

# **Modelo Probabilístico**

**Wendel Melo**

Faculdade de Computação  
Universidade Federal de Uberlândia

Recuperação da Informação

Adaptado do Material da Profª Vanessa Braganholo - IC/UFF

# Modelo Probabilístico

- Proposto em 1976 por Roberstson e Sparck Jones;
- Se baseia na premissa de que, para cada consulta do usuário, existe um conjunto resposta **ideal** (que contém apenas os documentos relevantes e nenhum outro mais);
- O objetivo é encontrar uma resposta para cada consulta que aproxime o conjunto ideal por meio de **formalismo probabilístico** e **refinamento iterativo**.

# Modelo Probabilístico

- Assim, tenta-se **estimar** a probabilidade do usuário considerar cada documento  $d_j$  como relevante;
- O modelo supõe que essa probabilidade de relevância depende apenas das representações da consulta e dos documentos (o que pode ser demasiadamente simplista, pois desconsidera variáveis externas ao sistema);
- Desse modo, as consultas são então vistas como especificações das propriedades do conjunto resposta ideal.

# Modelo Probabilístico – Funcionamento

- 1) Um conjunto inicial de documentos é recuperado;
- 2) O usuário inspeciona os documentos e indica os relevantes (em geral, só os primeiros do ranking são analisados);
- 3) A informação obtida no passo 2 é usada para refinar os parâmetros do modelo em busca do conjunto resposta ideal;
- 4) Repetindo-se o processo, espera-se conjunto resposta do algoritmo melhor. Assim, volta-se ao passo 1.

# Modelo Probabilístico

- Observe que o modelo é iterativo, e foi originalmente concebido para receber intervenção do usuário (o que contraria, de certo modo, a filosofia de RI);
- Assim, o sistema busca evoluir sua resposta por meio do “aprendizado” obtido com o usuário (o escopo desse “aprendizado” é apenas a consulta sendo respondida);
- Posteriormente, foram propostos esquemas de refinamento iterativo automático para evitar intervenções do usuário.

# Modelo Probabilístico

- Cada documento é representado por um vetor de pesos binários que indicam presença ou ausência dos termos de indexação:

$$d_j = (w_{1j}, w_{2j}, \dots, w_{Tj})$$

onde:

$$w_{ij} = \begin{cases} 1, & \text{se o termo } k_i \text{ aparece em } d_j, \\ 0, & \text{caso contrário.} \end{cases}$$

# Modelo Probabilístico

- Assim, dada uma consulta  $q$ , sejam:  
  
 $R$ : conjunto de documentos relevantes à  $q$ , isto é, o conjunto ideal, (o qual não se conhece ainda);  
  
 $\bar{R}$ : o conjunto de documentos não relevantes a  $q$ , (o qual também não se conhece ainda).  
  
• Desse modo, para cada documento  $d_j$ , são estimadas as probabilidades de  $d_j$  pertencer a  $R$ , e de  $d_j$  pertencer a  $\bar{R}$ .

# Modelo Probabilístico

- Dessa forma, a similaridade entre cada documento  $d_j$  e a consulta  $q$  é dada por uma razão entre as duas probabilidades, denominada *probabilidade de relevância*

$$\text{sim}(d_j, q) = \frac{P(d_j \text{ relevante a } q)}{P(d_j \text{ não relevante a } q)} = \frac{P(R \mid d_j)}{P(\bar{R} \mid d_j)}$$



# Modelo Probabilístico

- Dessa forma, a similaridade entre cada documento  $d_j$  e a consulta  $q$  é dada por uma razão entre as duas probabilidades, denominada *probabilidade de relevância*

$$\text{sim}(d_j, q) = \frac{P(d_j \text{ relevante a } q)}{P(d_j \text{ não relevante a } q)} = \frac{\overbrace{P(R \mid d_j)}^{\text{Prob de } d_j \text{ ser relevante a } q}}{\underbrace{P(\bar{R} \mid d_j)}_{\text{Prob de } d_j \text{ não ser relevante a } q}}$$

- Essa razão é adotada com o objetivo de minimizar a probabilidade de um julgamento errôneo;
- Como calcular essas probabilidades?

# Modelo Probabilístico

- Dessa forma, a similaridade entre cada documento  $d_j$  e a consulta  $q$  é dada por uma razão entre as duas probabilidades, denominada *probabilidade de relevância*

$$\text{sim}(d_j, q) = \frac{P(d_j \text{ relevante a } q)}{P(d_j \text{ não relevante a } q)} = \frac{\overbrace{P(R \mid d_j)}^{\text{Prob de } d_j \text{ ser relevante a } q}}{\underbrace{P(\bar{R} \mid d_j)}_{\text{Prob de } d_j \text{ não ser relevante a } q}}$$

- Essa razão é adotada com o objetivo de minimizar a probabilidade de um julgamento errôneo;
- Como calcular essas probabilidades? **Não se sabe ao certo!**

# Modelo Probabilístico

Para se tentar estimar as probabilidades de relevância e não relevância usa-se:

$P(k_i|R)$  : Probabilidade de se obter o termo  $k_i$  em um documento aleatório do conjunto de documentos relevantes.

$P(k_i|\bar{R})$  : Probabilidade de se obter o termo  $k_i$  em um documento aleatório do conjunto de documentos **não** relevantes.

# Modelo Probabilístico

Para se tentar estimar as probabilidades de relevância e não relevância usa-se:

$P(k_i|R)$  : Probabilidade de se obter o termo  $k_i$  em um documento aleatório do conjunto de documentos relevantes.

$P(k_i|\bar{R})$  : Probabilidade de se obter o termo  $k_i$  em um documento aleatório do conjunto de documentos **não** relevantes.

Assim, após algumas assumições, algebrismos e uso de logaritmos:

$$sim(d_j, q) \sim \sum_{i=1}^T w_{ij} w_{iq} \left( \log \frac{P(k_i|R)}{1 - P(k_i|R)} + \log \frac{1 - P(k_i|\bar{R})}{P(k_i|\bar{R})} \right)$$

# Modelo Probabilístico

$$\text{sim}(d_j, q) \sim \sum_{i=1}^T \underbrace{w_{ij}w_{iq}} \left( \log \frac{P(k_i|R)}{1 - P(k_i|R)} + \log \frac{1 - P(k_i|\bar{R})}{P(k_i|\bar{R})} \right)$$

Note que só é preciso computar as parcelas referentes a termos que apareçam tanto na consulta  $q$  quanto no documento  $d_j$

# Modelo Probabilístico

$$\text{sim}(d_j, q) \sim \sum_{i=1}^T \underbrace{w_{ij}w_{iq}} \left( \log \frac{P(k_i|R)}{1 - P(k_i|R)} + \log \frac{1 - P(k_i|\bar{R})}{P(k_i|\bar{R})} \right)$$

Note que só é preciso computar as parcelas referentes a termos que apareçam tanto na consulta  $q$  quanto no documento  $d_j$

- A questão agora é: Como calcular  $P(k_i|R)$  e  $P(k_i|\bar{R})$  ?

# Modelo Probabilístico

$$\text{sim}(d_j, q) \sim \sum_{i=1}^T \underbrace{w_{ij}w_{iq}} \left( \log \frac{P(k_i|R)}{1 - P(k_i|R)} + \log \frac{1 - P(k_i|\bar{R})}{P(k_i|\bar{R})} \right)$$

Note que só é preciso computar as parcelas referentes a termos que apareçam tanto na consulta  $q$  quanto no documento  $d_j$

- A questão agora é: Como calcular  $P(k_i|R)$  e  $P(k_i|\bar{R})$  ?
  - Não se sabe ao certo!

# Modelo Probabilístico

- Na prática, adota-se inicialmente:

$$P(k_i|R) = 0,5 \quad P(k_i|\bar{R}) = \frac{n_i}{N} \quad \left. \begin{array}{l} \} \text{Nº de docs com o termo } k_i \\ \} \text{Nº total de docs} \end{array} \right\}$$

- A partir daí, calcula-se as similaridades e recupera-se um conjunto inicial de documentos. Este conjunto é então utilizado para atualizar as probabilidades iterativamente:

$$P(k_i|R) = \frac{V_i}{V} \quad P(k_i|\bar{R}) = \frac{n_i - V_i}{N - V}$$

Onde  $V$  é o nº de docs inicialmente recuperados (podem ser só os primeiros do ranking) e  $V_i$  é o nº de docs inicialmente recuperados que contém o termo  $k_i$ .



# Modelo Probabilístico

- Para evitar problemas com  $V$  e  $V_i$  muito pequenos, um fator de ajuste  $\varphi_i$  pode ser adicionado à formula:

$$P(k_i|R) = \frac{V_i + \varphi_i}{V + 1}$$

$$P(k_i|\bar{R}) = \frac{n_i - V_i + \varphi_i}{N - V + 1}$$

- Pode-se utilizar:

$$\varphi_i = 0,5$$

ou

$$\varphi_i = \frac{n_i}{N}$$

# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos,  $A$ ,  $B$  e  $C$ . Consulta:  $A \ C$

D1	AAAB
D2	AAC
D3	AA
D4	BB
D5	BC

$$\text{sim}(d_j, q) \sim \sum_{i=1}^T w_{ij} w_{iq} \left( \log \frac{P(k_i|R)}{1 - P(k_i|R)} + \log \frac{1 - P(k_i|\bar{R})}{P(k_i|\bar{R})} \right)$$

$w_{ij} = 1$  se o termo  $k_i$  está em  $d_j$ , e 0 caso contrário.

Inicialmente, adotamos:

$$P(k_i|R) = 0,5 \quad P(k_i|\bar{R}) = \frac{n_i}{N} \quad \left. \begin{array}{l} \text{Nº de docs com o termo } k_i \\ \text{Nº total de docs} \end{array} \right\}$$

$$P(A|R) = 0,5$$

$$P(B|R) = 0,5$$

$$P(C|R) = 0,5$$

$$P(A|\bar{R}) = \frac{3}{5} = 0,6$$

$$P(B|\bar{R}) = \frac{3}{5} = 0,6$$

$$P(C|\bar{R}) = \frac{2}{5} = 0,4$$

# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos,  $A$ ,  $B$  e  $C$ . Consulta:  $A\ C$

D1	AAAB
D2	AAC
D3	AA
D4	BB
D5	BC C

$$P(A|R) = 0,5$$

$$P(C|R) = 0,5$$

$$P(A|\bar{R}) = \frac{3}{5} = 0,6$$

$$P(C|\bar{R}) = \frac{2}{5} = 0,4$$

$$\text{sim}(d_j, q) \sim \sum_{i=1}^T w_{ij} w_{iq} \left( \log \frac{P(k_i|R)}{1 - P(k_i|R)} + \log \frac{1 - P(k_i|\bar{R})}{P(k_i|\bar{R})} \right)$$

Primeiro, calculamos as parcelas referentes aos logaritmos. Como a consulta só é composta por  $A$  e  $C$ , só precisamos calcular as parcelas referentes a estes dois termos.

$$\log \frac{P(A|R)}{1 - P(A|R)} = \log \frac{0,5}{1 - 0,5} = \log 1 = 0$$

$$\log \frac{P(C|R)}{1 - P(C|R)} = \log \frac{0,5}{1 - 0,5} = \log 1 = 0$$

$$\log \frac{1 - P(A|\bar{R})}{P(A|\bar{R})} = \log \frac{1 - 0,6}{0,6} = \log \frac{0,4}{0,6} = -0,176$$

$$\log \frac{1 - P(C|\bar{R})}{P(C|\bar{R})} = \log \frac{1 - 0,4}{0,4} = \log \frac{0,6}{0,4} = 0,176$$

# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos,  $A$ ,  $B$  e  $C$ . Consulta:  $A$   $C$

D1	AAAB
D2	AAC
D3	AA
D4	BB
D5	BC C

$$\text{sim}(d_j, q) \sim \sum_{i=1}^T w_{ij} w_{iq} \left( \log \frac{P(k_i|R)}{1 - P(k_i|R)} + \log \frac{1 - P(k_i|\bar{R})}{P(k_i|\bar{R})} \right)$$

Para o cálculo de  $\text{sim}(d_j, q)$  só precisamos calcular as parcelas referentes aos termos que aparecem em ambos  $d_j$  e  $q$ .

Similaridade entre  $d_1$  e  $q$ :

$$\text{sim}(d_1, q) \sim w_{A1} w_{Aq} \left( \log \frac{P(A|R)}{1 - P(A|R)} + \log \frac{1 - P(A|\bar{R})}{P(A|\bar{R})} \right)$$

$$\text{sim}(d_1, q) \sim 0 + -0,176$$

$$\text{sim}(d_1, q) \sim -0,176$$

$$\log \frac{P(A|R)}{1 - P(A|R)} = 0$$

$$\log \frac{1 - P(A|\bar{R})}{P(A|\bar{R})} = -0,176$$

Calculados no slide anterior

# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos,  $A$ ,  $B$  e  $C$ . Consulta:  $A$   $C$

D1	AAAB
D2	AAC
D3	AA
D4	BB
D5	BC

Para o cálculo de  $sim(d_j, q)$  só precisamos calcular as parcelas referentes aos termos que aparecem em ambos  $d_j$  e  $q$ .

Similaridade entre  $d_2$  e  $q$ :

$$sim(d_2, q) \sim w_{A2}w_{Aq} \left( \log \frac{P(A|R)}{1 - P(A|R)} + \log \frac{1 - P(A|\bar{R})}{P(A|\bar{R})} \right) + w_{C2}w_{Cq} \left( \log \frac{P(C|R)}{1 - P(C|R)} + \log \frac{1 - P(C|\bar{R})}{P(C|\bar{R})} \right)$$

$$sim(d_2, q) \sim (0 + -0,176) + (0 + 0,176)$$

$$\log \frac{P(A|R)}{1 - P(A|R)} = 0$$

$$\log \frac{P(C|R)}{1 - P(C|R)} = 0$$

$$sim(d_2, q) \sim 0$$

$$\log \frac{1 - P(A|\bar{R})}{P(A|\bar{R})} = -0,176$$

$$\log \frac{1 - P(C|\bar{R})}{P(C|\bar{R})} = 0,176$$

# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos,  $A$ ,  $B$  e  $C$ . Consulta:  $A$   $C$

D1	AAAB
D2	AAC
D3	AA
D4	BB
D5	BC

Para o cálculo de  $\text{sim}(d_j, q)$  só precisamos calcular as parcelas referentes aos termos que aparecem em ambos  $d_j$  e  $q$ .

Similaridade entre  $d_3$  e  $q$ :

$$\text{sim}(d_3, q) \sim w_{A3}w_{Aq} \left( \log \frac{P(A|R)}{1 - P(A|R)} + \log \frac{1 - P(A|\bar{R})}{P(A|\bar{R})} \right)$$

$$\text{sim}(d_3, q) \sim 0 + -0,176$$

$$\text{sim}(d_3, q) \sim -0,176$$

$$\log \frac{P(A|R)}{1 - P(A|R)} = 0$$

$$\log \frac{1 - P(A|\bar{R})}{P(A|\bar{R})} = -0,176$$

# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos,  $A$ ,  $B$  e  $C$ . Consulta:  $A\ C$

D1	AAAB
D2	AAC
D3	AA
D4	BB
D5	BC C

Para o cálculo de  $\text{sim}(d_j, q)$  só precisamos calcular as parcelas referentes aos termos que aparecem em ambos  $d_j$  e  $q$ .

Similaridade entre  $d_4$  e  $q$ :

$$\text{sim}(d_4, q) \sim 0$$

# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos, A, B e C. Consulta: A C

D1	AAAB
D2	AAC
D3	AA
D4	BB
D5	BC C

Para o cálculo de  $\text{sim}(d_j, q)$  só precisamos calcular as parcelas referentes aos termos que aparecem em ambos  $d_j$  e  $q$ .

Similaridade entre  $d_5$  e  $q$ :

$$\text{sim}(d_5, q) \sim +w_{C5}w_{Cq} \left( \frac{P(C|R)}{1 - P(C|R)} + \log \frac{1 - P(C|\bar{R})}{P(C|\bar{R})} \right)$$

$$\text{sim}(d_5, q) \sim 0 + 0,176$$

$$\text{sim}(d_5, q) \sim 0,176$$

$$\log \frac{P(C|R)}{1 - P(C|R)} = 0$$

$$\log \frac{1 - P(C|\bar{R})}{P(C|\bar{R})} = 0,176$$



# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos, *A*, *B* e *C*. Consulta: *A C*

D1	A A A B
D2	A A C
D3	A A
D4	B B
D5	B C C

Assim, temos as seguintes similaridades

Doc	Sim
D1	-0,176
D2	0
D3	-0,176
D4	0
D5	0,176

A ordem do ranqueamento fica: D5, D2, D4, D1, D3

# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos, A, B e C. Consulta: A C

D1	AAAB
D2	AAC
D3	AA
D4	BB
D5	BC C

A ordem do ranqueamento fica: D5, D2, D4, D1, D3

Atualizamos então as probabilidades  $P(k_i | R)$  e  $P(k_i | \bar{R})$  segundo as fórmulas:

$$P(k_i | R) = \frac{V_i + 0,5}{V + 1} \quad P(k_i | \bar{R}) = \frac{n_i - V_i + 0,5}{N - V + 1}$$

Onde  $V$  é o nº de docs inicialmente recuperados (podem ser só os primeiros do ranking),  $V_i$  é o nº de docs inicialmente recuperados que contém o termo  $k_i$ ,  $N$  é o nº total de docs e  $n_i$  é o nº total de docs com o termo  $k_i$ .

No nosso exemplo, usaremos apenas os 3 primeiros do ranking para atualizar as probabilidades.

# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos, A, B e C. Consulta: A C

D1	AAAB
D2	AAC
D3	AA
D4	BB
D5	BC

$$P(k_i|R) = \frac{V_i + 0,5}{V + 1} \quad P(k_i|\bar{R}) = \frac{n_i - V_i + 0,5}{N - V + 1}$$

Assim,  $V = 3$  (D5, D2 e D4),  $V_A = 1$ ,  $V_B = 2$  e  $V_C = 2$

$$P(A|R) = \frac{1 + 0,5}{3 + 1} = \frac{1,5}{4} = 0,375$$

$$P(A|\bar{R}) = \frac{3 - 1 + 0,5}{5 - 3 + 1} = \frac{2,5}{3} = 0,833$$

$$P(C|R) = \frac{2 + 0,5}{3 + 1} = \frac{2,5}{4} = 0,625$$

$$P(C|\bar{R}) = \frac{2 - 2 + 0,5}{5 - 3 + 1} = \frac{0,5}{3} = 0,167$$

Recalculamos então as similaridades de todos os documentos

# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos, A, B e C. Consulta: A C

D1	AAAB
D2	AAC
D3	AA
D4	BB
D5	BC C

$$P(A|R) = 0,375$$

$$P(C|R) = 0,625$$

$$P(A|\bar{R}) = 0,833$$

$$P(C|\bar{R}) = 0,167$$

$$sim(d_j, q) \sim \sum_{i=1}^T w_{ij} w_{iq} \left( \log \frac{P(k_i|R)}{1 - P(k_i|R)} + \log \frac{1 - P(k_i|\bar{R})}{P(k_i|\bar{R})} \right)$$

Primeiro, calculamos as parcelas referentes aos logaritmos. Como a consulta só é composta por A e C, só precisamos calcular as parcelas referentes a estes dois termos.

$$\log \frac{P(A|R)}{1 - P(A|R)} = \log \frac{0,375}{1 - 0,375} = \log 0,6 = -0,222$$

$$\log \frac{P(C|R)}{1 - P(C|R)} = \log \frac{0,625}{1 - 0,625} = \log 1,667 = 0,222$$

$$\log \frac{1 - P(A|\bar{R})}{P(A|\bar{R})} = \log \frac{1 - 0,833}{0,833} = \log 0,200 = -0,699$$

$$\log \frac{1 - P(C|\bar{R})}{P(C|\bar{R})} = \log \frac{1 - 0,167}{0,167} = \log 4,988 = 0,698$$

# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos, A, B e C. Consulta: A C

D1	AAAB
D2	AAC
D3	AA
D4	BB
D5	BC C

Para o cálculo de  $sim(d_j, q)$  só precisamos calcular as parcelas referentes aos termos que aparecem em ambos  $d_j$  e  $q$ .

Similaridade entre  $d_1$  e  $q$ :

$$sim(d_1, q) \sim w_{A1}w_{Aq} \left( \log \frac{P(A|R)}{1 - P(A|R)} + \log \frac{1 - P(A|\bar{R})}{P(A|\bar{R})} \right)$$

$$sim(d_1, q) \sim -0,222 + -0,698$$

$$sim(d_1, q) \sim -0,920$$

$$\log \frac{P(A|R)}{1 - P(A|R)} = -0,222$$

$$\log \frac{1 - P(A|\bar{R})}{P(A|\bar{R})} = -0,698$$

# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos, A, B e C. Consulta: A C

D1	AAAB
D2	AAC
D3	AA
D4	BB
D5	BC

Para o cálculo de  $sim(d_j, q)$  só precisamos calcular as parcelas referentes aos termos que aparecem em ambos  $d_j$  e  $q$ .

Similaridade entre  $d_2$  e  $q$ :

$$sim(d_2, q) \sim w_{A2}w_{Aq} \left( \log \frac{P(A|R)}{1 - P(A|R)} + \log \frac{1 - P(A|\bar{R})}{P(A|\bar{R})} \right) + w_{C2}w_{Cq} \left( \log \frac{P(C|R)}{1 - P(C|R)} + \log \frac{1 - P(C|\bar{R})}{P(C|\bar{R})} \right)$$

$$sim(d_2, q) \sim -0,222 + -0,698 + 0,222 + 0,698 \quad \boxed{\log \frac{P(A|R)}{1 - P(A|R)} = -0,222} \quad \boxed{\log \frac{P(C|R)}{1 - P(C|R)} = 0,222}$$

$$sim(d_2, q) \sim 0$$

$$\boxed{\log \frac{1 - P(A|\bar{R})}{P(A|\bar{R})} = -0,698} \quad \boxed{\log \frac{1 - P(C|\bar{R})}{P(C|\bar{R})} = 0,698}$$

# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos, A, B e C. Consulta: A C

D1	AAAB
D2	AAC
D3	AA
D4	BB
D5	BC C

Para o cálculo de  $sim(d_j, q)$  só precisamos calcular as parcelas referentes aos termos que aparecem em ambos  $d_j$  e  $q$ .

Similaridade entre  $d_3$  e  $q$ :

$$sim(d_3, q) \sim w_{A3}w_{Aq} \left( \log \frac{P(A|R)}{1 - P(A|R)} + \log \frac{1 - P(A|\bar{R})}{P(A|\bar{R})} \right)$$

$$sim(d_3, q) \sim -0,222 + -0,698$$

$$sim(d_3, q) \sim -0,920$$

$$\log \frac{P(A|R)}{1 - P(A|R)} = -0,222$$

$$\log \frac{1 - P(A|\bar{R})}{P(A|\bar{R})} = -0,698$$

# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos, A, B e C. Consulta: A C

D1	AAAB
D2	AAC
D3	AA
D4	BB
D5	BC

Para o cálculo de  $\text{sim}(d_j, q)$  só precisamos calcular as parcelas referentes aos termos que aparecem em ambos  $d_j$  e  $q$ .

Similaridade entre  $d_4$  e  $q$ :

$$\text{sim}(d_4, q) \sim 0$$



# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos, A, B e C. Consulta: A C

D1	AAAB
D2	AAC
D3	AA
D4	BB
D5	BC C

Para o cálculo de  $sim(d_j, q)$  só precisamos calcular as parcelas referentes aos termos que aparecem em ambos  $d_j$  e  $q$ .

Similaridade entre  $d_5$  e  $q$ :

$$sim(d_5, q) \sim +w_{C5}w_{Cq} \left( \frac{P(C|R)}{1 - P(C|R)} + \log \frac{1 - P(C|\bar{R})}{P(C|\bar{R})} \right)$$

$$sim(d_5, q) \sim 0,222 + 0,698$$

$$sim(d_5, q) \sim 0,920$$

$$\log \frac{P(C|R)}{1 - P(C|R)} = 0,222$$

$$\log \frac{1 - P(C|\bar{R})}{P(C|\bar{R})} = 0,698$$

# Modelo Probabilístico - Exemplo

- Base de 5 documentos e três termos, A, B e C. Consulta: A C

D1	AAAB
D2	AAC
D3	AA
D4	BB
D5	BC C

Assim, temos as novas seguintes similaridades:

Doc	Sim
D1	-0,920
D2	0
D3	-0,920
D4	0
D5	0,920

A ordem do novo ranqueamento fica: D5, D2, D4, D1, D3;

Com nosso exemplo de brinquedo, as similaridades ficam “estranhas”, mas em uma base realística, o modelo se comporta melhor;

O processo poderia ser repetido por mais iterações, mas vamos parar por aqui.

# Vantagens do Modelo Probabilístico

- Documentos ordenados em ordem decrescente de probabilidade de relevância.
  - No entanto, essa probabilidade pode ser incorretamente estimada e depende de fatores externos.
- Refinamento iterativo pode captar características pessoais do usuário.
  - Todavia, na prática, o modelo é implementado sem a realimentação do usuário.

# Desvantagens do Modelo Probabilístico

- Necessidade de “adivinhar” valores iniciais para  $P(k_i|R)$  e  $P(k_i|\bar{R})$
- Não leva em conta ponderação de termos, em especial a frequência dos termos em um documento (TF);
- Falta de normalização pelo tamanho do documento.
- Refinamento iterativo pode produzir resultados ruins se for mau influenciado pelo resultado da primeira iteração.