



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



# Recuperação da Informação (MAB605)

Realimentação de relevância e  
Expansão de consulta

Profa. Giseli Rabello Lopes

# Roteiro

---

- Introdução
- Um framework para métodos de realimentação
  - Realimentação de relevância explícita
  - Realimentação explícita através de cliques
  - Realimentação implícita através de análise local
  - Realimentação implícita através de análise global
- Referências

# Introdução

---

- A maioria dos usuários encontra dificuldades para formular consultas bem projetadas para fins de recuperação
- No entanto, usuários muitas vezes precisam reformular suas consultas para obter os resultados que lhes interessam
  - Assim, a primeira formulação da consulta deve ser tratada como uma tentativa inicial de recuperar informações relevantes
  - Os documentos inicialmente recuperados poderiam ser analisados por relevância e usados para melhorar a formulação da consulta inicial

# Introdução

---

- O processo de modificação da consulta é comumente referenciado como
  - **Realimentação de relevância** (*relevance feedback*), quando o usuário fornece explicitamente informações sobre os documentos relevantes para a consulta, ou
  - **Expansão de consulta** (*query expansion*), quando informações relacionadas à consulta são utilizadas para expandi-las
- Nos referimos a ambos deles como ***métodos de realimentação***

# Introdução

---

- Duas abordagens básicas de métodos de realimentação:
  - **Realimentação explícita**, em que a informação para reformulação da consulta é fornecida diretamente pelos usuários, e
  - **Realimentação implícita**, em que a informação para a reformulação da consulta é derivada implicitamente pelo sistema

# Um *framework* para métodos de *realimentação*

---

- Considere um conjunto de documentos  $D_r$  que são sabidamente relevantes para a consulta corrente  $q$
- Na realimentação de relevância, os documentos em  $D_r$  são usados para transformar  $q$  em uma consulta modificada  $q_m$
- Entretanto, obter informações sobre a relevância dos documentos para uma consulta requer a direta interferência do usuário
  - A maioria dos usuários não estão dispostos a fornecer esta informação, particularmente na Web

# Um *framework* para métodos de *realimentação*

---

- Devido a esse custo elevado, a ideia de realimentação de relevância foi flexibilizada ao longo dos anos
- Ao invés de solicitar ao usuário os documentos relevantes, pode-se:
  - Analisar documentos que os usuários tenham clicado; ou
  - Observar os termos pertencentes aos documentos do topo do conjunto de resultados
- Em ambos os casos, é esperado que o ciclo de realimentação produza resultados de melhor qualidade

# Um *framework* para métodos de *realimentação*

---

- Um **ciclo de realimentação** é composto por duas etapas básicas:
  - Determinar a informação de realimentação que está relacionada, ou que se espera que esteja relacionada à consulta original  $q$  e
  - Determinar como transformar a consulta  $q$  de modo a utilizar essa informação de forma eficaz



# Um *framework* para métodos de *realimentação*

---

- A primeira etapa pode ser realizada de duas formas distintas:
  - Obter a informação de realimentação ***explicitamente*** dos usuários
  - Obter a informação de realimentação ***implicitamente*** a partir dos resultados da consulta ou de fontes externas, como um tesouro (*thesaurus*)

# Um *framework* para métodos de *realimentação*

---

- Em um ciclo de **realimentação de relevância explícita**, a informação de realimentação é provida diretamente pelos usuários
- Entretanto, coletar informação de realimentação é caro e consome tempo

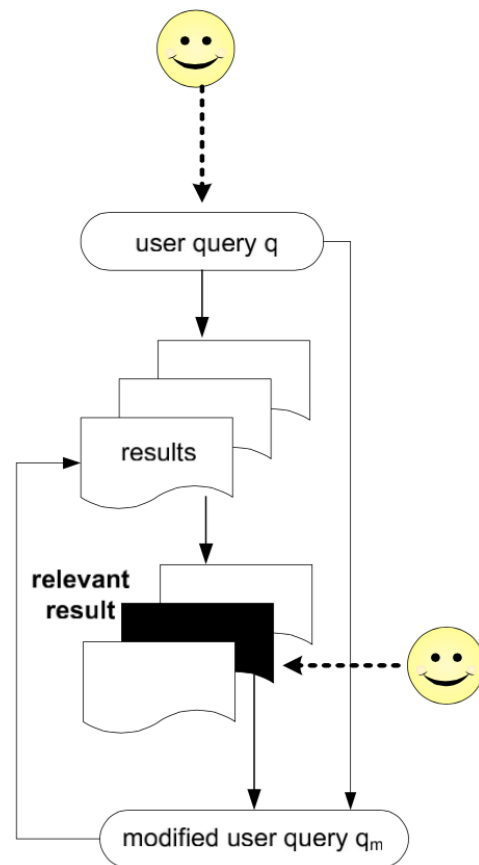
# Um *framework* para métodos de *realimentação*

---

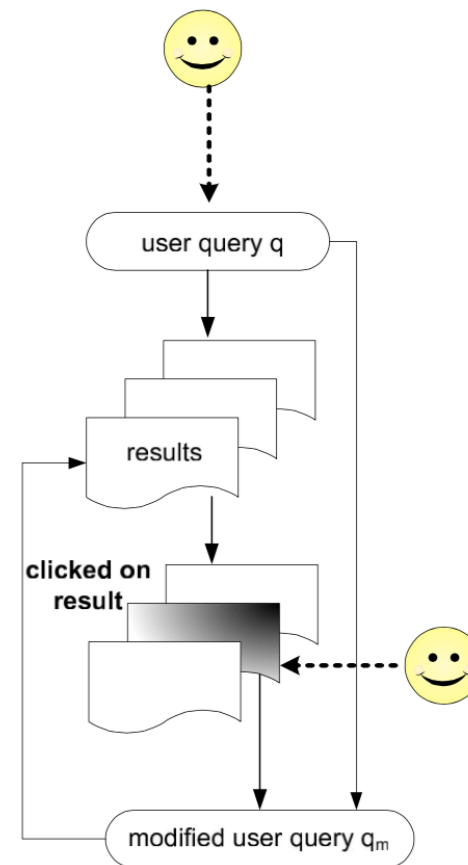
- Na Web, os **cliques do usuário** sobre os resultados da busca constituem uma nova fonte de informação de realimentação
- Um clique indica que um documento é de interesse para o usuário no contexto da consulta atual
  - Note que um clique não necessariamente indica que um documento é relevante para a consulta

# Informação explícita de realimentação

## Explicit Feedback



(a) relevance feedback



(b) click feedback

# Realimentação de relevância explícita

---

- Em um ciclo clássico de realimentação de relevância, uma lista de documentos recuperados é apresentada ao usuário
- Então, o usuário examina os documentos e marca aqueles que são relevantes
- Na prática, apenas os top 10 (ou 20) documentos precisam ser examinados

# Realimentação de relevância explícita

---

- A ideia principal consiste em
  - Selecionar termos importantes dos documentos que foram identificados como relevantes, e
  - Aumentar a importância desses termos em uma nova formulação da consulta

# Realimentação de relevância explícita

---

- **Efeito esperado:** a nova consulta será movida para mais perto dos documentos relevantes e para mais longe dos documentos não relevantes
- Experimentos iniciais mostraram melhorias na precisão em pequenas coleções de teste

# Realimentação de relevância explícita

---

- A realimentação de relevância apresenta as seguintes características:
  - Evita que o usuário tenha que se envolver com o processo de reformulação da consulta (tudo que o usuário tem que fornecer são os julgamentos de relevância)
  - Divide a tarefa de busca em uma sequência de pequenos passos que são mais fáceis de entender

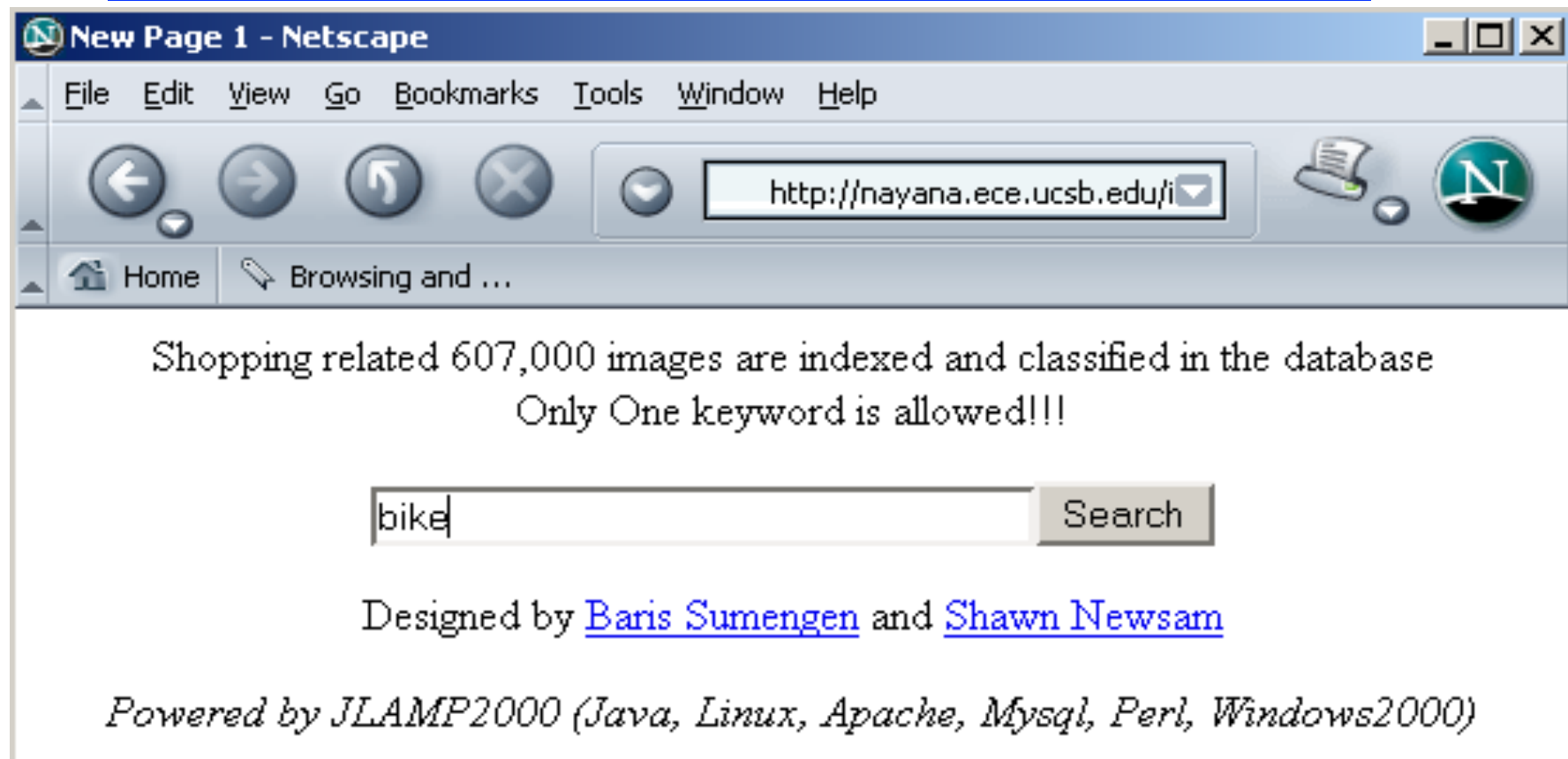


# Exemplo de realimentação de relevância explícita [Manning et al., 2008]

- Máquina de busca por imagens

<http://nayana.ece.ucsb.edu/imsearch/imsearch.html>













*Site não existe mais!*



# Resultados para a consulta inicial

[Manning et al., 2008]






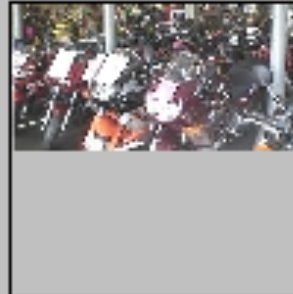






Browse Search Prev Next Random

					
(144473, 16458)	(144457, 252140)	(144456, 262857)	(144456, 262863)	(144457, 252134)	(144483, 265154)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
					
(144483, 264644)	(144483, 265153)	(144518, 257752)	(144538, 525937)	(144456, 249611)	(144456, 250064)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0













# Realimentação de relevância

[Manning et al., 2008]

[Browse](#)
[Search](#)
[Prev](#)
[Next](#)
[Random](#)

					
(144473, 16458) 0.0 0.0 0.0	(144457, 252140) 0.0 0.0 0.0	(144456, 262857) 0.0 0.0 0.0	(144456, 262863) 0.0 0.0 0.0	(144457, 252134) 0.0 0.0 0.0	(144483, 265154) 0.0 0.0 0.0
					
(144483, 264644) 0.0 0.0 0.0	(144483, 265153) 0.0 0.0 0.0	(144518, 257752) 0.0 0.0 0.0	(144538, 525937) 0.0 0.0 0.0	(144456, 249611) 0.0 0.0 0.0	(144456, 250064) 0.0 0.0 0.0

# Resultados depois da realimentação de relevância [Manning et al., 2008]

<a href="#">Browse</a> <a href="#">Search</a> <a href="#">Prev</a> <a href="#">Next</a> <a href="#">Random</a>					
					
(144538, 523493) 0.54182 0.231944 0.309876	(144538, 523835) 0.56319296 0.267304 0.295889	(144538, 523529) 0.584279 0.280881 0.303398	(144456, 253569) 0.64501 0.351395 0.293615	(144456, 253568) 0.650275 0.411745 0.23853	(144538, 523799) 0.66709197 0.358033 0.309059
					
(144473, 16249) 0.6721 0.393922 0.278178	(144456, 249634) 0.675018 0.4639 0.211118	(144456, 253693) 0.676901 0.47645 0.200451	(144473, 16328) 0.700339 0.309002 0.391337	(144483, 265264) 0.70170796 0.36176 0.339948	(144478, 512410) 0.70297 0.469111 0.233859

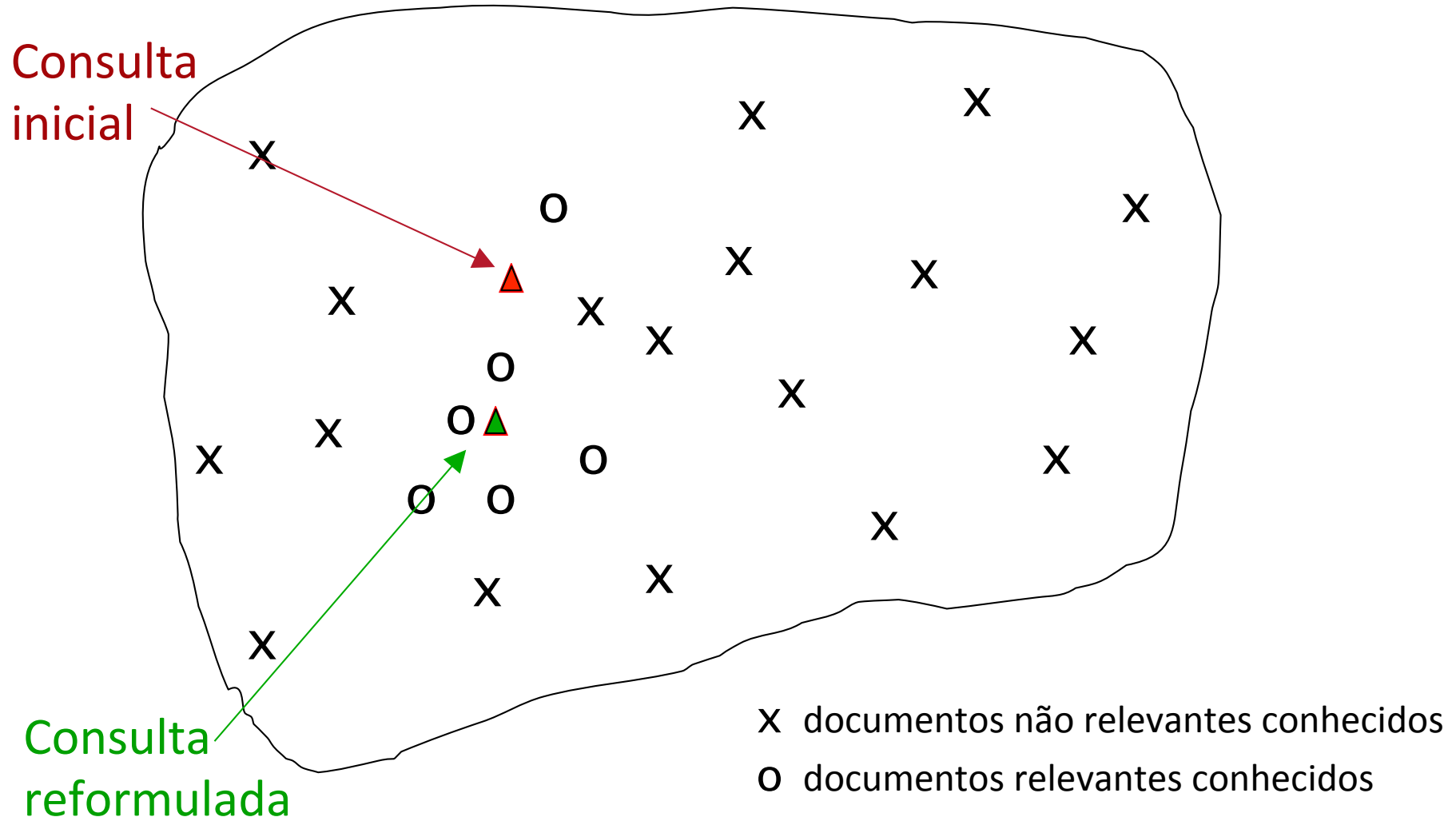


# O método de Rocchio (para o modelo vetorial)

---

- Documentos identificados como relevantes (para uma dada consulta) têm similaridades entre si
- Além disso, documentos não relevantes têm vetores de pesos de termos que são diferentes dos de documentos relevantes
- A ideia básica do método de Rocchio é reformular a consulta de tal forma que ela fique:
  - Mais próxima da vizinhança dos documentos relevantes no espaço vetorial, e
  - Longe da vizinhança dos documentos não relevantes

# Realimentação de relevância em uma consulta inicial [Manning et al., 2008]



# O método de Rocchio

---

- Vamos definir a terminologia sobre o processamento de uma dada consulta  $q$ , como segue:
  - $D_r$ : conjunto de documentos *relevantes* dentre os documentos recuperados
  - $N_r$ : número de documentos no conjunto  $D_r$
  - $D_n$ : conjunto de documentos *não relevantes* dentre os documentos recuperados
  - $N_n$ : número de documentos no conjunto  $D_n$
  - $C_r$ : conjunto de documentos *relevantes* dentre todos os documentos da coleção
  - $N$ : número de documentos da coleção
  - $\alpha, \beta, \gamma$ : constantes de sintonização (*tuning*)

# O método de Rocchio

---

- Considere que o conjunto  $C_r$  é conhecido *a priori*
- Então, o melhor vetor de consulta para distinguir documentos relevantes dos documentos não relevantes é dado por

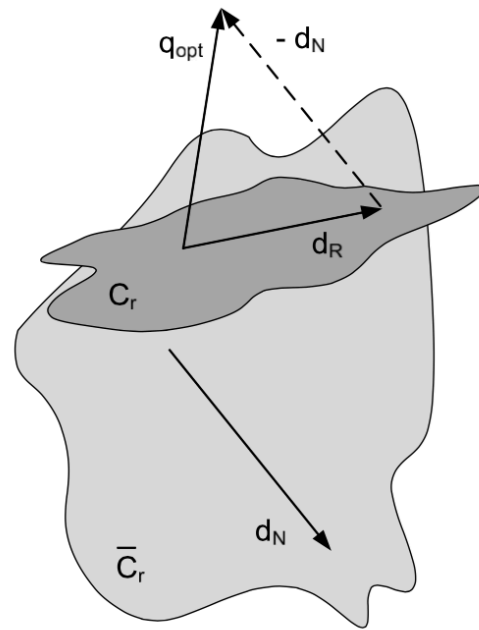
$$\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{\forall \vec{d}_j \in C_r} \vec{d}_j - \frac{1}{N - |C_r|} \sum_{\forall \vec{d}_j \notin C_r} \vec{d}_j$$

- Onde
  - $|C_r|$  refere-se à cardinalidade do conjunto  $C_r$
  - $\vec{d}_j$  é um vetor de termos com pesos associados ao documento  $d_j$
  - $\vec{q}_{opt}$  é o vetor de termos com pesos ótimo para a consulta  $q$

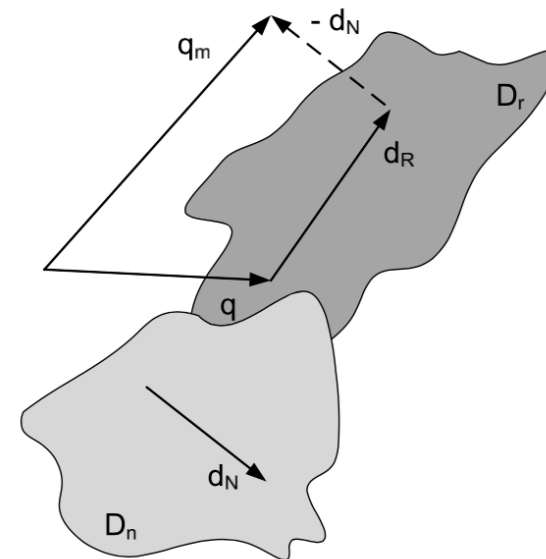


# O método de Rocchio

- Entretanto, o conjunto  $C_r$  não é conhecido *a priori*
- Para resolver este problema, podemos formular a consulta inicial e incrementalmente mudar o vetor de consulta inicial



(a)



(b)

# O método de Rocchio

---

- Existem três modos clássicos e similares para calcular a consulta modificada  $\vec{q}_m$  como segue,

$$\textit{Standard\_Rocchio} : \quad \vec{q}_m = \alpha \vec{q} + \frac{\beta}{N_r} \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{N_n} \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

$$\textit{Ide\_Regular} : \quad \vec{q}_m = \alpha \vec{q} + \beta \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \gamma \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

$$\textit{Ide\_Dec\_Hi} : \quad \vec{q}_m = \alpha \vec{q} + \beta \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \gamma \max\_rank(D_n)$$

- onde  $\max\_rank(D_n)$  é o documento não relevante mais bem ranqueado

# O método de Rocchio

---

- As principais vantagens das técnicas de realimentação de relevância anteriores são simplicidade e bons resultados
  - **Simplicidade:** os pesos modificados dos termos são computados diretamente do conjunto de documentos recuperados
  - **Bons resultados:** o vetor de consulta modificado reflete a porção da semântica da consulta pretendida (experimentalmente observado)

# Exemplo [Lin, 2006]

$$Q' = \alpha Q + \beta \sum_{i=1}^{n_1} \frac{R_i}{n_1} - \gamma \sum_{i=1}^{n_2} \frac{S_i}{n_2}$$

Pesos arbitrários, obtidos experimentalmente (ver TREC)

Comum deixá-lo em zero para usar somente os termos relevantes (Grossman e Frieder 2004)

Tipicamente:  $\gamma < \beta$

Consulta original

0	4	0	8	0	0
---	---	---	---	---	---

$\alpha = 1.0$

--	--	--	--	--	--

Feedback positivo

2	4	8	0	0	2
---	---	---	---	---	---

$\beta = 0.5$

--	--	--	--	--	--

(+)

Feedback negativo

8	0	4	4	0	16
---	---	---	---	---	----

$\gamma = 0.25$

--	--	--	--	--	--

(-)

Nova consulta

--	--	--	--	--	--

# Exemplo [Lin, 2006]

$$Q' = \alpha Q + \beta \sum_{i=1}^{n_1} \frac{R_i}{n_1} - \gamma \sum_{i=1}^{n_2} \frac{S_i}{n_2}$$

Pesos arbitrários, obtidos experimentalmente (ver TREC)

Comum deixá-lo em zero para usar somente os termos relevantes (Grossman e Frieder 2004)

Tipicamente:  $\gamma < \beta$

Consulta original

0	4	0	8	0	0
---	---	---	---	---	---

$\alpha = 1.0$

0	4	0	8	0	0
---	---	---	---	---	---

Feedback positivo

2	4	8	0	0	2
---	---	---	---	---	---

$\beta = 0.5$

--	--	--	--	--	--

(+)

Feedback negativo

8	0	4	4	0	16
---	---	---	---	---	----

$\gamma = 0.25$

--	--	--	--	--	--

(-)

Nova consulta

--	--	--	--	--	--

# Exemplo [Lin, 2006]

$$Q' = \alpha Q + \beta \sum_{i=1}^{n_1} \frac{R_i}{n_1} - \gamma \sum_{i=1}^{n_2} \frac{S_i}{n_2}$$

Pesos arbitrários, obtidos experimentalmente (ver TREC)

Comum deixá-lo em zero para usar somente os termos relevantes (Grossman e Frieder 2004)

Tipicamente:  $\gamma < \beta$

Consulta original

0	4	0	8	0	0
---	---	---	---	---	---

$\alpha = 1.0$

0	4	0	8	0	0
---	---	---	---	---	---

Feedback positivo

2	4	8	0	0	2
---	---	---	---	---	---

$\beta = 0.5$

1	2	4	0	0	1
---	---	---	---	---	---

(+)

Feedback negativo

8	0	4	4	0	16
---	---	---	---	---	----

$\gamma = 0.25$

--	--	--	--	--	--

(-)

Nova consulta

--	--	--	--	--	--

# Exemplo [Lin, 2006]

$$Q' = \alpha Q + \beta \sum_{i=1}^{n_1} \frac{R_i}{n_1} - \gamma \sum_{i=1}^{n_2} \frac{S_i}{n_2}$$

Pesos arbitrários, obtidos experimentalmente (ver TREC)

Comum deixá-lo em zero para usar somente os termos relevantes (Grossman e Frieder 2004)

Tipicamente:  $\gamma < \beta$

Consulta original

0	4	0	8	0	0
---	---	---	---	---	---

$\alpha = 1.0$

0	4	0	8	0	0
---	---	---	---	---	---

Feedback positivo

2	4	8	0	0	2
---	---	---	---	---	---

$\beta = 0.5$

1	2	4	0	0	1
---	---	---	---	---	---

(+)

Feedback negativo

8	0	4	4	0	16
---	---	---	---	---	----

$\gamma = 0.25$

2	0	1	1	0	4
---	---	---	---	---	---

(-)

Nova consulta

--	--	--	--	--	--

# Exemplo [Lin, 2006]

$$Q' = \alpha Q + \beta \sum_{i=1}^{n_1} \frac{R_i}{n_1} - \gamma \sum_{i=1}^{n_2} \frac{S_i}{n_2}$$

Pesos arbitrários, obtidos experimentalmente (ver TREC)

Comum deixá-lo em zero para usar somente os termos relevantes (Grossman e Frieder 2004)

Tipicamente:  $\gamma < \beta$

Consulta original

0	4	0	8	0	0
---	---	---	---	---	---

$\alpha = 1.0$

0	4	0	8	0	0
---	---	---	---	---	---

Feedback positivo

2	4	8	0	0	2
---	---	---	---	---	---

$\beta = 0.5$

1	2	4	0	0	1
---	---	---	---	---	---

(+)

Feedback negativo

8	0	4	4	0	16
---	---	---	---	---	----

$\gamma = 0.25$

2	0	1	1	0	4
---	---	---	---	---	---

(-)

---

Nova consulta

-1	6	3	7	0	-3
----	---	---	---	---	----



# Avaliação de realimentação de relevância

---

- Considere que o vetor de consulta modificado  $\vec{q}_m$  produzido pela expansão de  $\vec{q}$  com documentos relevantes, de acordo com a fórmula de Rocchio
- Avaliação de  $\vec{q}_m$ :
  - Comparar os documentos recuperados por  $\vec{q}_m$  com o conjunto de documentos relevantes para  $\vec{q}$
  - Em geral, os resultados mostram melhorias espetaculares
  - Entretanto, uma parte dessa melhoria resulta de *ranks* maiores associados a documentos relevantes usados para expandir  $\vec{q}$  em  $\vec{q}_m$
  - Uma vez que o usuário já viu esses documentos, tal avaliação não é realista

# A coleção residual

---

- Uma abordagem mais realista é avaliar  $\vec{q}_m$  considerando apenas a **coleção residual**
  - Chamamos coleção residual o conjunto de todos os documentos menos o conjunto de documentos com realimentação provida pelo usuário
- Então, as curvas de recall-precision para  $\vec{q}_m$  tendem a ser mais baixas que as curvas para o vetor de consultas original  $\vec{q}$
- Esta não é uma limitação porque o principal propósito do processo é comparar estratégias distintas de realimentação de relevância

# Exercício - Modelo de Rocchio

---

- Partir da implementação desenvolvida nas aulas anteriores. Considere que um usuário examinou todos os documentos do *ranking* retornado pelo modelo vetorial para a consulta original  $q$  (considere que serão recuperados todos os documentos que tiverem um escore maior do que zero) e marcou como relevantes os documentos indicados em  $R$ . Com base nesta informação, calcule qual seria o *ranking* gerado por uma iteração de realimentação de relevância explícita, em que a informação para reformulação da consulta é fornecida diretamente pelos usuários, de acordo com o método de Rocchio padrão. Indique também qual seria a consulta modificada  $q_m$ .

Obs.: Por questões de simplicidade, ranqueie novamente todos os documentos da coleção.

---

# Exercício - Relembrando

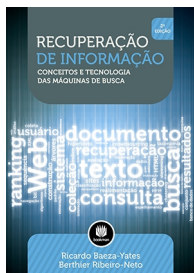
---

- Exemplo de entradas:

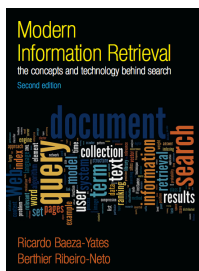
```
M=['O peã e o caval são pec de xadrez. O caval é o melhor  
do jog.'];  
'A jog envolv a torr, o peã e o rei.';  
'O peã lac o boi';  
'Caval de rodei!';  
'Polic o jog no xadrez.']; //conjunto de documentos  
stopwords=['a', 'o', 'e', 'é', 'de', 'do', 'no',  
'são']; //lista de stopwords  
q='xadrez peã caval torr'; //consulta  
separadores=[' ','.',',','!','?']; //separadores para  
tokenizacao  
R=[1; 2]; //identificador dos documentos relevantes para a  
consulta q  
alpha=1; beta=0.75; gama=0.15; //parâmetros do método de  
Rocchio
```

# Referências

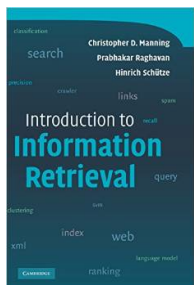
---



- 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/>

# Referências

---

- Lin, Jimmy. Relevance Feedback (lâminas de aula). Disponível em: [www.umiacs.umd.edu/~jimmylin/LBSC796-INFM718R-2006-Spring/lecture7.ppt](http://www.umiacs.umd.edu/~jimmylin/LBSC796-INFM718R-2006-Spring/lecture7.ppt). College of Information Studies, University of Maryland, 2006.



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



# Recuperação da Informação (MAB605)

## Dúvidas?

Profa. Giseli Rabello Lopes  
**[giseli@dcc.ufrj.br](mailto:giseli@dcc.ufrj.br)**  
CCMN - DCC - Sala E-2012



Material selecionado e traduzido de slides do capítulo 4 do livro [Baeza-Yates & Ribeiro-Neto, 2013]