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
... continuando

Aula Anterior!

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 sim(dj,q) entre o documento dj e a consulta q é definida por:

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

Ajuste para (R = ri = 0)

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

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

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

Modelo Probabilístico – procedimento automático

Modelo probabilístico – Ausência de informação de relevância

- Equação básica para a computação do ranking quando nenhuma informação de relevância é fornecida.

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

- **Se a informação de relevância for considerada, então é necessária a intervenção humana.**

Modelo probabilístico – Ausência de informação de relevância

- Croft e Harper (1979) propuseram evitar a intervenção humana utilizando procedimentos automáticos para refinar as estimativas do parâmetros R e ri.

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

- Propuseram duas suposições para estimar piR e qiR.
  - p – conjunto relevantes;
  - q – conjunto não relevantes;

Modelo probabilístico – Ausência de informação de relevância

- 1. Supor que pir é constante para todos os termos de indexação ki, igual a 0.5;
- 2. Supor que a distribuição dos termos de indexação entre os documentos não relevantes pode ser aproximada pela distribuição dos termos de indexação entre todos os documentos da coleção;
  - Temos que:

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

$$p_{iR} = 0{,}5; \qquad q_{iR} = \frac{n_i}{N}$$

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

Modelo probabilístico – Ausência de informação de relevância

- A partir da estimativa inicial, podemos recuperar documentos que **contenham os termos da consulta** e também fornecer uma ordenação probabilística para eles;
  - Se a consulta for "to do", conseguiremos ter uma ideia sobre os documentos da coleção que possuem os termos "to" e "do".

- Depois podemos tentar melhorar o ranking inicial.

Ausência de informação de relevância Melhorando o ranking inicial

- Seja:
  - D ou V =subconjunto dos documentos inicialmente recuperados e ranqueados pela nova equação (pode-se escolher um limiar, por exemplo, os M documentos melhor ranqueados);
  - Di ou Vi= subconjunto de D composto pelos documentos de D que contêm o termo de indexação ki;
- Podemos melhorar piR e qiR fazendo as seguintes suposições:
  - Aproximar piR pela distribuição de termos de indexação ki entre os documentos recuperados;
  - Aproximar qiR considerando que todos os documentos não recuperados são não relevantes.

Ausência de informação de relevância Melhorando o ranking inicial

- Ficamos com:

$$piR = \frac{D_i}{D} \qquad qiR = \frac{n_i - Di}{N - D}$$

  - Esse processo agora é repetido recursivamente;

- Espera-se que as estimativas para as probabilidades piR e qiR melhorem sem a intervenção humana.

Ajustes nas fórmulas

- Ajuste 1 (problemas para valores pequenos de D e Di):

$$piR = \frac{D_i + 0.5}{D + 1} \qquad qiR = \frac{n_i - Di + 0.5}{N - D + 1}$$

- Ajuste 2 (ao invés de 0,5 tomar a fração ni/N):

$$piR = \frac{D_i + \frac{n_i}{N}}{D + 1} \qquad qiR = \frac{n_i - Di + \frac{n_i}{N}}{N - D + 1}$$

Exemplo

- Necessidade de informação: "Que jogador foi o artilheiro do Brasil na Copa de 1994? Quantos gols ele marcou?"
  - Possível consulta: artilheiro ^brasil ^1994 ^ gols.

- Informações relevantes:
  - Número de documentos da coleção: 20
  - Documentos que aparecem o termos:
    - artilheiro: d1, d3, d6, d7, d11, d15, d18
    - brasil: d1, d3, d7
    - 1994: d1, d3, d6, d7, d9, d15, d19
    - gols: d1, d3, d6, d7, d11, d15, d16

Exemplo

Termo de índice	n_i	$P(k_i \mid R)$	$P(k_i \mid \overline{R}) = \frac{n_i}{N}$	$\log \frac{P(k_i \mid R)}{1 - P(k_i \mid R)} + \log \frac{1 - P(k_i \mid \overline{R})}{P(k_i \mid \overline{R})}$
artilheiro	7	0,500	0,350	0,893
brasil	3	0,500	0,150	2,502
1994	6	0,500	0,300	1,222
gols	6	0,500	0,300	1,222

log 0.5/(1-0.5) + log (1-0.35)/0.35

log 0.5/(1-0.5) + log (1-0.150)/0.150

log 0.5/(1-0.5) + log (1-0.300)/0.300

log 0.5/(1-0.5) + log (1-0.300)/0.300

Exemplo

Resultados obtidos na 1ª interação do Modelo

Doc.	Termos de índice	$sim(d_j, q)$
d_1	artilheiro, brasil, 1994, gols	5,840
d_3	artilheiro, brasil, 1994, gols	5,840
d_7	artilheiro, brasil, 1994, gols	5,840
d_{15}	artilheiro, 1994, gols	3,338
d_{11}	artilheiro, gols	2,115
d_6	artilheiro	0,893
d_9	1994	1,222
d_{16}	gols	1,222
d_{18}	artilheiro	0,893
d_{19}	1994	1,222

Exemplo

Probabilidades e cálculos relacionados, na 2ª interação do Modelo, para cada termo da consulta, considerando que o subconjunto *V* é composto pelos 5 primeiros documentos do conjunto inicial recuperado e utilizando a fórmula do segundo ajuste:

Termo de índice	n_i	V	V_i	$P(k_i \mid R)$	$P(k_i \mid \overline{R})$	$\log\frac{P(k_i \mid R)}{1-P(k_i \mid R)} + \log\frac{1-P(k_i \mid \overline{R})}{P(k_i \mid \overline{R})}$
artilheiro	7	5	5	0,891	0,212	4,931
brasil	3	5	3	0,525	0,038	4,806
1994	6	5	4	0,717	0,200	3,341
gols	6	5	5	0,883	0,138	5,559

$$P(k_i \mid R) = \frac{V_i + \frac{n_i}{N}}{V+1} \qquad P(k_i \mid \overline{R}) = \frac{n_i - V_i + \frac{n_i}{V}}{N - V + 1}$$

=(5 + 7/20)/(5+1) = 0.891

=(7-5+7/5)/(20-5+1) = 0.212

Exemplo

Resultados obtidos na 2ª interação do Modelo

Doc.	Termos de índice	$sim(d_j, q)$
d_1	artilheiro, brasil, 1994, gols	18,637
d_3	artilheiro, brasil, 1994, gols	18,637
d_7	artilheiro, brasil, 1994, gols	18,637
d_{15}	artilheiro, 1994, gols	13,831
d_{11}	artilheiro, gols	10,490
d_6	artilheiro	4,931
d_{16}	gols	5,559
d_{18}	artilheiro	4,931
d_9	1994	3,341
d_{19}	1994	3,341

Exemplo

- Pelos resultados obtidos, observa-se que:
  - da 1ª para a 2ª interação, só houve diferença no *ranking* dos 5 últimos documentos;
  - em uma possível 3ª interação, como o subconjunto *V* seria o mesmo da 2ª, não haveria diferença no *ranking* gerado (ou seja, o resultado da aplicação do Modelo estabilizou-se na 2ª interação);
  - A ideia do Modelo Probabilístico é fazer com que o *ranking* dos documentos melhore, a cada interação, a partir dos documentos do conjunto *V*, até um ponto de estabilização.

Atividade para entregar – 12/07

- Aplique o modelo probabilístico no exercício 1 da lista II.


Nunca deixe
de acreditar,
seja *Otimista!* Pois
de um dia para outro
a vida se transforma.
MUNDO DAS MENSAGENS