# Organização e Recuperação de Informação – GSI521

Prof. Dr. Rodrigo Sanches Miani – FACOM/UFU

# Modelo Vetorial

Organização e Recuperação de Informação (GSI521)

# Motivação

- Discutido nos seguintes trabalhos:
  - K. Spark Jones - 1974;
  - Salton e Yang – 1973;
  - Salton, Wong e Yang – 1975;
- Reconhecimeno de que a recuperação booleana é bastante limitada;
- Propor um método em que casamentos parciais entre consultas e documentos são possíveis (algo entre 0 e 1).

# Ideia básica

- Atribuir pesos não binários aos termos de indexação das consultas e documentos;
- Calcular o grau de similaridade entre cada documento armazenado no sistema e a consulta do usuário;
- Ordenar os documentos recuperados de forma decrescente de acordo com esse grau de similaridade.

# Modelo vetorial x Modelo booleano

Documentos recuperados de forma decrescente (ranqueados) fornecem uma resposta (no sentido da necessidade de informação do usuário) mais precisa do que a resposta fornecida pelo modelo booleano.

# Modelo vetorial - Definição

Para o modelo vetorial, o peso $w_{i,j}$ associado ao par termo-documento ($k_i$, $d_j$) é não negativo e não binário. Os termos de indexação são todos considerados mutuamente independentes e são representados por vetores unitários em um espaço com t dimensões, no qual t é o número de termos de indexação. As representações do documento $d_j$ e da consulta q são vetores com t dimensões dadas por

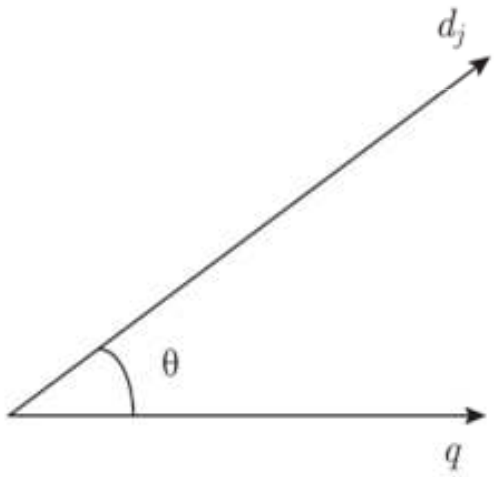
$$\vec{d_j} = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$

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

onde $w_{i,q}$ é o peso associado ao par termo-consulta ($k_i$, q), com $w_{i,q} \geq 0$.

# Modelo vetorial - Definição

- Portanto, um documento $d_j$ e uma consulta de usuário q são representados como vetores com t dimensões:




# Modelo vetorial – Grau de similaridade

- O grau de similaridade do documento $d_j$ em relação à consulta q é dado à partir da correlação entre os vetores $d_j$ e q;

- Um meio de quantificar essa correlação é através do cálculo do cosseno entre os vetores $d_j$ e q:
  - Por que cosseno?

# Modelo vetorial – Grau de similaridade

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

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

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 ranquado será o documento $d_j$ com relação a consulta q;
  - Valores próximos de 1 para cos($\theta$) representam maior "proporcionalidade" entre os vetores $d_j$ e q.

# Modelo vetorial – Grau de similaridade

Os pesos no modelo vetorial são basicamente os pesos TF-IDF:

$$w_{i,q} = (1 + \log f_{i,q}) \times \log\left(\frac{N}{n_i}\right)$$

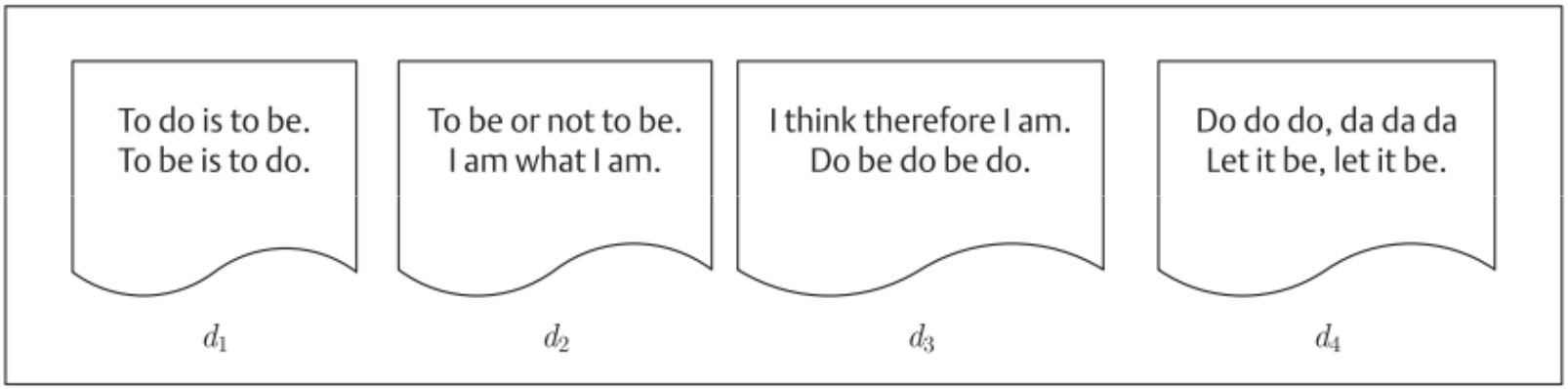
$$w_{i,j} = (1 + \log f_{i,j}) \times \log\left(\frac{N}{n_i}\right)$$

onde $f_{i,q}$ é a frequência do termo ki no texto da consulta q.

*Importante: as equações acima só devem ser aplicadas para valores de frequência de termo maior do que zero. Se a frequência do termo for zero, o respectivo peso também deve ser zero.*

# Modelo vetorial – Exemplo

Considere a coleção de documentos abaixo:




e a consulta q = “to do”. Vamos calcular o grau de similaridade entre cada documento e a consulta.

# Ponderação TF

| # | termo | $f_{i,1}$ | $f_{i,2}$ | $f_{i,3}$ | $f_{i,4}$ | $TF_{i,1}$ | $TF_{i,2}$ | $TF_{i,3}$ | $TF_{i,4}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | to | 4 | 2 | – | – | 3 | 2 | – | – |
| 2 | do | 2 | – | 3 | 3 | 2 | – | 2,585 | 2,585 |
| 3 | is | 2 | – | – | – | 2 | – | – | – |
| 4 | be | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 5 | or | – | 1 | – | – | – | 1 | – | – |
| 6 | not | – | 1 | – | – | – | 1 | – | – |
| 7 | I | – | 2 | 2 | – | – | 2 | 2 | – |
| 8 | am | – | 2 | 1 | – | – | 2 | 1 | – |
| 9 | what | – | 1 | – | – | – | 1 | – | – |
| 10 | think | – | – | 1 | – | – | – | 1 | – |
| 11 | therefore | – | – | 1 | – | – | – | 1 | – |
| 12 | da | – | – | – | 3 | – | – | – | 2,585 |
| 13 | let | – | – | – | 2 | – | – | – | 2 |
| 14 | it | – | – | – | 2 | – | – | – | 2 |
| Tamanho do documento (# palavras) | | 10 | 11 | 10 | 12 | | | | |

# Ponderação IDF

| # | termo | $n_i$ | $IDF_i = \log(N/n_i)$ |
|---|---|---|---|
| 1 | to | 2 | 1 |
| 2 | do | 3 | 0,415 |
| 3 | is | 1 | 2 |
| 4 | be | 4 | 0 |
| 5 | or | 1 | 2 |
| 6 | not | 1 | 2 |
| 7 | I | 2 | 1 |
| 8 | am | 2 | 1 |
| 9 | what | 1 | 2 |
| 10 | think | 1 | 2 |
| 11 | therefore | 1 | 2 |
| 12 | da | 1 | 2 |
| 13 | let | 1 | 2 |
| 14 | it | 1 | 2 |

# Ponderação TF-IDF

| # | termo | $d_1$ | $d_2$ | $d_3$ | $d_4$ |
|---|---|---|---|---|---|
| 1 | to | 3 | 2 | – | – |
| 2 | do | 0,830 | – | 1,073 | 1,073 |
| 3 | is | 4 | – | – | – |
| 4 | be | – | – | – | – |
| 5 | or | – | 2 | – | – |
| 6 | not | – | 2 | – | – |
| 7 | I | – | 2 | 2 | – |
| 8 | am | – | 2 | 1 | – |
| 9 | what | – | 2 | – | – |
| 10 | think | – | – | 2 | – |
| 11 | therefore | – | – | 2 | – |
| 12 | da | – | – | – | 5,170 |
| 13 | let | – | – | – | 4 |
| 14 | it | – | – | – | 4 |
| Tamanho do documento (normas dos vetores) | | 5,068 | 4,899 | 3,762 | 7,738 |

# 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 |

# Modelo vetorial – Exemplo

- O documento $d_1$ é o documento mais bem ranqueado, porque possui todos os termos da consulta;

- Os documentos $d_3$ e $d_4$ contêm apenas o termo da consulta "do" mas o documento $d_4$ recebe um escore menor. Porque?
  - A norma do vetor é maior (normalização pelo tamanho do documento).

# Modelo vetorial x Sistema de RI

- Consulta do usuário
  - Permite o uso de expressões booleanas;
- Recuperação de documentos
  - Documentos são recuperados usando a ponderação de termos TF-IDF;
- Ranqueamento dos documentos
  - Com base no grau de similaridade entre o documento e a consulta (cosseno do ângulo entre os vetores).

# Modelo vetorial – Características

- Apesar de simples, o modelo vetorial consegue bons resultados com coleções genéricas pois utiliza:
  - Esquema de ponderação de termos;
  - Método para normalizar o documento de acordo com o seu tamanho;
- Fornece resultados ranqueados que dificilmente podem ser melhorados sem o uso de expansão de consultas ou de realimentação de relevância;
- Um dos modelos de RI mais populares.

# Modelo vetorial – Vantagens

1. Ponderação de termos melhora a qualidade da recuperação (quando comparado ao modelo booleano);
2. Estratégia de casamento parcial entre a consulta e o documento permite a recuperação de documentos que aproximam as condições da consulta;
3. A normalização pelo tamanho do documento está naturalmente embutida no modelo.

# Modelo vetorial – Desvantagem

1. Termos de indexação são considerados mutuamente independentes

# Comentários

Organização e Recuperação de Informação (GSI521)

# No decorrer da aula vimos...

- Como utilizar a ponderação TF-IDF como base para a construção de um modelo de RI (modelo vetorial);
- Um método para calcular o grau de similaridade entre um documento e uma consulta, baseado no cosseno formado entre os ângulos de tais vetores;

# No decorrer da aula vimos...

- A simplicidade e a possibilidade do casamento parcial entre um documento e uma consulta são características fundamentais do modelo vetorial;

- Modelo vetorial é uma boa estratégia de ranqueamento para coleções genéricas.

# Próximas aulas

- Modelo probabilístico.
- Aulas práticas.