GSI024 - Organização e recuperação de informação

Prof. Dr. Rodrigo Sanches Miani (FACOM/UFU)

Última atualização - Junho/2022

Modelo Probabilístico

Agenda

“Modelo vetorial”

Ideia básica

Definição

Estimativa das probabilidades e exemplo

Vantagens x Desvantages

Aula passada

Modelo vetorial – Grau de similaridade

Grau de similaridade entre um determinado documento e uma consulta, no modelo vetorial, é dado por:

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

onde o numerador representa o produto interno entre os dois vetores e o denominador representa o produto da norma dos dois vetores.

Modelo vetorial – Grau de similaridade

- O grau de similaridade (sim(d_j,q)) varia entre 0 e 1;
  - Ao invés de adotar um critério binário, os documentos são ordenados com base no grau de similaridade;
  - Assim, um documento pode ser recuperado, mesmo que ele satisfaça a consulta apenas parcialmente.

- Quanto mais próximo de 1, mais bem ranqueado será o documento d_j com relação a consulta q;
  - Valores próximos de 1 para cos(θ) representam maior “proporcionalidade” entre os vetores d_j e q.

Modelo vetorial – Exemplo

Doc	Computação do escore	Escore
d_1	$\frac{1\times 3+0,415\times 0,830}{5.068}$	0,660
d_2	$\frac{1\times 2+0,415\times 0}{4,899}$	0,408
d_3	$\frac{1\times 0+0,415\times 1,073}{3,762}$	0,118
d_4	$\frac{1\times 0+0,415\times 1,073}{7,738}$	0,058

Ideia básica

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 precise;
- 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* (*odds ratio*);

- Modo de quantificar o quão forte a presença (ou ausência) da propriedade A está associada a presença (ou ausência) da propriedade B;
- Relação: documento dj ser relevante a q e o documento dj não ser relevante a q;
- O = proporção de sucessos / proporção de falhas;
- Razão de possibilidades = 1 indica que a condição ou evento sob estudo é igualmente provável de ocorrer nos dois grupos;

- Grau de similaridade = chance de relevância ou razão de possibilidade.

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;

- **Pergunta:** O quão forte é a relação entre homens beberem vinho e mulheres beberem vinho?

Ideia fundamental - Exemplo

- A razão de chances ou razão de possibilidades é definida como a razão entre a chance de um evento ocorrer em um grupo e a chance de ocorrer em outro grupo;

- Chance ou possibilidade é a probabilidade de ocorrência deste evento dividida pela probabilidade da não ocorrência do mesmo evento;

- Razão de chances: $$\frac{p/(1-p)}{q/(1-q)} = \frac{p(1-q)}{q(1-p)}.$$

Ideia fundamental - Exemplo

- A chance (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;

- 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.

Definição

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

$$\vec{d_j} = (w_{1,j},\ w_{2,j},\ \ldots,\ w_{t,j})$$

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 $\overline{R}$ o complemento de R (o conjunto de documentos não relevantes). A similaridade sim(d_j,q) entre o documento d_j e a consulta q é definida por:

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

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

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

- 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}

Estimativa das probabilidades

Estimar as probabilidades do conjunto de documentos relevantes

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 relevantes para a consulta q (na opinião do usuário) e r_i o número de documentos relevantes que contêm o termo k_i.

- Caso 1 – Documentos que contêm k_i (relevantes e não relevantes);
- Caso 2 – Documentos que não contêm k_i (relevantes e não relevantes);
- Caso 3 – Todos os documentos.

Estimar as probabilidades do conjunto de documentos relevantes

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 relevantes para a consulta q (na opinião do usuário) e r_i o número de documentos relevantes que contêm o termo k_i.

Caso	Relevantes	Não relevantes	Total
Documentos que contêm ki			
Documentos que não contêm ki			
Todos os documentos			

Estimar as probabilidades do conjunto de documentos relevantes

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 relevantes para a consulta q (na opinião do usuário) e r_i o número de documentos relevantes que contêm o termo k_i.

Caso	Relevantes	Não relevantes	Total
Documentos que contêm ki	r_i		n_i
Documentos que não contêm ki			
Todos os documentos	R		N

Estimar as probabilidades do conjunto de documentos relevantes

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 relevantes para a consulta q (na opinião do usuário) e r_i o número de documentos relevantes que contêm o termo k_i.

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

Estimar as probabilidades do 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:

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

- 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.

Estimar as probabilidades do conjunto de documentos relevantes

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

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

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:

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

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

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

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):

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

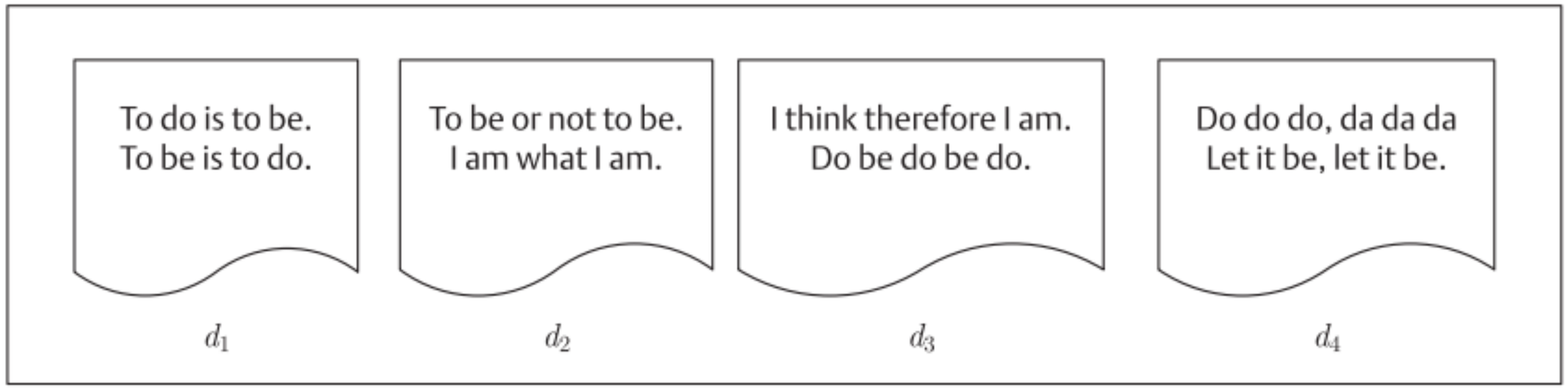
Dessa forma, um termo que ocorre em todos os documentos (ni = N) produz um peso igual a zero (log(1)=0) e não existem mais pesos negativos.

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

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

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

Consulta q = “to do”.



Alternativa para estimar R e r_i

As equações anteriores consideram que R=ri=0. Uma alternativa para estimar R e ri mais cuidadosamente é:

1. Fazer a busca inicial utilizando a equação com R=ri=0;
2. Selecionar os 10-20 documentos mais bem ranqueados;
3. Inspecionar os documentos para obter novas estimativas para R e ri;
4. Remover esses 10-20 documentos da coleção;
5. Reprocessar a consulta com as novas estimativas.

Vantagens x Desvantagens

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:
   - Experimentos realizados por Croft indicam que o modelo probabilístico fornece melhor qualidade de recuperação;
   - 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.

Comentários

No decorrer da aula vimos…

- Como construir um modelo de RI usando fundamentos de probabilidade;

- A partir de uma consulta do usuário, existe um conjunto de documentos que contém exatamente os documentos relevantes (**resposta ideal**) e nenhum outro.

No decorrer da aula vimos…

- Problema:
  - As propriedades para descrever um conjunto relevante 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.

No decorrer da aula vimos…

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

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

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

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

Próximas aulas

- Pré-processamento de documentos e aula prática;

- Avaliação da recuperação da informação.