Universidade Federal do Rio de Janeiro (UFRJ)
Departamento de Ciência da Computação (DCC)

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO - UFRJ -

Recuperação da Informação (MAB605)

Modelo Vetorial

Profa. Giseli Rabello Lopes

Roteiro

- Introdução
- Modelo vetorial
  - Normalização
  - Representação
  - Similaridade
- Referências

Modelo Vetorial

- *Vector Space Model (VSM)*
  - Proposto por Gerard Salton no final dos anos 60
  - Propõe ranking dos resultados
    - Ordenado pelo grau de similaridade de cada documento em relação à consulta
    - Possibilita "casamento parcial"
  - Representação (*bag of words*):
    - Documentos e consultas
      - Vetores de termos com associação de peso

Variantes do esquema TF-IDF – Relembrando

- Esquemas recomendados de ponderação TF-IDF [Salton, 1971]

Esquema de ponderação	Pesos para os termos dos docs	Pesos para os termos das consultas
1	$f_{i,j} * \log \frac{N}{n_i}$	$(0.5 + 0.5 \frac{f_{i,q}}{max_i \ f_{i,q}}) * \log \frac{N}{n_i}$
2	$1 + \log f_{i,j}$	$\log(1 + \frac{N}{n_i})$
3	$(1 + \log f_{i,j}) * \log \frac{N}{n_i}$	$(1 + \log f_{i,q}) * \log \frac{N}{n_i}$

Normalização pelo tamanho dos documentos [Baeza-Yates & Ribeiro-Neto, 2013]

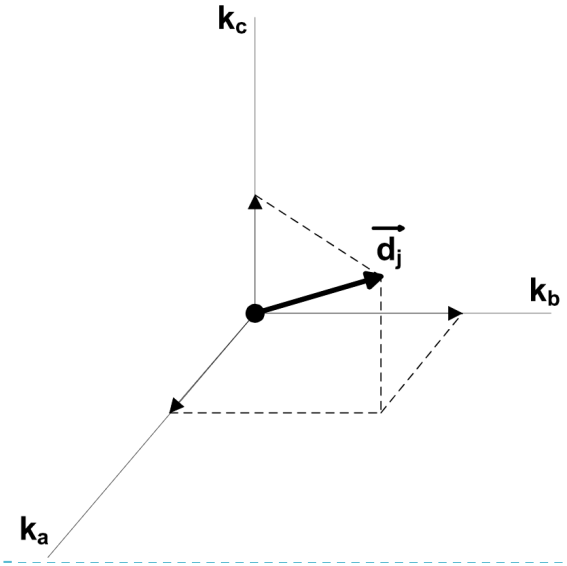
- O tamanho dos documento pode variar bastante
- Isso é um problema porque documentos longos têm mais chance de serem recuperados por uma consulta
- Para compensar esse efeito indesejado, podemos dividir o número de ordem (*rank*) de cada documento pelo seu tamanho
- Esse procedimento consistentemente leva a um ranqueamento melhor, e é chamado **normalização pelo tamanho dos documentos**

Normalização pelo tamanho dos documentos [Baeza-Yates & Ribeiro-Neto, 2013]

- Métodos de normalização pelo tamanho dos documentos dependem da representação adotada para os documentos:
  - **Tamanho em bytes**: considera que cada documento é representado simplesmente como um fluxo (*stream*) de bytes
  - **Número de palavras**: cada documento é representado como uma única *string*, e o tamanho do documento é o número de palavras nele contidas
  - **Norma**: documentos são representados como vetores de termos com pesos associados

Normalização pelo tamanho dos documentos [Baeza-Yates & Ribeiro-Neto, 2013]

- Documentos representados como vetores de termos com pesos associados
  - Cada termo da coleção é associado com um vetor unitário ortonormal $\vec{k}_i$ em um espaço t-dimensional
  - Para cada termo k_i de um documento d_j é associado o componente do vetor de termos $w_{i,j} \times \vec{k}_i$



Normalização pelo tamanho dos documentos [Baeza-Yates & Ribeiro-Neto, 2013]

- A representação de um documento $\vec{d_j}$ é um vetor composto pelos vetores de todos os seus termos

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

- O tamanho do documento é dado pela norma (módulo) desse vetor, que é computada como segue

$$|\vec{d_j}| = \sqrt{\sum_{i}^{t} w_{i,j}^2}$$

Três variantes de tamanhos de documentos para uma coleção de exemplo

To do is to be.
To be is to do.

d_1

To be or not to be.
I am what I am.

d_2

I think therefore I am.
Do be do be do.

d_3

Do do do, da da da.
Let it be, let it be.

d_4

$$w_{i,j} = \begin{cases} (1 + \log f_{i,j}) \times \log \frac{N}{n_i} & \texttt{if } f_{i,j} > 0 \\ 0 & \texttt{otherwise} \end{cases}$$

		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

Relembrando: TF-IDF

	d_1	d_2	d_3	d_4
size in bytes	33	37	41	43
number of words	10	11	10	12
vector norm	5.068	4.899	3.762	7.738

Obs.: Para calcular o tamanho em bytes, consideramos que há um caracter de fim de linha ao final de cada linha e um caracter de fim de arquivo ao final de cada documento. Para calcular a norma, nesse exemplo, é usada a 3ª variação de TF-IDF.

Modelo Vetorial

- Definição formal
  - Documentos representados por vetores
    - t dimensões

$$w_{i,j} \in R \mid w_{i,j} \geq 0$$

$$w_{i,q} \in R \mid w_{i,q} \geq 0$$

q = conjunto de termos

$$sim(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{\left|\vec{d}_j\right| \left|\vec{q}\right|}$$

Produto escalar (interno)

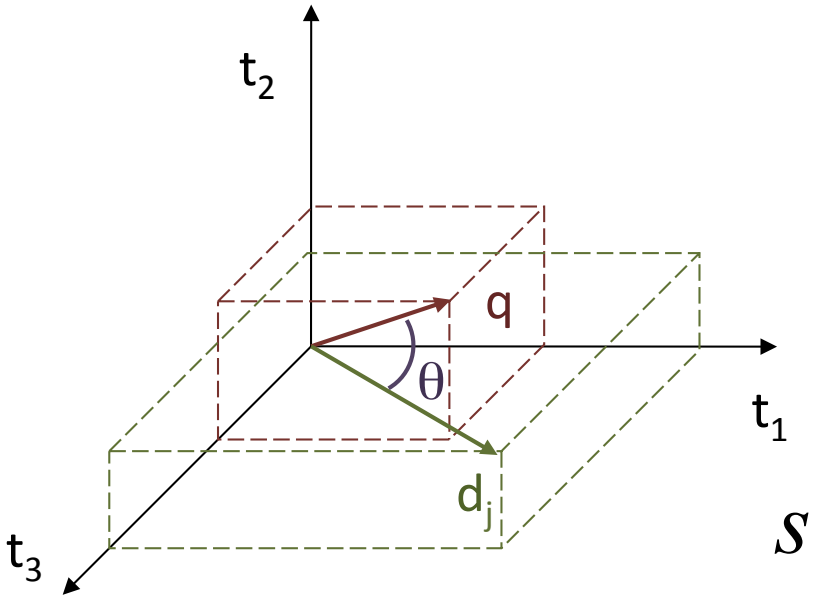
número de termos

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

Multiplicação entre módulos dos vetores (comprimento)

Modelo Vetorial

- Interpretação geométrica



$$sim(d_j, q) = \cos(\theta) = \frac{\vec{d}_j \bullet \vec{q}}{\left|\vec{d}_j\right| \left|\vec{q}\right|}$$

$$sim(d_j, q) = [0,1]$$

Exemplo [Baeza-Yates & Ribeiro-Neto, 2013]

- Ponderação dos pesos (esquema 3 tabela do slide 4)

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

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

  - Se a frequência do termo for zero, o respectivo peso também será zero

- Para calcular o *rank* = $sim(d_j,q)$ do modelo vetorial:
  - Como $|\vec{q}|$ não afeta o ranqueamento (ordenação dos documentos) ele foi desconsiderado no cálculo do exemplo
  - O fator $|\vec{d_j}|$ faz a normalização pelo tamanho do documento (existem outras formas de normalização)

Outro exemplo

[Baeza-Yates & Ribeiro-Neto, 2013]

Relembrando: TF-IDF

		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

To do is to be.
To be is to do.
d_1

To be or not to be.
I am what I am.
d_2

I think therefore I am.
Do be do be do.
d_3

Do do do, da da da.
Let it be, let it be.
d_4

- Consulta: "to do"

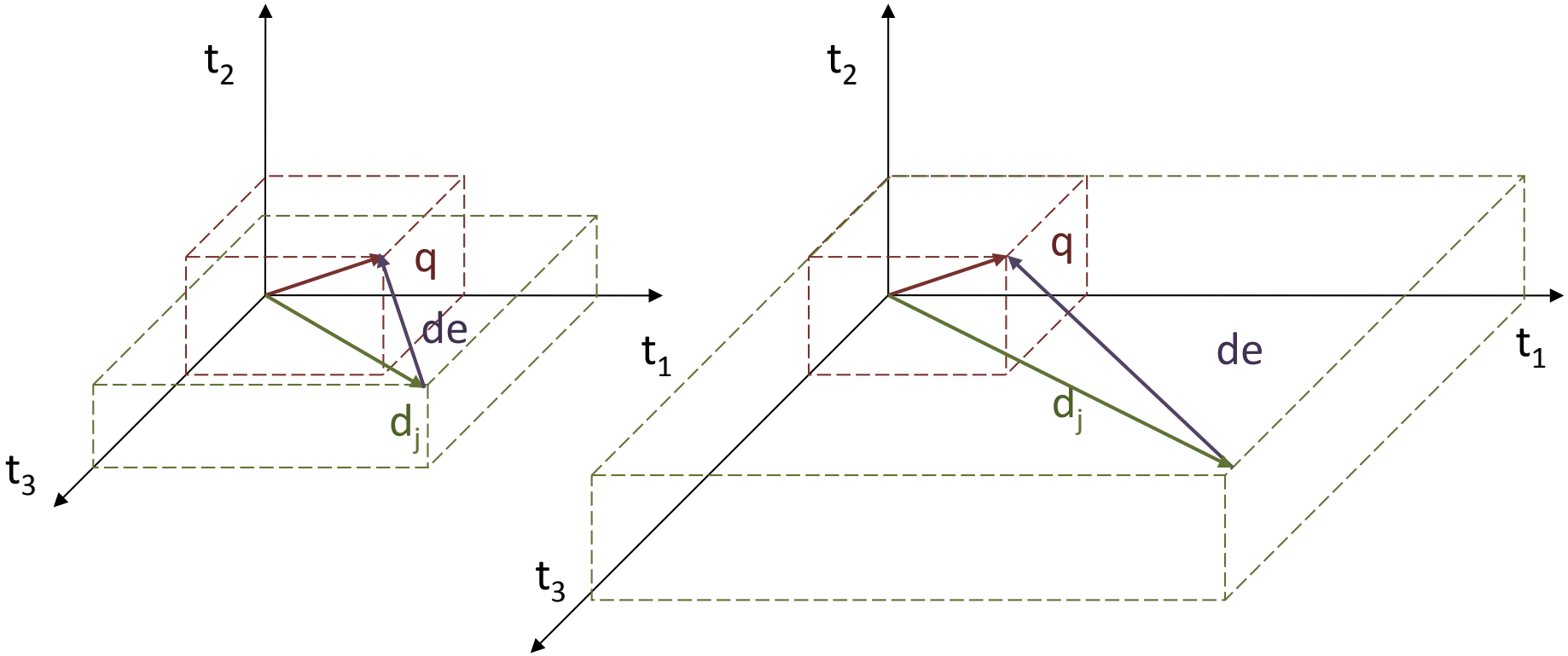
doc	rank computation	rank
d_1	$\frac{1*3+0.415*0.830}{5.068}$	0.660
d_2	$\frac{1*2+0.415*0}{4.899}$	0.408
d_3	$\frac{1*0+0.415*1.073}{3.762}$	0.118
d_4	$\frac{1*0+0.415*1.073}{7.738}$	0.058

Relembrando: IDF

	term	n_i	$idf_i = \log(N/n_i)$
1	to	2	1
2	do	3	0.415

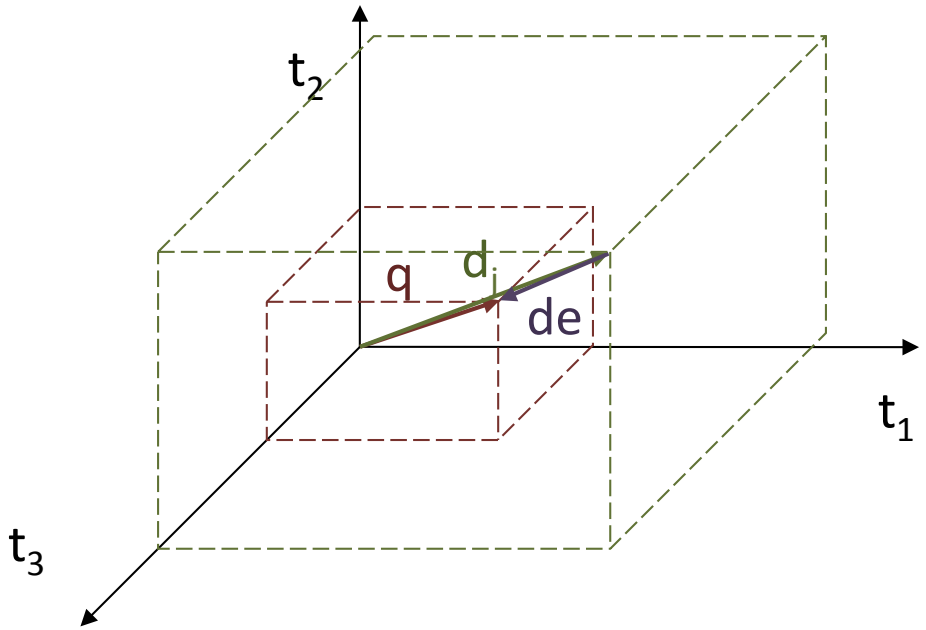
Por que não utilizar distância euclidiana?

- Distância deveria ser igual em ambas as figuras



Por que não utilizar distância euclidiana?

- Deveria ser zero na figura

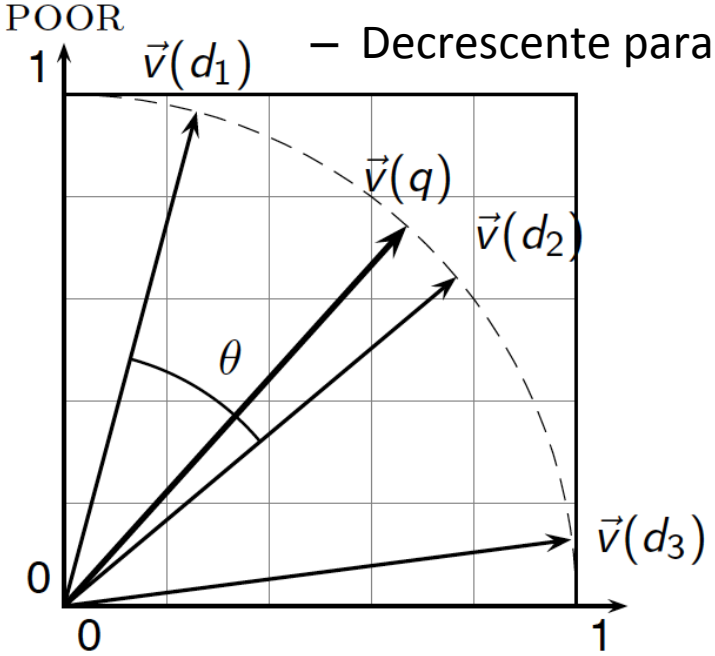


Por que não utilizar distância euclidiana?

- Não representa a similaridade entre os vetores
  - Distância pode ser maior que os vetores
- Pesos representam a importância de cada termo em relação aos outros termos
  - Vetores com comprimentos diferentes, mas com mesmo ângulo possuem distribuições de termos equivalente

Por que não utilizar distância euclidiana?

- A similaridade é inversamente proporcional ao ângulo entre os vetores
  - O cosseno do ângulo é uma boa função de similaridade
    - Imagem varia no intervalo [0,1]
    - Decrescente para o domínio no intervalo [0°, 90°]



Vetores normalizados

Cada componente é dividido pelo comprimento do vetor

[Manning et al., 2008]

Modelo Vetorial

- Principais vantagens:
  - Esquema de atribuição de pesos aos termos melhora a performance da recuperação
  - Estratégia de casamento parcial
    - Permite recuperação de documentos que se “aproximam” das condições da consulta
  - Ordenação dos documentos de acordo com o grau de similaridade em relação à consulta
- Desvantagem:
  - Conceitualmente, não considera a correlação entre os termos

Exercício - Modelo Vetorial

- Partir da implementação desenvolvida na aula anterior (ponderação de termos). Após utilizar a ponderação TF-IDF para atribuição dos pesos dos termos, aplique o modelo vetorial e compute a similaridade entre a consulta e cada um dos documentos.
- Por fim, gere o *ranking* final dos documentos para uma consulta q especificada (ordem que os documentos seriam ranqueados).

Exercício - Relembrando

- Exemplo de entradas:

```
M=['O peã e o caval são pec de xadrez. O caval é
o melhor do jog.';
'A jog envolv a torr, o peã e o rei.';
'O peã lac o boi';
'Caval de rodei!';
'Polic o jog no xadrez.']; //conjunto de
documentos
stopwords=['a', 'o', 'e', 'é', 'de', 'do', 'no',
'são']; //lista de stopwords
q='xadrez peã caval torr'; //consulta
separadores=[' ',',',',','.','!','?']; //separadores
para tokenizacao
```

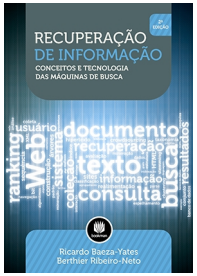
Exercício - Relembrando

- Sua implementação deve:
  - Tokenizar os documentos utilizando os separadores adequados
  - Normalizar termos (ex. caixa-baixa) e eliminar stopwords das consultas e documentos
  - Usar uma solução de indexação utilizando uma variação da matriz de incidências (obs.: guarde a frequência de aparecimento dos termos em cada documento)

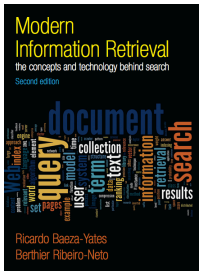
Exercício - Relembrando

- Implemente a ponderação TF-IDF para atribuição dos pesos dos termos (você deve utilizar o 3º esquema de ponderação sugerido por [Salton, 1971] – slide 3).

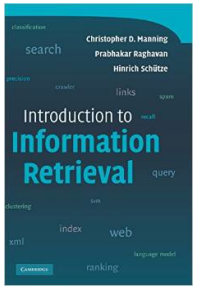
Referências



- Baeza-Yates, R.; Ribeiro-Neto, B. Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca. 2 ed. Bookman, 2013.



- Baeza-Yates, R.; Ribeiro-Neto, B. Modern Information Retrieval. Wokingham, UK: Addison-Wesley, 2 ed., 2011.



- Manning, C. D.; Raghavan, P.; Schütze, H. Introduction to Information Retrieval. Cambridge University Press, 2008.

  Online edition 2009: http://nlp.stanford.edu/IR-book/

Universidade Federal do Rio de Janeiro (UFRJ)
Departamento de Ciência da Computação (DCC)

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO - UFRJ -

Recuperação da Informação (MAB605)

Dúvidas?

Profa. Giseli Rabello Lopes
giseli@dcc.ufrj.br
CCMN - DCC - Sala E-2012