Universidade Federal de Santa Maria
1960
Frederico Westphalen

Modelo Probabilístico

Tópicos Especiais em Recuperação de Informações

Profª. Solange Pertile

21/09/15

Fontes:
Prof. Viviane Moreira (UFRGS)
Prof. Jairo de Souza (UFJF)

Relembrando Modelo Booleano...

- Consultas usando operadores booleanos
- Documentos casam ou não casam
- Boa para usuários especialistas com um entendimento preciso das suas necessidades e da coleção
- Boa para aplicações: aplicações podem facilmente processar milhares de resultados.
- Ruim para a maioria dos usuários
- Incapazes de escrever consultas Booleanas (ou são, mas acham muito trabalhoso)
- Não querem procurar em milhares de resultados
    - Principalmente quando se trata de busca na Web

Relembrando Modelo Vetorial...

- ✓Vector space model (VSM)
- ✓Associa peso aos termos de indexação.
- ✓Atribui escores aos documentos.
- ✓Possibilita *ranking* dos resultados da consulta.

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.

Universidade Federal de Santa Maria
1960
Frederico Westphalen

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 precise;
- **IMPORTANTE**: é necessário estimar, no início, a descrição do **conjunto resposta ideal.**

Ranqueamento

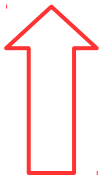
- Como calcular a medida de similaridade? Como criar uma função que irá ranquear os resultados?
  - *d* é um documento da coleção
  - *R* representa que o documento é relevante
  - *NR* representa que o documento não é relevante
  - Os documentos serão ranqueados de acordo com a estimativa de probabilidade de que eles sejam relevantes à consulta

$$P(R \mid d,q)$$

$$\text{d é relevante se } P(R \mid d, q) > P(\text{N}R \mid d, q)$$

Ranqueamento

- De maneira mais específica, precisamos saber como as estatísticas que podemos calcular influenciam na relevância do documento



tf, df, tamanho do doc,etc.

- **Como computar estas probabilidades??**

Ranqueamento

- A medida usada para ranquear os documentos é o Retrieval Status Value (RSV)/Valor do Estado da Recuperação.

$$RSV = \log \prod_{x_i = q_i = 1} \frac{p_i(1-u_i)}{u_i(1-p_i)} = \sum_{x_i = q_i = 1} \log \frac{p_i(1-u_i)}{u_i(1-p_i)}$$

Probabilidade do termo ocorrer em um documento relevante

Probabilidade do termo ocorrer em um documento não-relevante

$$c_i = \log \frac{p_i(1-u_i)}{u_i(1-p_i)} = \log\left(\frac{p_i}{1-p_i}\right) + \log\left(\frac{1-u_i}{u_i}\right)$$

Relevantes

Não- Relevantes

Para cada termo *i*, como calcular c_i?

Independência binária

- Este é o modelo tradicionalmente usado com o Princıpio de ranking probabilıstico.
  - **Binary** quer dizer booleano, i.e. documentos são representados como vetores de termos cujas incidências são representadas por 0 ou 1. É possível que mais de um documento tenha portanto o mesmo vetor
  - **Independence** significa que os termos ocorrem nos documentos de maneira independente; i.e. o fato de um termo aparecer em um documento não tem nenhum impacto sobre a ocorrência de outros termos no documento.

Estimando C_i – Peso dos Termos

- **Para cada termo *i* consultar a tabela de contingência:**

Docs	Relevante	Não relevante	Total
Termo presente x_i=1	***s***	***df_i-s***	***df_i***
Termo ausente x_i=0	***S-s***	***N-df_i-S+s***	***N-df_i***
Total	***S***	***N-S***	***N***

- **Assumindo que: *p_i=s/S e u_i= (df_i-s)/(N-S)***

$$c_i = \log \frac{s/(S-s)}{(df_i - s)/(N - df_i - S + s)}$$

Estimando C_i – Peso dos Termos

- **Para cada termo *i* consultar a tabela de contingência:**

Docs	Relevante	Não relevante	Total
Termo presente x_i=1	***s***	***df_i-s***	***df_i***
Termo ausente x_i=0	***S-s***	***N-df_i-S+s***	***N-df_i***
Total	***S***	***N-S***	***N***

- **Assumindo que: *p_i=s/S e u_i= (df_i-s)/(N-S)***

Soma-se 0.5 a cada termo na presença de incerteza

$$c_i = \log \frac{(s+0.5)/(S-s+0.5)}{(df_i-s+0.5)/(N-df_i-S+s+0.5)}$$

Estimando C_i – Peso dos Termos

Exemplo:

relevantes

Termos \ Docs	Tom e Jerry	Super Mouse	Garfield	Scooby Doo	PiuPiu e Frajola	Mônica
Cachorro	1	0	1	1	0	1
Casa	1	0	1	0	1	1
Gato	1	0	1	0	1	0
Menino	0	0	0	1	0	1
Passarinho	0	0	0	0	1	0
Rato	1	1	0	0	0	0

Consulta: cachorro gato rato

$$c_i = \log \frac{(s+0.5)/(S-s+0.5)}{(df_i - s+0.5)/(N-df_i-S+s+0.5)}$$

Termo	c_i
Cachorro	*- 0.3679*
Gato	*1.066*
Rato	*-0.6989*

Estimando C_i – Peso dos Termos

Exemplo:

relevantes

Termos \ Docs	Tom e Jerry	Super Mouse	Garfield	Scooby Doo	PiuPiu e Frajola	Mônica
Cachorro	1	0	1	1	0	1
Casa	1	0	1	0	1	1
Gato	1	0	1	0	1	0
Menino	0	0	0	1	0	1
Passarinho	0	0	0	0	1	0
Rato	1	1	0	0	0	0

Consulta: cachorro gato rato

$$c_i = \log \frac{(s+0.5)/(S-s+0.5)}{(df_i - s+0.5)/(N-df_i-S+s+0.5)}$$

$$c_{cachorro} = \log \frac{(1+0.5)/(2-1+0.5)}{(4-1+0.5)/(6-4-2+1+0.5)} = \log \frac{1}{2,33} = -0,3679$$

$$c_{gato} = \log \frac{(2+0.5)/(2-2+0.5)}{(3-2+0.5)/(6-3-2+2+0.5)} = \log \frac{5}{0,4285} = 1,0669$$

$$c_{rato} = \log \frac{(0+0.5)/(2-0+0.5)}{(2-0+0.5)/(6-2-2+0+0.5)} = \log \frac{0,2}{1} = -0.6989$$

Termo	c_i
Cachorro	*- 0.3679*
Gato	*1.066*
Rato	*-0.6989*

Estimando C_i – Peso dos Termos

Exemplo:

relevantes

Termos \ Docs	Tom e Jerry	Super Mouse	Garfield	Scooby Doo	PiuPiu e Frajola	Mônica
Cachorro	1	0	1	1	0	1
Casa	1	0	1	0	1	1
Gato	1	0	1	0	1	0
Menino	0	0	0	1	0	1
Passarinho	0	0	0	0	1	0
Rato	1	1	0	0	0	0

Consulta: cachorro gato rato

$$c_i = \log \frac{s/(S-s)}{(df_i - s)/(N - df_i - S + s)}$$

$$c_{cachorro} = \log \frac{(1+0.5)/(2-1+0.5)}{(4-1+0.5)/(6-4-2+1+0.5)} = \log \frac{1}{2,33} = -0,3679$$

$$c_{gato} = \log \frac{(2+0.5)/(2-2+0.5)}{(3-2+0.5)/(6-3-2+2+0.5)} = \log \frac{5}{0,4285} = 1,0669$$

$$c_{rato} = \log \frac{(0+0.5)/(2-0+0.5)}{(2-0+0.5)/(6-2-2+0+0.5)} = \log \frac{0,2}{1} = -0.6989$$

Termo	c_i
Cachorro	*- 0.3679*
Gato	*1.066*
Rato	*-0.6989*

Ranque dos documentos???

Ranqueamento – Peso dos documentos

Termo	c_i
Cachorro	*- 0.3679*
Gato	*1.0669*
Rato	*-0.6989*

$$RSV = \sum_{x_i = q_i = 1} c_i$$

Doc	RSV	Rank
Tom e Jerry	- *0.3679*(cachorro) + 1.0669(gato) - 0.6989(rato) = 0.0001	
Super Mouse	- 0.6989(rato) = -0.6989	
Garfield	- *0.3679*(cachorro) + 1.0669(gato) = 0.6990	
Scooby Doo	- 0.*3679*(cachorro) = - 0.3679	
PiuPiu e Frajola	1.0669(gato) = 1.0669	
Mônica	- 0.3679(cachorro) = - 0.3679	

Ranqueamento – Peso dos documentos

Termo	c_i
Cachorro	*- 0.3679*
Gato	*1.0669*
Rato	*-0.6989*

$$RSV = \sum_{x_i = q_i = 1} c_i$$

Doc	RSV	Rank
Tom e Jerry	- *0.3679*(cachorro) + 1.0669(gato) - 0.6989(rato) = 0.0001	3°
Super Mouse	- 0.6989(rato) = -0.6989	6°
Garfield	- *0.3679*(cachorro) + 1.0669(gato) = 0.6990	2°
Scooby Doo	- 0.*3679*(cachorro) = - 0.3679	5°
PiuPiu e Frajola	1.0669(gato) = 1.0669	1°
Mônica	- 0.3679(cachorro) = - 0.3679	4°

Questões importantes

- Como saber quantos documentos relevantes existem e quais são eles?
  - Pode-se fazer uma estimativa inicial
  - Exemplo: $p_i = 0.5$ e $u_i = df_t / N$
  - Ou podemos fazer uma recuperação inicial de acordo com o modelo vetorial e supor que os top *k* documentos são relevantes
  - O processo pode ser iterativo (com ou sem a interferência do usuário).
  - Neste caso, a cada iteração as estimativas ficam mais precisas.

Vantagens x Desvantagens

- **Vantagens**
  - Os documentos são ranqueados de acordo com sua probabilidade de serem relevantes, com base na informação disponível ao sistema.
- **Desvantagens**
  - Relevância de um documento é afetada por diversos fatores externos, não somente na informação disponível ao sistema;
  - Necessidade de estimar a separação inicial dos documentos em conjuntos relevantes e não relevantes;
  - Não leva em consideração a frequência na qual um termo de indexação ocorre em um documento;
  - Falta de normalização pelo tamanho dos documentos.

Comparação entre os modelo

- O modelo booleano é considerado como o mais fraco
- Há controvérsias quanto ao melhor modelo
- Croft realizou experimentos e concluiu que o modelo probabilístico é melhor
- Logo após, Salton & Buckley realizaram experimentos e concluíram que o modelo vetorial *deve* ser melhor para coleções gerais.