seu significado.
  - Para ilustrar, o termo “veículo”é menos específico que os termos “caminhão”e “carro”.
  - Logo, se a indexação fosse feita manualmente, é esperado que o termo “veículo”fosse usado para indexar mais documentos do que os termos “caminhão”e “carro”.
  - Entretanto, usualmente não é viável interpretar manualmente a especificidade de um tempo.

Ponderação da Frequência Inversa de Documentos

- Uma alternativa a esse problema é considerar a especificidade como uma função da utilização dos termos.
- Logo, a **exaustividade estatística** da descrição de um documento pode ser quantificada como o número de termos de indexação que ele possui.
- Já a **especificidade estatística** de um termo é a função do inverso do número de documentos nos quais ele ocorre.

Ponderação da Frequência Inversa de Documentos

- Naturalmente, se as descrições dos documentos ficarem mais longas, a especificidade dos termos tende a ficar mais baixa.
- Em específico, se um termo ocorre em todos os documentos da coleção sua especificidade é mínima e o termo não é útil para a recuperação.
- Esse raciocínio leva àa ideia da ponderação de termos por especificidade. Para isso, os pesos dos termos podem ser representados como uma função das frequências relativas dos termos.

Ponderação da Frequência Inversa de Documentos

- Considere um termo i, sua **Frequência Inversa de Documentos (IDF)** é:

$$IDF_i = \log \frac{N}{n_i} \tag{7}$$

onde n_i é o número de documentos nos quais o termo i ocorre e N é o número de documentos da coleção.

Ponderação da Frequência Inversa de Documentos

- Como exemplo, a tabela abaixo apresenta os valores de IDF para cada um dos termos presentes nos documentos da coleção referência:

#	Termo	n_i	$IDF_i = \log(N/n_i)$
1	se	1	2,00
2	eu	4	0,00
3	for	1	2,00
4	vou	2	1,00
5	não	3	0,42
6	sou	2	1,00
7	o	1	2,00
8	que	2	1,00
9	pensam	2	1,00
10	louco	1	2,00
11	pra	1	2,00
12	casa	1	2,00
13	agora	1	2,00

Ponderação TF-IDF

- O esquema de ponderação de termos mais empregado em recuperação de informação é a **ponderação TF-IDF**, onde o peso $w_{i,j}$ do termo associado ao para (k_i, d_j) é dado por:

$$w_{i,j} = \begin{cases} (1 + \log f_{i,j}) \times \log \frac{N}{n_i} & \text{se } f_{i,j} > 0 \\ 0 & \text{caso contrário} \end{cases} \quad (8)$$

Ponderação TF-IDF

- Valores de TF-IDF para a coleção de exemplo:

#	Termo	d_1	d_2	d_3	d_4
1	se	4,00	0,00	0,00	0,00
2	eu	0,00	0,00	0,00	0,00
3	for	4,00	0,00	0,00	0,00
4	vou	2,00	0,00	0,00	2,58
5	não	0,84	0,42	0,42	0,00
6	sou	0,00	2,58	2,00	0,00
7	o	0,00	4,00	0,00	0,00
8	que	0,00	2,00	1,00	0,00
9	pensam	0,00	1,00	1,00	0,00
10	louco	0,00	0,00	4,00	0,00
11	pra	0,00	0,00	0,00	2,00
12	casa	0,00	0,00	0,00	2,00
13	agora	0,00	0,00	0,00	2,00

6. Normalização pelo Tamanho dos Documentos

Normalização pelo Tamanho dos Documentos

- Em uma coleção o tamanho dos documentos pode variar consideravelmente.
- Isso é problemático, pois documentos mais longos têm mais chances de serem recuperados em consulta simplesmente por ter mais palavras (e, consequentemente, termos de indexação).
- A fim de compensar esse efeito indesejado, podemos dividir o número de ordem de cada documento pelo seu tamanho - procedimento chamado de **normalização pelo tamanho dos documentos**.

Normalização pelo Tamanho dos Documentos

- As três formas mais usuais de normalização pelo tamanho dos documentos são:

  Tamanho em Bytes é meramente o número de bytes que o documento ocupa em memória. Na maioria dos sistemas 1 caractere equivale a 1 byte.

  Número de Palavras decompõe o documento em palavras e considera a contagem de palavras contidas nesse documento.

  Norma do Vetor é dada pela raiz da somatória dos quadrados dos pesos de cada termo k_i no documento d_j:

$$|\vec{d_j}| = \sqrt{\sum_{i}^{t} w_{i,j}^2} \tag{9}$$

7. Modelo Vetorial

Modelo Vetorial

- Ao contrário do modelo Booleano, permite computar um **grau de similaridade** entre documentos e consultas.
- Os documentos podem então ser **ranqueados** de acordo com esse grau de similaridade.

- De fato, documentos ranqueados fornecem uma resposta mais precisa (no sentido de satisfazer a necessidade de informação do usuário) do que o modelo Booleano.

Modelo Vetorial

- No **modelo vetorial**, os termos de indexação são representados por vetores unitários em um espaço com t dimensões, no qual t é o número de termos de indexação.

- As representações do documento d_j e da consulta q são vetores com t dimensões, dadas por:

$$\begin{aligned} \vec{d_j} &= (w_{1,j}, w_{2,j}, ..., w_{t,j}) \\ \vec{q} &= (w_{1,q}, w_{2,q}, ..., w_{t,q}) \end{aligned} \tag{10}$$

onde $w_{i,q}$ é o peso associado ao par termo-consulta (k_i, q), sendo $w_{i,q} \geq 0$.

Modelo Vetorial

- Assim, um documento d_j e uma consulta de usuário q são representadas como vetores com t dimensões.
- O modelo vetorial então calcula o grau de similaridade do documento d_j com relação à consulta q sob a forma de correlação entre os vetores $\vec{d_j}$ e $\vec{q}$.
- Essa correlação pode ser quantificada, por exemplo, pelo cosseno do ângulo entre esses dois vetores:

$$\begin{aligned} sim(d_j, q) &= \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} \\ &= \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \end{aligned} \tag{11}$$

onde $|\vec{d_j}|$ e $|\vec{q}|$ são as normas dos vetores do documento e da consulta e $\vec{d_j} \bullet \vec{q}$ é o produto interno dos dois vetores.

Modelo Vetorial

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \quad (12)$$

- Note que o fator $|\vec{q}|$ não afeta o ranqueamento uma vez que é o mesmo para todos os documentos. Já fator $|\vec{d_j}|$ faz a normalização pelo tamanho do documento.
- Os pesos adotados no modelo vetorial são aqueles obtidos pelo TF-IDF apresentado anteriormente, assim temos:

$$w_{i,q} = (1 + \log f_{i,q}) \times \log \left(\frac{N}{n_i}\right)$$
$$w_{i,j} = (1 + \log f_{i,j}) \times \log \left(\frac{N}{n_i}\right) \quad (13)$$

Modelo Vetorial

- Um documento pode ser recuperado mesmo que satisfaça apenas **parcialmente** a uma consulta.
- Dessa forma, pode-se inclusive definir um **limiar** para o grau de similaridade $sim(d_j, q)$ e recuperar apenas documentos com similaridade maior que esse limiar.

Modelo Vetorial

- Cálculo e ranqueamento do modelo vetorial para a consulta "eu sou". Termos que não aparecem na consulta ($w_{i,q} = 0$) foram omitidos.

Doc	Computação do Escore	Escore
d_1	$\frac{0{,}00 \times 0{,}00 + 0{,}00 \times 1{,}00}{6{,}057}$	$0,000$
d_2	$\frac{0{,}00 \times 0{,}00 + 2{,}58 \times 1{,}00}{5{,}278}$	$0,489$
d_2	$\frac{0{,}00 \times 0{,}00 + 2{,}00 \times 1{,}00}{4{,}709}$	$0,424$
d_2	$\frac{0{,}00 \times 0{,}00 + 0{,}00 \times 1{,}00}{4{,}322}$	$0,000$

8. Modelo Probabilístico

Modelo Probabilístico

- A ideia principal do **modelo probabilístico** é de que, a partir de uma consulta do usuário, existe um conjunto de documentos que contém exatamente os documentos relevantes e nenhum outro.
- Esse conjunto é chamado conjunto de **resposta ideal**.
- No modelo probabilístico, uma consulta q é um subconjunto dos termos de indexação.
- Um documento d_j é representado por um vetor de pesos binários que indicam a presença ou ausência de termos de indexação:

$$\vec{d_j} = (w_{1,j}, w_{2,j}, ..., w_{t,j}) \tag{14}$$

onde $w_{i,j} = 1$ se o termo k_i ocorre no documento d_j e 0 caso contrário.

Modelo Probabilístico

- Seja R um conjunto de documentos estimados como relevantes para uma consulta q, e seja $\overline{R}$ o conjunto de documentos estimados como não relevantes (complemento de $\overline{R}$).

- $P(R|\vec{d_j}, q)$ é a probabilidade de que o documento d_j com a representação $\vec{d_j}$ **seja** relevante para a consulta q e $P(\overline{R}|\vec{d_j}, q)$ a probabilidade de que o documento d_j **não seja** relevante para a consulta q.

- A similaridade $sim(d_j, q)$ entre o documento d_j e a consulta q é dada por:

$$sim(d_j, q) = \frac{P(R|\vec{d_j}, q)}{P(\overline{R}|\vec{d_j}, q)} \tag{15}$$

Modelo Probabilístico

- Seja R um conjunto de documentos estimados como relevantes para uma consulta q, e seja $\overline{R}$ o conjunto de documentos estimados como não relevantes (complemento de $\overline{R}$).
- $P(R|\vec{d_j}, q)$ é a probabilidade de que o documento d_j com a representação $\vec{d_j}$ **seja** relevante para a consulta q e $P(\overline{R}|\vec{d_j}, q)$ a probabilidade de que o documento d_j **não seja** relevante para a consulta q.
- A similaridade $sim(d_j, q)$ entre o documento d_j e a consulta q é dada por:

$$sim(d_j, q) = \frac{P(R|\vec{d_j}, q)}{P(\overline{R}|\vec{d_j}, q)} \tag{16}$$

Modelo Probabilístico

- Através da regra de Bayes temos:

$$sim(d_j, q) = \frac{P(\vec{d_j}|R, q) \times P(R, q)}{P(\vec{d_j}|\overline{R}, q) \times P(\overline{R}, q)} = \frac{P(\vec{d_j}|R, q) \times P(R|q)}{P(\vec{d_j}|\overline{R}, q) \times P(\overline{R}|q)} \quad (17)$$

- Como $P(R|q)$ e $P(\overline{R}|q)$ são os mesmos para todos os documentos, podemos escrever:

$$sim(d_j, q) \approx \frac{P(\vec{d_j}|R, q)}{P(\vec{d_j}|\overline{R}, q)} \quad (18)$$

Modelo Probabilístico

- Dado que a representação do documento d_j é composta por valores binários que indicam se os termos estão presentes ou não no documento, e assumindo que os termos de indexação são independentes, obtemos:

$$sim(d_j, q) \approx \frac{(\prod_{k_i|w_{i,j}=1} P(k_i|R,q) \times (\prod_{k_i|w_{i,j}=0} P(\overline{k_i}|R,q))}{(\prod_{k_i|w_{i,j}=1} P(k_i|\overline{R},q) \times (\prod_{k_i|w_{i,j}=0} P(\overline{k_i}|\overline{R},q))} \quad (19)$$

onde $P(k_i|R,q)$ é a probabilidade de que o termo de indexação k_i esteja presente em um documento aleatoriamente selecionado a partir do conjunto R de documentos relevantes à consulta q, e $P(\overline{k_i}|R,q)$ é a probabilidade que o termo de indexação k_i não esteja presente em um documento aleatoriamente selecionado do conjunto R. As probabilidades associadas ao conjunto $\overline{R}$ têm significados análogos.

Modelo Probabilístico

- Através de transformações baseadas em logaritmos e produtórios da Equação 19, temos a **expressão-chave** para a computação do ranking no modelo probabilístico:

$$sim(d_j, q) \approx \sum_{k_i \in q \wedge k_i \in d_j} \log \left(\frac{P(k_i|R, q)}{1 - P(k_i|R, q)} \right) + \log \left(\frac{P(k_i|\overline{R}, q)}{1 - P(k_i|\overline{R}, q)} \right) \quad (20)$$

Modelo Probabilístico

- A princípio, a **limitação** desse modelo é que não conhecemos de antemão o conjunto R.
- Ademais, não é viável solicitar ao usuário que defina o conjunto R manualmente.
- Dessa forma, faz-se necessário definir um método para computar de forma autônoma as probabilidades $P(k_i|R, q)$ e $P(k_i|\overline{R}, q)$.

Tabela de Contingência das Incidências de Termos

- Seja N o número de documentos da coleção, n_i o número de documentos que contêm o termo k_i, R o número total de documentos que atendem à consulta q e r_i o número de documentos relevantes que contêm o termo k_i, temos a seguinte **tabela de contingência das incidências de termos**:

Caso	Relevantes	Não relevantes	Total
Documentos que contêm k_i	r_i	$n_i - r_i$	n_i
Documentos que não contêm k_i	$R - r_i$	$N - n_i - (R - r_i)$	$N - n_i$
Todos os documentos	R	$N - R$	N

Modelo Probabilístico

- Assumindo que a informação da tabela de contingência das incidências de termos está disponível para uma dada consulta (o que não é verdade pois não se tem informação sobre quais documentos são relevantes ou não para uma consulta), poderíamos escrever:

$$P(k_i|R,q) = \frac{r_i}{R}$$
$$P(k_i|\overline{R},q) = \frac{n_i - r_i}{N - R} \tag{21}$$

e reescrever a Equação 20 como:

$$sim(d_j,q) \approx \sum_{k_i \in q \wedge k_i \in d_j} \log \frac{r_i(N - n_i - R + r_i)}{(R - r_i)(n_i - r_i)} \tag{22}$$

Modelo Probabilístico

- Para que essa equação seja efetiva é necessário ainda estimar quais são os documentos relevantes para a consulta.
- Para lidar com valores pequenos de r_i, é conveniente somar $0,5$ a cada um dos termos da equação anterior, levando a:

$$sim(d_j, q) \approx \sum_{k_i \in q \wedge k_i \in d_j} \log \frac{(r_i + 0,5)(N - n_i - R + r_i + 0,5)}{(R - r_i + 0,5)(n_i - r_i + 0,5)} \quad (23)$$

- A Equação 23 ainda não pode ser computada sem estimativas de r_i e R. Uma possibilidade é **considerar** $R = r_i = 0$, assim temos:

$$sim(d_j, q) \approx \sum_{k_i \in q \wedge k_i \in d_j} \log \frac{N - n_i + 0,5}{n_i + 0,5} \quad (24)$$

Modelo Probabilístico

- Essa equação permite valores negativos de ranqueamento, e pode ter comportamentos anômalos.
- Uma alternativa para sanar essa anomalia é eliminar o fator n_i do numerador da Equação 24, levando a:

$$sim(d_j, q) \approx \sum_{k_i \in q \wedge k_i \in d_j} \log \frac{N + 0,5}{n_i + 0,5} \tag{25}$$

Modelo Probabilístico

- Escores obtidos no modelo probabilístico, computado pela Equação 25, para a consulta "eu sou":

Doc	Computação do Escore	Escore
d_1	$\log\left(\frac{4+0,5}{4+0,5}\right)$	$0,000$
d_2	$\log\left(\frac{4+0,5}{4+0,5}\right) + \log\left(\frac{4+0,5}{2+0,5}\right)$	$0,848$
d_3	$\log\left(\frac{4+0,5}{4+0,5}\right) + \log\left(\frac{4+0,5}{2+0,5}\right)$	$0,848$
d_4	$\log\left(\frac{4+0,5}{4+0,5}\right)$	$0,000$

9. Modelo Alternativo BM25

BM25 (*Best Match* 25)

- O modelo vetorial tem um ótimo desempenho para coleções genéricas pois é baseada nos seguintes fatores:
  1. frequência inversa de documentos;
  2. frequência de termos;
  3. normalização pelo tamanho dos documentos.

- O Modelo Probabilístico clássico considera **apenas o fator (1)** em sua equação...

- O **Modelo BM25** foi criado como resultados de uma série de experimentos sobre variações da fórmula clássica do Modelo Probabilístico (eq. 24).

Fórmulas BM1, BM11 e BM15

- A primeira ideia para melhorar o ranking foi introduzir a **frequência de termos** na Equação 24.

$$\mathcal{F}_{i,j} = S_1 \times \frac{f_{i,j}}{K_1 + f_{i,j}} \tag{26}$$

- onde
  - $f_{i,j}$ é a frequência do termo k_i no documento d_j.
  - K_1 é uma constante que pode ser definida experimentalmente para uma coleção em particular.
  - S_1 é uma constante de escala relativa a K_1, usualmente $S_1 = (K_1 + 1)$.

- Note que se $K_1 = 0$, o fator completo se torna 1 e não produz efeito no ranking.

Fórmulas BM1, BM11 e BM15

- O próximo passo foi introduzir a **normalização pelo tamanho dos documentos**:

$$\mathcal{F}'_{i,j} = S_1 \times \frac{f_{i,j}}{\frac{K_1 \times len(d_j)}{avg\_doclen} + f_{i,j}} \tag{27}$$

- onde
  - $len(d_j)$ é o tamanho do documento.
  - avg_doclen é o tamanho médio dos documentos.

Fórmulas BM1, BM11 e BM15

- Adicionalmente, um **fator de correção** $\mathcal{G}_{j,q}$ dependente do tamanho dos documentos e da consulta é considerado:

$$\mathcal{G}_{j,q} = K_2 \times len(q) \times \frac{avg\_doclen - len(d_j)}{avg\_doclen + len(d_j)} \tag{28}$$

- onde
  - $len(q)$ é o tamanho da consulta (número de termos).
  - K_2 é uma segunda constante ajustada como parâmetro.

Fórmulas BM1, BM11 e BM15

- Um raciocínio análogo pode ser aplicado às **frequências dos termos dentro das consultas**, levando a um fator adicional dado por

$$\mathcal{F}_{i,q} = S_3 \times \frac{f_{i,q}}{K_3 + f_{i,q}} \tag{29}$$

- onde
  - $f_{i,q}$ é a frequência do termo k_i na consulta q.
  - K_3 é uma constante.
  - S_3 é uma constante de escala relativa a K_3.

- Usualmente $S_3 = K_3 + 1$.

Fórmulas BM1, BM11 e BM15

- A introdução desses fatores na equação 24 leva a várias formas BM:

$$sim_{BM1}(d_j, q) \approx \sum_{k_i \in q \wedge k_i \in d_j} \log \frac{N - n_i + 0,5}{n_i + 0,5} \tag{30}$$

$$sim_{BM15}(d_j, q) \approx \mathcal{G}_{j,q} + \sum_{k_i \in q \wedge k_i \in d_j} \mathcal{F}_{i,j} \times \mathcal{F}_{i,q} \times \log \frac{N - n_i + 0,5}{n_i + 0,5} \tag{31}$$

$$sim_{BM11}(d_j, q) \approx \mathcal{G}_{j,q} + \sum_{k_i \in q \wedge k_i \in d_j} \mathcal{F}'_{i,j} \times \mathcal{F}_{i,q} \times \log \frac{N - n_i + 0,5}{n_i + 0,5} \tag{32}$$

Fórmulas BM1, BM11 e BM15

- Evidências empíricas sugerem $K_2 = 0$, o que elimina o fator de correção $\mathcal{G}_{j,q}$ dessas equações.
- Além disso, boas estimativas para as constantes de escala são $S_1 = K_1 + 1$ e $S_3 = K_3 + 1$, com evidências sugerindo que valores altos de K_3 são bons.
- Para consultas curtas podemos usar $f_{i,q} = 1$ para todos os termos.

Fórmulas BM1, BM11 e BM15

- Essa considerações levam a equações mais simples:

$$sim_{BM1}(d_j, q) \approx \sum_{k_i \in q \wedge k_i \in d_j} \log \frac{N - n_i + 0,5}{n_i + 0,5} \tag{33}$$

$$sim_{BM15}(d_j, q) \approx \sum_{k_i \in q \wedge k_i \in d_j} \frac{(K_1 + 1) f_{i,j}}{(K_1 + f_{i,j})} + \log \frac{N - n_i + 0,5}{n_i + 0,5} \tag{34}$$

$$sim_{BM11}(d_j, q) \approx \sum_{k_i \in q \wedge k_i \in d_j} \frac{(K_1 + 1) f_{i,j}}{(\frac{K_1 \times len(d_j)}{avg\_doclen} + f_{i,j})} + \log \frac{N - n_i + 0,5}{n_i + 0,5} \tag{35}$$

Fórmula BM25

- A BM25 foi criada como uma **combinação das fórmulas BM11 e BM15**.
- A motivação foi combinar os itens como segue:

$$\mathcal{B}_{i,j} = \frac{(K_1 + 1) f_{i,j}}{K_1((1 - b) + b \frac{len(d_j)}{avg\_doclen}) + f_{i,j}} \tag{36}$$

- onde
  - b é uma nova constante introduzida com valores entre 0 e 1.
  - Se $b = 0$, a equação acima é reduzida ao fator de frequência de termos usado na BM15.
  - Se $b = 1$, ela é equivalente ao fator de frequência de termos usado na BM11.
  - Para valores de b entre 0 e 1, a equação provê uma combinação de BM11 com BM15.

Fórmula BM25

- A equação de ranqueamento do **Modelo BM25** pode então ser escrita como:

$$sim_{BM25}(d_j, q) \approx \sum_{k_i \in q \wedge k_i \in d_j} \mathcal{B}_{i,j} \times \log \frac{N - n_i + 0,5}{n_i + 0,5} \tag{37}$$

- onde K_1 e b são constantes empíricas
  - $K_1 = 1$ funciona bem em coleções genéricas.
  - b deve ser mantido próximo a 1 para enfatizar a normalização pelo tamanho de documento (ex.: $b = 0.75$).
  - Esses valores devem ser ajustados para coleções através de experimentação e **avaliação da recuperação**.

Fórmula BM25

- Ao contrário do Modelo Probabilístico a fórmula BM25 pode ser calculada sem informação prévia sobre a relevância.
- Há um consenso que o BM25 supera o modelo Vetorial clássico para coleções genéricas.

10. Exercícios

Exercícios

1. A visibilidade de aplicativos e páginas Web é essencial para que o produto obtenha novos usuários. Suponha que você tenha que criar a descrição de um aplicativo com as mesmas funcionalidades do YouTube (porém menos poluído de propagandas), chamado NoAdTube. Elabore uma descrição para esse aplicativo de modo que: (i) demonstre suas principais funcionalidades aos usuários; (ii) tenha um bom escore para ranqueamento em consultas; (iii) esteja no limite de palavras permitidas pela PlayStore (entre 30 e 150).

Exercícios

2 Apresente as principais vantagens e desvantagens dos modelos Booleano, vetorial e probabilístico para o ranqueamento de documentos com relação a uma consulta.

3 Calcule a similaridade entre cada documento da coleção abaixo e a consulta "eu bebo" de acordo com os modelos Booleano, vetorial e probabilístico.

Ai se eu te pego, ai, ai se eu te pego.	Aí eu bebo, aí eu bebo! Bebo pra carai!	Só eu que bebo? Será que só eu que bebo.	Ai, ai, ai, ai esse amor, é bom demais
d_1	d_2	d_3	d_4

Referências

- Baeza-Yates, R; Ribeiro-Neto, B. *Modern Information Retrieval: The Concepts and Technology behind Search*. 2ªedição. Pearson, 2011 (capítulo 2).