



INSTITUTO FEDERAL
MINAS GERAIS

RECUPERAÇÃO DE INFORMAÇÃO

Profa. Patrícia Proença

patricia.proenca@ifmg.edu.br



ATENÇÃO!!!

↓ O material a seguir é uma videoaula apresentada pela professora PATRÍCIA APARECIDA PROENÇA AVILA, como material pedagógico do IFMG, dentro de suas atividades curriculares ofertadas em ambiente virtual de aprendizagem. Seu uso, cópia e ou divulgação em parte ou no todo, por quaisquer meios existentes ou que vierem a ser desenvolvidos, somente poderá ser feito, mediante autorização expressa deste docente e do IFMG. Caso contrário, estarão sujeitos às penalidades legais vigentes”.


↓ Conforme Art. 2º§1º da Nota Técnica nº
1/2020/PROEN/Reitoria/IFMG (SEI 0605498, Processo nº
23208.002340/2020-04

MODELO PROBABILÍSTICO





Modelo probabilístico

- ▶ Proposto em 1976 por Robertson e Sparck;
 - ▶ Propõe uma solução ao problema de RI com base na teoria das probabilidades.
- 



Ideia fundamental

- ▶ A partir de uma consulta do usuário, existe um **conjunto de documentos** que contém exatamente os **documentos relevantes (resposta ideal)** e nenhum outro;
- ▶ Dada uma descrição desse **conjunto resposta ideal**, poderíamos recuperar os documentos relevantes;
- ▶ Quais são essas propriedades dessa descrição?
 - ▶ Resposta: não sabemos! Tudo que sabemos é que existem termos de indexação para caracterizar tais propriedades.



Ideia fundamental

- Problema:
 - Essas propriedades não são conhecidas na hora da consulta!
 - É necessário um esforço para conseguir uma **estimativa inicial dessas propriedades**.
- Essa estimativa inicial nos permite gerar uma descrição probabilística preliminar do **conjunto resposta ideal**, que pode ser utilizado para recuperar um primeiro conjunto de documentos.



Ideia fundamental

- ▶ Por exemplo:
 - ▶ O usuário pode ver os documentos recuperados e decidir quais são relevantes e quais não são;
 - ▶ O sistema pode então utilizar essa informação para refinar a descrição do conjunto resposta ideal;
 - ▶ Repetindo-se esse processo muitas vezes, espera-se que a descrição do conjunto resposta ideal fique mais precisa;
- ▶ **IMPORTANTE:** é necessário estimar, no início, a descrição do **conjunto resposta ideal**.



Ideia fundamental - Similaridade

- ▶ Como calcular a medida de similaridade?
 - ▶ Como criar uma função que irá ranquear os resultados?
- ▶ Será usada a chance ou *razão de possibilidade*;
 - ▶ Modo de quantificar o quão forte a presença (ou ausência) da propriedade A está associada a presença (ou ausência) da propriedade B;
 - ▶ Relação: documento d_j ser relevante a q e o documento d_j não ser relevante a q ; (proporção de sucessos / proporção de falhas);

Ideia fundamental - Exemplo

- ▶ Suponhamos que em uma amostra de 100 homens, 90 beberam vinho na semana anterior;
- ▶ Em um grupo similar de 100 mulheres, apenas 20 beberam vinho no mesmo período;
- ▶ Portanto, a probabilidade de um homem beber vinho é de 90 para 10, ou 9:1, enquanto que a chance de uma mulher beber vinho é de 20 para 80, ou $1:4 = 0,25:1$;
 - ▶ Razão: proporção de sucessos / proporção de falhas
- ▶ Podemos calcular então a razão de chances como sendo $9/0.25$, ou 36, mostrando que homens tem muito mais chances de beber vinho do que mulheres.
 - ▶ O quão forte a presença de homens que bebem vinho está associada a presença de mulheres beberem vinho;

Definição

- 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 a ausência de termos de indexação, como segue

$$\vec{d}_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$$

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

Definição

- Seja R um conjunto de documentos inicialmente estimado como relevante para o usuário para a consulta q . Seja o complemento de R (o conjunto de documentos não relevantes). A similaridade $\text{sim}(d_j, q)$ entre o documento d_j e a consulta q é definida por:

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



Expressão chave para a computação do ranking no modelo probabilístico

- ▶ Ao aplicarmos:
 - ▶ Regra de Bayes;
 - ▶ Hipótese de independência;
 - ▶ Uso de logaritmos;
 - ▶ Simplificação de notação;
 - ▶ Conversão de produtório de logaritmo para somatório de logaritmo;

Expressão chave para a computação do ranking no modelo probabilístico

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

- ▶ p_{iR} é a probabilidade que o termo de indexação k_i esteja em um documento aleatoriamente selecionado a partir do conjunto R de relevantes à consulta q .
- ▶ q_{iR} é a probabilidade que o termo de indexação k_i esteja presente em um documento aleatoriamente selecionado a partir do conjunto de não relevantes à consulta q .
- ▶ **Como não conhecemos o conjunto R no princípio do processo, é necessário definir um método para, inicialmente, computar as probabilidades p_{iR} e q_{iR} .**

Estimar as probabilidades relacionadas ao conjunto de documentos relevantes

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

- Seja N o número de documentos da coleção e n_i o número de documentos que contêm o termo k_i .
- Seja R o número total de documentos para a consulta q (na opinião do usuário) e r_i o número de documentos relevantes que contêm o termo k_i .

Estimar as probabilidades relacionadas ao conjunto de documentos relevantes

- Se a informação na tabela estivesse disponível para qualquer consulta, poderíamos escrever:

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

- e reescrever a equação original da seguinte forma:

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

Estimar as probabilidades relacionadas ao conjunto de documentos relevantes

- Para lidar com valores pequenos de r_i , é conveniente somar 0,5 a cada um dos termos da fórmula anterior:

$$\text{sim}(d_j, q) \sim \sum_{k_i[q, d_j]} \log \left(\frac{(r_i + 0,5)(N - n_i - R + r_i + 0,5)}{(R - r_i + 0,5)(n_i - r_i + 0,5)} \right)$$

- Essa fórmula é conhecida como equação Robertson-Spark Jones e é considerada a equação de ranqueamento clássica para o modelo probabilístico.
- Comporta-se bem para estimativas particulares como $R = r_i$.

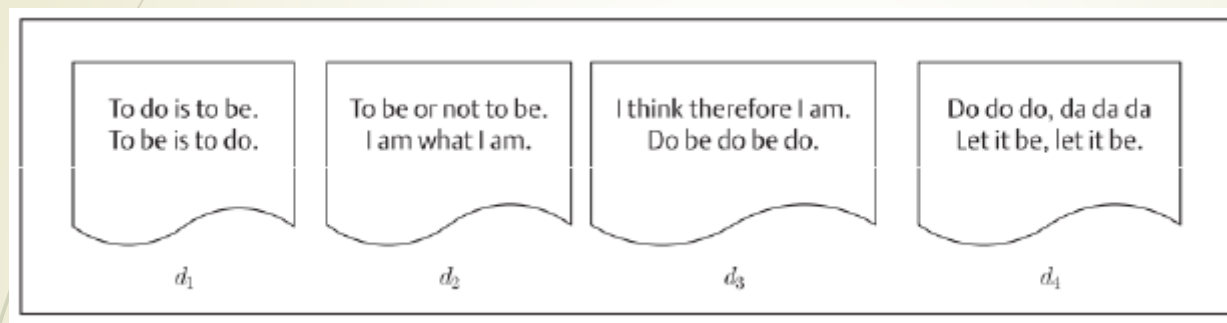
Estimativa ($R = r_i = 0$)

- Ausência de informação quanto à relevância dos documentos:

$$\text{sim}(d_j, q) \sim \sum_{k_i \in q \wedge k_i \in d_j} \log \left(\frac{N - n_i + 0,5}{n_i + 0,5} \right)$$

- Essa equação apresenta problemas quando $n_i > N/2$.

Estimativa ($R = r_i = 0$) - Exemplo



#	termo	$f_{i,1}$	$f_{i,2}$	$f_{i,3}$	$f_{i,4}$
1	to	4	2	—	—
2	do	2	—	3	3
3	is	2	—	—	—
4	be	2	2	2	2
5	or	—	1	—	—
6	not	—	1	—	—
7	I	—	2	2	—
8	am	—	2	1	—
9	what	—	1	—	—
10	think	—	—	1	—
11	therefore	—	—	1	—
12	da	—	—	—	3
13	let	—	—	—	2
14	it	—	—	—	2
Tamanho do documento (# palavras)		10	11	10	12

Estimativa ($R = r_i = 0$) - Exemplo

➤ Consulta: **to do**

$$\text{sim}(d_j, q) \sim \sum_{k_i \in q \wedge k_i \in d_j} \log \left(\frac{N - n_i + 0,5}{n_i + 0,5} \right)$$

termo	$f_{i,1}$	$f_{i,2}$	$f_{i,3}$	$f_{i,4}$
to	4	2	—	—
do	2	—	3	3
is	2	—	—	—
be	2	2	2	2
or	—	1	—	—
not	—	1	—	—
I	—	2	2	—
am	—	2	1	—
what	—	1	—	—
think	—	—	1	—
therefore	—	—	1	—
da	—	—	—	3
let	—	—	—	2
it	—	—	—	2

Doc	Computação do escore	Escore
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

Ajuste para ($R = r_i = 0$)

- Para evitar o comportamento anômalo mostrado anteriormente, podemos eliminar o fator n_i do numerador da equação anterior, conforme sugerido por Robertson e Walker (1997):

$$\text{sim}(d_j, q) \sim \sum_{k_i \in q \wedge k_i \in d_j} \log \left(\frac{N + 0,5}{n_i + 0,5} \right)$$

- Dessa forma, um termo que ocorre em todos os documentos ($n_i = N$) produz um peso igual a zero ($\log(1)=0$) e não existem mais pesos negativos.

Ajuste para ($R = r_i = 0$) - Exemplo

$$\text{sim}(d_j, q) \sim \sum_{k_i \in q \wedge k_i \in d_j} \log \left(\frac{N + 0,5}{n_i + 0,5} \right)$$


Doc	Computação do escore	Escore
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

Alternativa para estimar R e r_i

- ▶ As equações anteriores consideram que $R=r_i=0$.
- ▶ Uma alternativa para estimar R e r_i mais cuidadosamente é:
 - ▶ 1) Fazer a busca inicial utilizando a equação com $R=r_i=0$;
 - ▶ 2) Selecionar os 10-20 documentos mais bem ranqueados;
 - ▶ 3) Inspeccionar os documentos para obter novas estimativas para R e r_i ;
 - ▶ 4) Remover esses 10-20 documentos da coleção;
 - ▶ 5) Reprocessar a consulta com as novas estimativas.




Vantagem do modelo probabilístico

- Os documentos são ranqueados de acordo com sua probabilidade de serem relevantes, com base na informação disponível ao sistema.
- 



Desvantagens do modelo probabilístico

- 1) Relevância de um documento é afetada por diversos fatores externos, não somente na informação disponível ao sistema;
 - 2) Necessidade de estimar a separação inicial dos documentos em conjuntos relevantes e não relevantes;
 - 3) Não leva em consideração a frequência na qual um termo de indexação ocorre em um documento;
 - 4) Falta de normalização pelo tamanho dos documentos.
- 

Comparação entre os modelos clássicos

- 1) Modelo booleano é considerado o mais fraco entre os modelos clássicos;
- 2) O maior problema do modelo booleano é a falta de casamento parcial entre a consulta e os documentos;
- 3) Existe controvérsia quanto ao modelo probabilístico ser melhor do que o vetorial:
 - a) Experimentos realizados por Croft indicam que o modelo probabilístico fornece melhor qualidade de recuperação;
 - b) Outros experimentos conduzidos por Salton e Buckley contestam esses resultados.
- 4) Com coleções genéricas, o modelo vetorial fornece um modelo de RI razoável e robusto para fins de comparação.



MUNDO
MENSAGENS

*Feliz início de semana!
Que suas forças estejam
restabelecidas para agarrar
as oportunidades que os
próximos dias trarão.*