

GSI024 - Organização e recuperação de informação

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

Última atualização - Junho/2022

Realimentação de relevância e
expansão de consultas

Agenda

“Realimentação de relevância e expansão de consultas”

Introdução e contextualização

Métodos de realimentação explícitos

- Método de Rocchio

- Cliques

Métodos de realimentação implícitos

- Local (análise de contexto)

- Global (tesauro de similaridade)

QP-3

QP-3

- Provavelmente será na semana que vem...
- Avaliação + Realimentação + Alguma coisa de modelos de RI

Aula passada

Coleção de referência

- Coleções de referência permitem comparar diretamente os resultados produzidos por diferentes funções de ranqueamento;
- Os julgamentos de relevância são produzidos por humanos especialistas e idealmente devem fornecer uma decisão de relevância para cada par necessidade de informação-documento;
- Claramente, isso só é viável para coleções de documento pequenas, como as dos experimentos Cranfield.

Precisão e revocação

As medidas de precisão e revocação são definidas da seguinte forma:

Precisão (fração dos documentos recuperados que é relevante):

$$p = |R \cap A| / |R|$$

Revocação (fração dos documentos relevantes que foi recuperada):

$$r = |R \cap A| / |A|$$



- Na Web, é comum medir a média da precisão quando $n = 5$ ou 10 documentos tenham sido vistos;
- Os valores típicos para n são normalmente precisão na posição 5 ($P@5$), precisão na posição 10 ($P@10$) e precisão na posição 20 ($P@20$);
- Essas métricas fornecem uma avaliação da impressão do usuário sobre os resultados.

MAP - Exemplo

Calcular o MAP para o conjunto de consultas q_1 ($R_1 = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89} \text{ e } d_{123}\}$) e q_2 ($R_2 = \{d_3, d_{56}, d_{129}\}$).

1. d_{123} •
2. d_{84}
3. d_{56} •
4. d_6
5. d_8

6. d_9 •
7. d_{511}
8. d_{129}
9. d_{187}
10. d_{25} •

11. d_{38}
12. d_{48}
13. d_{250}
14. d_{113}
15. d_3 •

1. d_{425}
2. d_{87}
3. d_{56} •
4. d_{32}
5. d_{124}

6. d_{615}
7. d_{512}
8. d_{129} •
9. d_4
10. d_{130}

11. d_{193}
12. d_{715}
13. d_{810}
14. d_5
15. d_3 •

Coeficiente de Spearman - Exemplo

Documentos	$s_{1,j}$	$s_{2,j}$	$s_{i,j} - s_{2,j}$	$(s_{i,j} - s_{2,j})^2$
d_{123}	1	2	-1	1
d_{84}	2	3	-1	1
d_{56}	3	1	+2	4
d_6	4	5	-1	1
d_8	5	4	+1	1
d_9	6	7	-1	1
d_{511}	7	8	-1	1
d_{129}	8	10	-2	4
d_{187}	9	6	+3	9
d_{25}	10	9	+1	1
Soma dos quadrados das distâncias				24

Introdução

Introdução

- Sem um conhecimento detalhado da coleção de documentos, a maioria dos usuários acha difícil formular consultas bem projetadas para fins de recuperação;
- Exemplo: usuários de sistemas de RI muitas vezes precisam reformular suas consultas para obter os resultados que interessam.
 - A primeira consulta deve ser tratada como uma tentativa inicial de recuperar informações relevantes!
 - Formulações melhores da consulta podem ser escritas para recuperar mais documentos úteis.

Introdução

Como melhorar a formulação da consulta inicial utilizando a informação que está relacionada com a intenção “por trás” da consulta?

Introdução

Como melhorar a formulação da consulta inicial utilizando a informação que está relacionada com a intenção “por trás” da consulta?

- 1) Realimentação explícita – quando o usuário fornece explicitamente informações sobre os documentos relevantes para uma consulta;

Introdução

Como melhorar a formulação da consulta inicial utilizando a informação que está relacionada com a intenção “por trás” da consulta?

- 1) Realimentação explícita – quando o usuário fornece explicitamente informações sobre os documentos relevantes para uma consulta;
- 2) Realimentação implícita – quando informações relacionadas à consulta são utilizadas implicitamente pelo sistema.

Métodos de realimentação - Definição

Definição: A realimentação de relevância refere-se a um ciclo de realimentação em que documentos que são conhecidamente relevantes para a consulta q em questão são usados para transformá-la em uma consulta modificada q_m .

- A expectativa é que a consulta q_m retornará um maior número de documentos relevantes para q .

Métodos de realimentação - Problemas

- Obter informações sobre a relevância dos documentos em relação a consulta:
 - É caro;
 - Exige a interferência direta do usuário;
- Exemplo: um sistema de RI poderia perguntar aos usuários se os 10 primeiros resultados para uma determinada consulta são de fato relevantes. Será que os usuários estão dispostos a fornecer essa informação?

Métodos de realimentação – Possível solução

- Em vez de pedir aos usuários que marquem os documentos relevantes, poderíamos analisar documentos que:
 - Eles tenham clicado;
 - Ou observar os termos pertencentes aos documentos do topo do conjunto dos resultados.
- Em ambos os casos, se supusermos que a informação recolhida está relacionada à consulta original, esperamos que o ciclo de realimentação produza resultados de melhor qualidade!

Ciclo de realimentação - Etapas

- 1) Determinar a informação de realimentação que está relacionada, ou que se espera que esteja relacionada à consulta original;
- 2) Determinar como transformar a consulta q de modo a utilizar essa informação de forma eficaz.

Ciclo de realimentação - Etapas

A etapa 1) pode ser realizada de duas formas distintas:

- A. Obter **explicitamente** a informação de realimentação a partir dos usuários;
- B. Obter **implicitamente** a informação de realimentação a partir dos resultados da consulta ou de fontes externas, como um tesouro.

Informações explícitas de realimentação



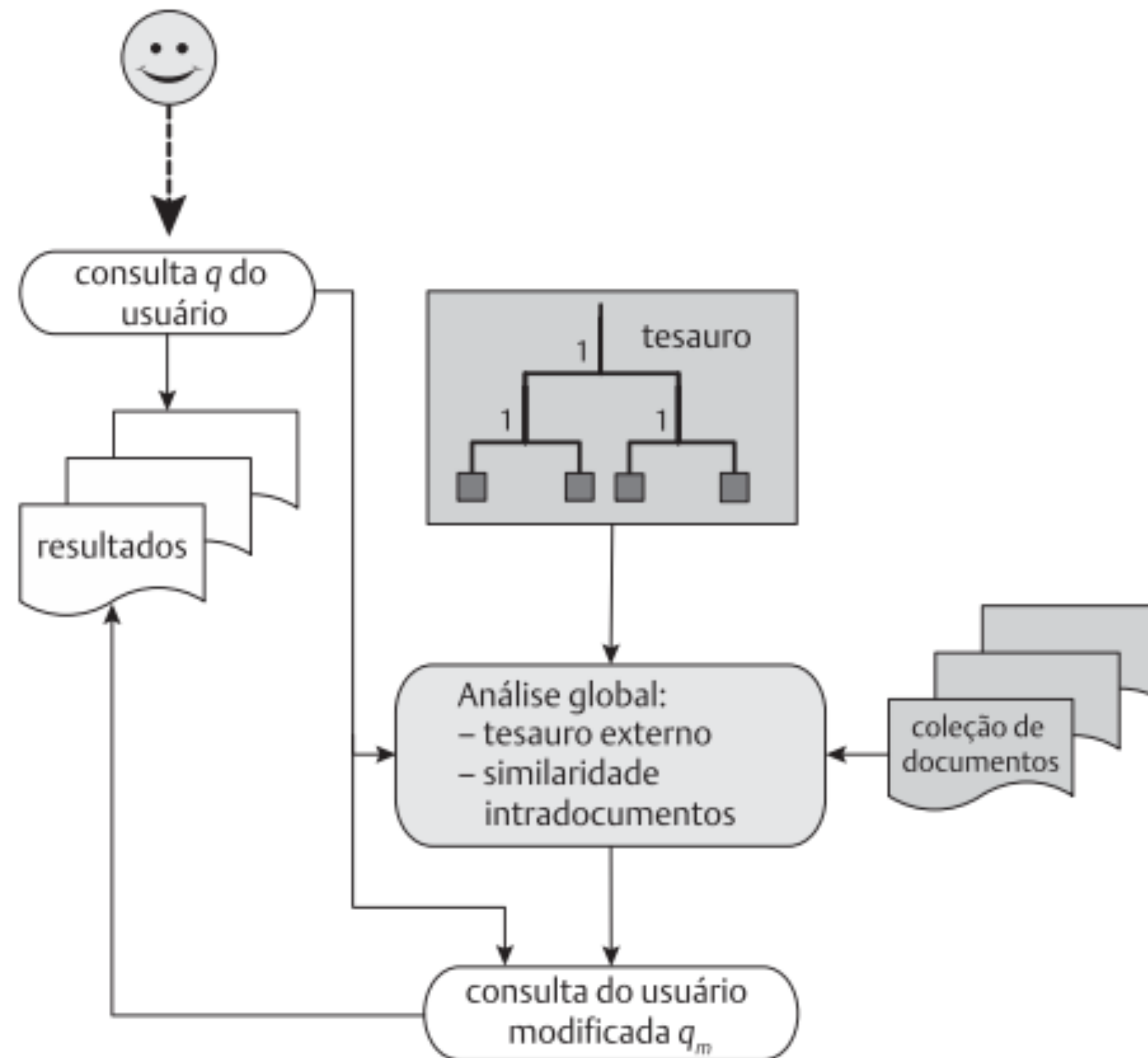
Informações explícitas de realimentação



Informações implícitas de realimentação



Informações implícitas de realimentação



Métodos de realimentação explícitos

Realimentação explícita - Cliques

- Usuários de máquinas de busca na Web não só inspecionam os resultados de suas consultas, como também clicam sobre eles;
- Os cliques podem ser coletados em grandes números, sem interferir nas ações dos usuários;
- Pergunta: os dados de cliques podem ser usados para decidir sobre a relevância do resultado para consultas futuras?

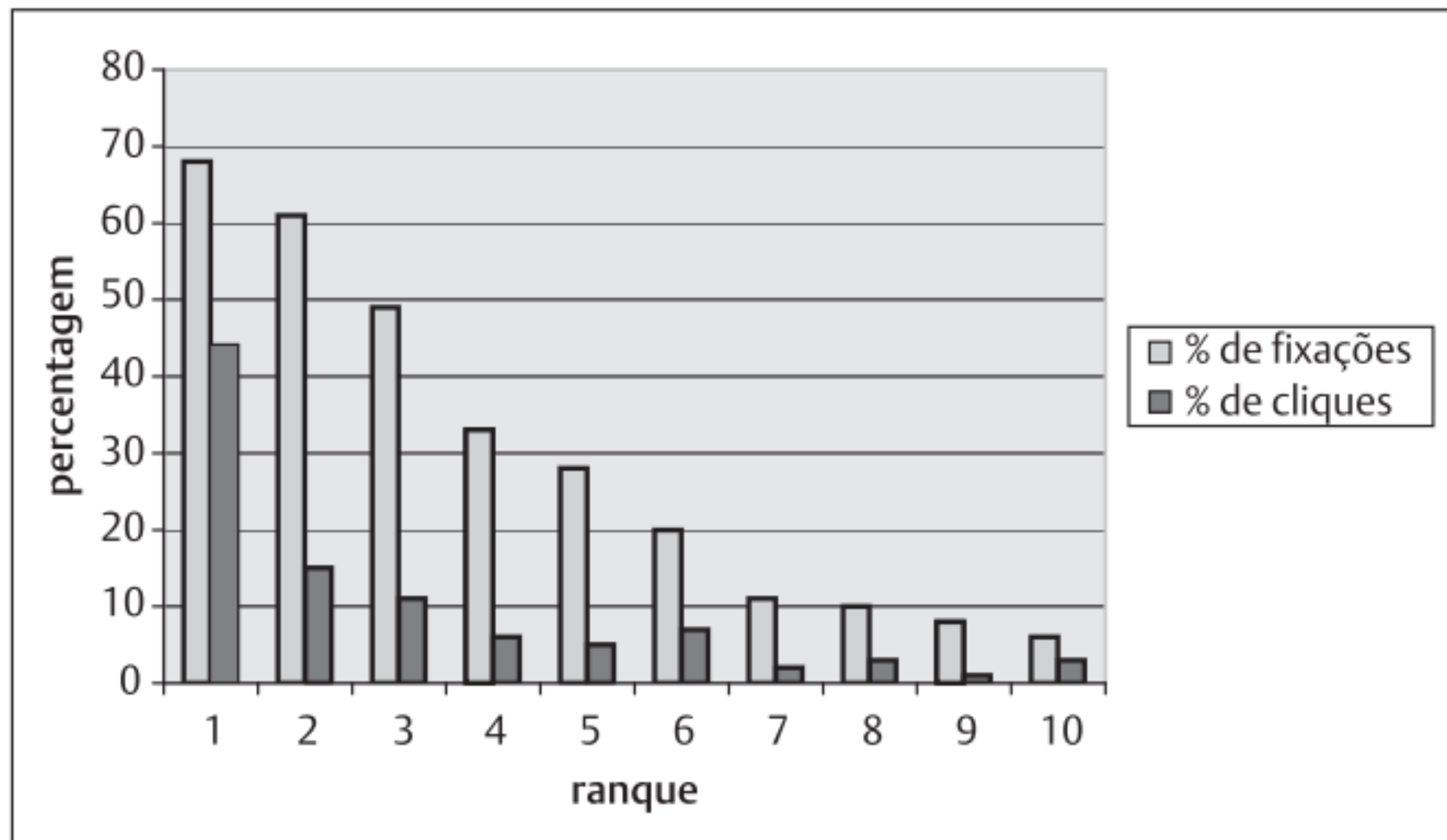
Realimentação explícita – Comportamento do usuário

- Experimentos com um grupo de 29 indivíduos;
- Resultados:
 - Usuários escaneiam os resultados da consulta de cima para baixo;
 - Inspeccionam o primeiro e o segundo resultados de imediato;
 - Tendem a escanear em detalhe as primeiras cinco ou seis respostas que aparecem na área visível da tela;

Realimentação explícita – Comportamento do usuário

- Resultados:
 - 60-70% das tarefas, os usuários têm uma fixação no primeiro ou no segundo resultado – para o quarto resultado a frequência cai pela metade;
 - Usuários inspecionam as duas primeiras respostas quase que igualmente, mas eles clicam quase três vezes mais no primeiro resultado;
 - Indicativo de que o usuário tende a confiar na máquina de busca!

Realimentação explícita - Cliques



Realimentação explícita - Cliques

- Participantes recebem dois conjuntos distintos de resultados:
 - O ranking normal retornado pela máquina de busca;
 - Um ranking modificado, no qual os dois melhores resultados têm a sua posição trocada.
- O que acontece?

Realimentação explícita - Cliques

- Participantes recebem dois conjuntos distintos de resultados:
 - O ranking normal retornado pela máquina de busca;
 - Um ranking modificado, no qual os dois melhores resultados têm a sua posição trocada.
- O que acontece?
- Usuários clicam quase três vezes mais no primeiro resultado do que no segundo!
 - A posição do resultado tem uma grande influência na decisão do usuário.

Realimentação explícita - Cliques

- Interpretar cliques como um **indicativo direto** de relevância não é uma boa abordagem...
- Ideia: interpretar cliques como métricas de **PREFERÊNCIA** do usuário;
- Exemplo: Se você olha para o *snippet* de um resultado e decide ignorá-lo e clicar em um resultado mais abaixo no ranking, é apropriado dizer que este usuário prefere o resultado clicado ao mostrado mais acima do ranking.

Cliques dentro de uma mesma consulta (exemplo)

$r_1 \quad r_2 \quad \sqrt{r_3} \quad r_4 \quad \sqrt{r_5} \quad r_6 \quad r_7 \quad r_8 \quad r_9 \quad \sqrt{r_{10}}$

- Skip-Above: supõe que o usuário prefere o resultado em que ele clicou a todos os outros resultados em que ele não clicou e que aparecerem antes;
- Skip-Previous: usuário prefere o resultado clicado ao resultado imediatamente anterior no ranking que não foi clicado.

Cliques em uma cadeia de consultas (exemplo 2)

$$\begin{array}{cccccccccc} r_1 & r_2 & r_3 & r_4 & r_5 & r_6 & r_7 & r_8 & r_9 & r_{10} \\ s_1 & \checkmark s_2 & s_3 & s_4 & \checkmark s_5 & s_6 & s_7 & s_8 & s_9 & s_{10} \end{array}$$

- Top-One-No-Click-Earlier: supõe que o usuário prefere qualquer resposta do segundo conjunto de resultados à primeira resposta do primeiro conjunto de resultados;
- Top-Two-No-Click-Earlier: supõe que o usuário prefere qualquer resposta do segundo conjunto de resultados às duas primeiras respostas do primeiro conjunto de resultados;

Realimentação de relevância explícita

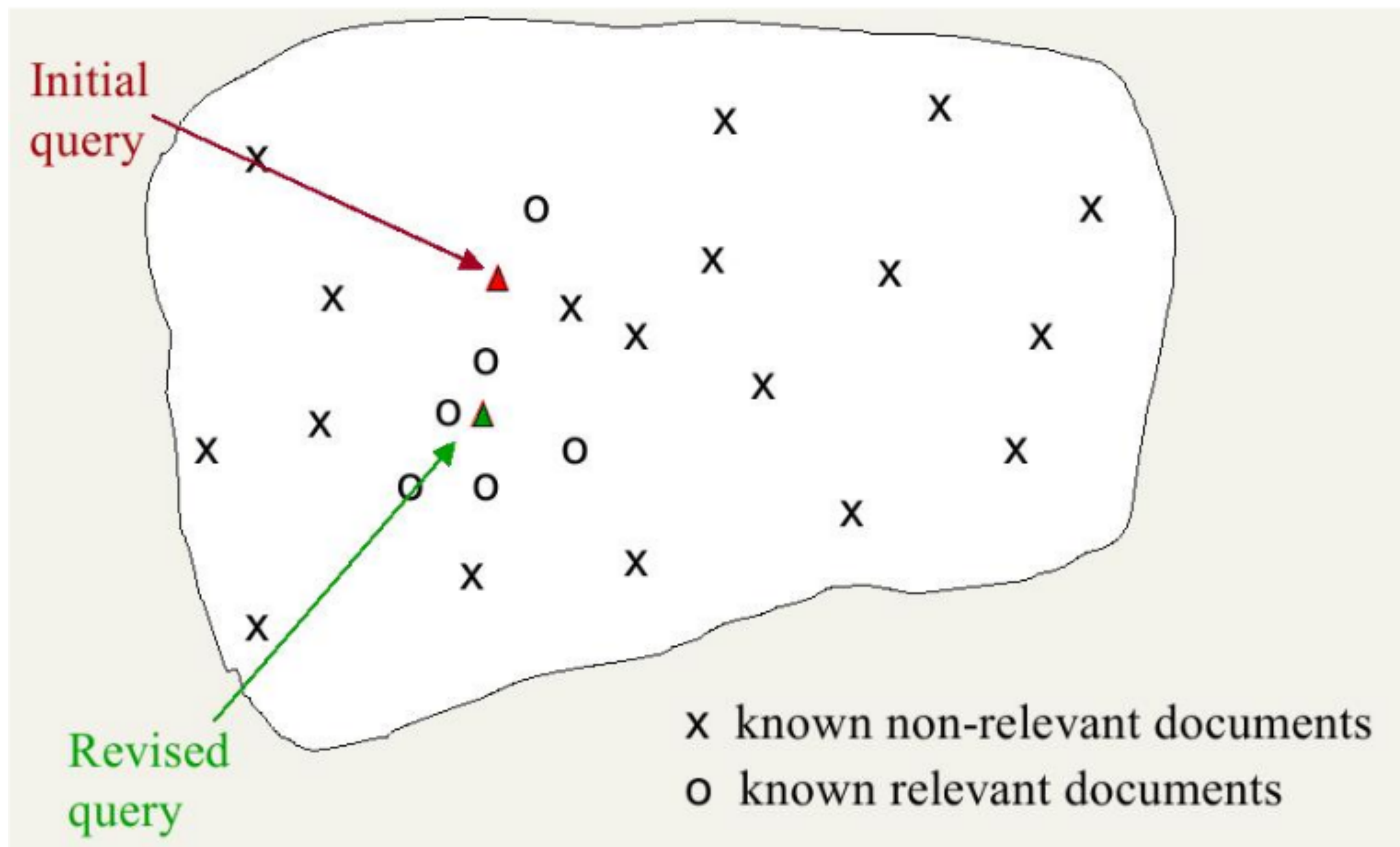
Ideia: reformulação da consulta à partir de informações do usuário sobre a relevância de documentos.

1. Usuário submete a consulta original;
2. Uma lista de documentos recuperados é apresentada ao usuário;
3. O usuário examina os documentos e marca aqueles que são relevantes;
4. Com base na informação fornecida pelo usuário, o sistema computa uma nova consulta;
5. A nova consulta é submetida ao sistema.

Realimentação de relevância explícita

- O objetivo principal consiste em:
 - Selecionar termos importantes dos documentos que foram identificados como relevantes pelos usuários;
 - Aumentar a importância desses termos em uma nova formulação da consulta.
- Espera-se que a nova consulta seja movida para mais perto dos documentos relevantes e para mais longe dos documentos não relevantes.

Realimentação de relevância explícita - Objetivo



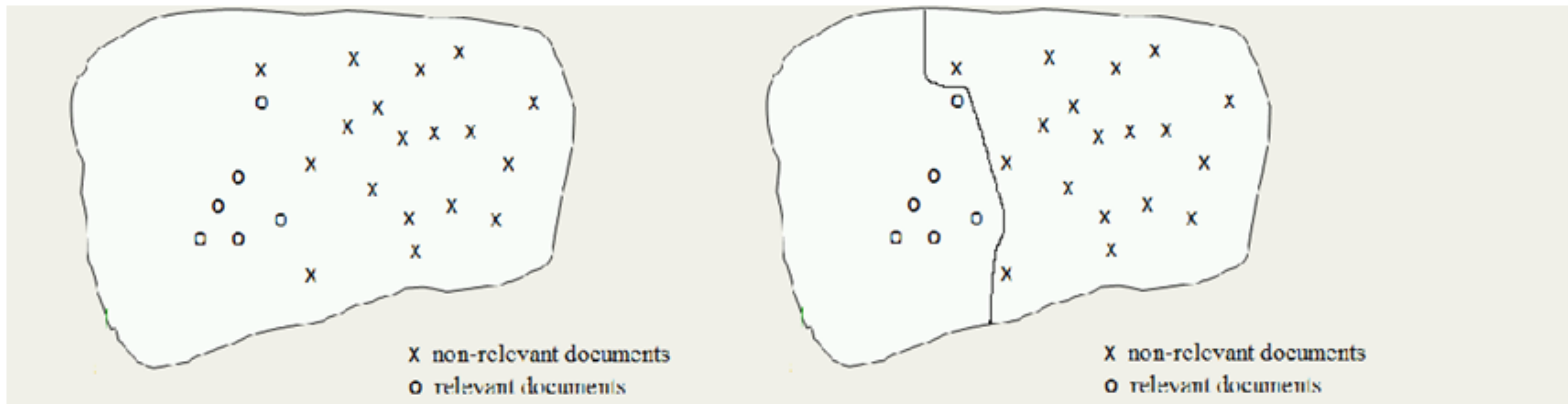
Realimentação de relevância explícita - Características

- Evita que o usuário tenha que se envolver com o processo de reformulação da consulta (ele somente precisa fornecer julgamentos de relevância para os documentos);
- Divide a tarefa da busca em uma sequência de pequenos passos que são mais fáceis de aprender.

Método de Rocchio - Premissas

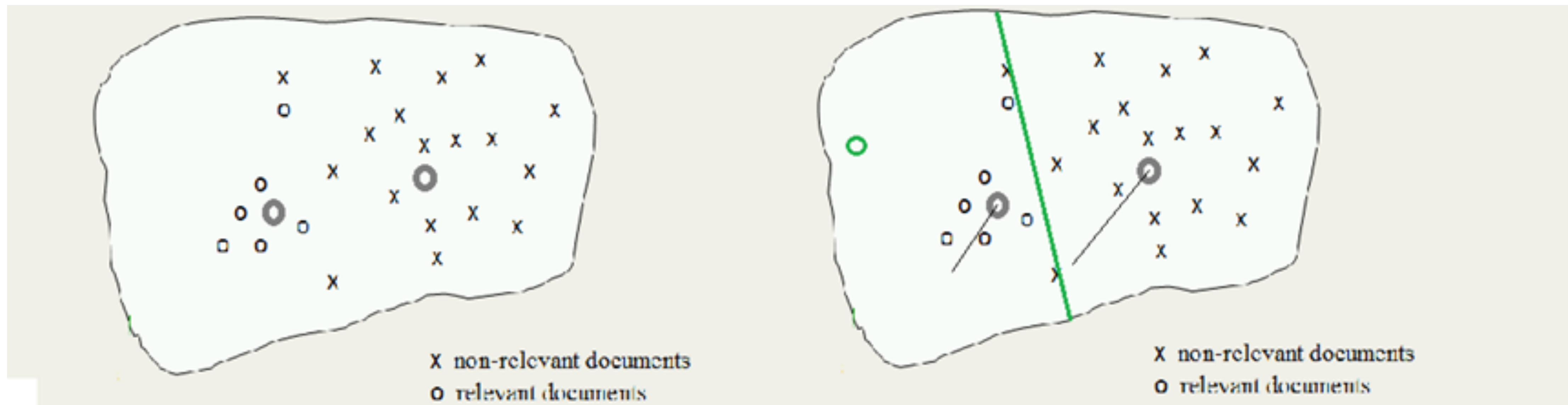
1. Considera que os documentos identificados como relevantes (para uma determinada consulta) têm semelhanças entre si;
 2. Documentos não relevantes têm vetores de termos **diferentes** dos vetores dos documentos relevantes;
- Ideia: reformular a consulta, de forma que ela aproxime-se dos documentos relevantes e afaste-se dos documentos não relevantes.

Realimentação de relevância explícita - Ideia



Realimentação de relevância explícita - Ideia

Vetor diferença entre os centróides dos vetores dos documentos relevantes e não relevantes:



Método de Rocchio – Consulta ótima

Conjunto completo dos documentos relevantes C_r para uma determinada consulta q é conhecido de antemão. Nesse caso, o melhor vetor consulta para distinguir os documentos relevantes dos não relevantes é o vetor diferença entre cada um dos centroides:

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

N = documentos da coleção.

Método de Rocchio (padrão)

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

D_r = documentos relevantes entre os documentos recuperados.

N_r = número de documentos no conjunto D_r .

D_n = documentos não relevantes entre os documentos recuperados.

N_n = número de documentos no conjunto D_n .

q = consulta original.

Outros termos são constantes de ajuste.

Método de Rocchio (padrão) - Exemplo

new query vector = $\alpha \cdot$ original query vector +
 $\beta \cdot$ relevant document vectors -
 $\gamma \cdot$ non-relevant document vectors

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

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

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

Método de Rocchio (padrão) - Exemplo

new query vector = $\alpha \cdot$ original query vector +
 $\beta \cdot$ relevant document vectors -
 $\gamma \cdot$ non-relevant document vectors

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

 $\alpha = 1$

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

 $\beta = 0.5$

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

 $\gamma = 0.25$

Typically $\beta > \gamma$,
since positive
feedback is more
meaningful.

Método de Rocchio (padrão) - Exemplo

new query vector = $\alpha \cdot$ original query vector +
 $\beta \cdot$ relevant document vectors -
 $\gamma \cdot$ non-relevant document vectors

<table><tr><td>0</td><td>4</td><td>0</td><td>8</td><td>0</td><td>0</td></tr></table>	0	4	0	8	0	0	$\alpha = 1$		<table><tr><td>0</td><td>4</td><td>0</td><td>8</td><td>0</td><td>0</td></tr></table>	0	4	0	8	0	0
0	4	0	8	0	0										
0	4	0	8	0	0										
<table><tr><td>2</td><td>4</td><td>8</td><td>0</td><td>0</td><td>2</td></tr></table>	2	4	8	0	0	2	$\beta = 0.5$	+	<table><tr><td>1</td><td>2</td><td>4</td><td>0</td><td>0</td><td>1</td></tr></table>	1	2	4	0	0	1
2	4	8	0	0	2										
1	2	4	0	0	1										
<table><tr><td>8</td><td>0</td><td>4</td><td>4</td><td>0</td><td>16</td></tr></table>	8	0	4	4	0	16	$\gamma = 0.25$	-	<table><tr><td>2</td><td>0</td><td>1</td><td>1</td><td>0</td><td>4</td></tr></table>	2	0	1	1	0	4
8	0	4	4	0	16										
2	0	1	1	0	4										

Typically $\beta > \gamma$,
since positive
feedback is more
meaningful.

Método de Rocchio (padrão) - Exemplo

new query vector = $\alpha \cdot$ original query vector +
 $\beta \cdot$ relevant document vectors -
 $\gamma \cdot$ non-relevant document vectors

<table><tr><td>0</td><td>4</td><td>0</td><td>8</td><td>0</td><td>0</td></tr></table>	0	4	0	8	0	0	$\alpha = 1$		<table><tr><td>0</td><td>4</td><td>0</td><td>8</td><td>0</td><td>0</td></tr></table>	0	4	0	8	0	0
0	4	0	8	0	0										
0	4	0	8	0	0										
<table><tr><td>2</td><td>4</td><td>8</td><td>0</td><td>0</td><td>2</td></tr></table>	2	4	8	0	0	2	$\beta = 0.5$	+	<table><tr><td>1</td><td>2</td><td>4</td><td>0</td><td>0</td><td>1</td></tr></table>	1	2	4	0	0	1
2	4	8	0	0	2										
1	2	4	0	0	1										
<table><tr><td>8</td><td>0</td><td>4</td><td>4</td><td>0</td><td>16</td></tr></table>	8	0	4	4	0	16	$\gamma = 0.25$	-	<table><tr><td>2</td><td>0</td><td>1</td><td>1</td><td>0</td><td>4</td></tr></table>	2	0	1	1	0	4
8	0	4	4	0	16										
2	0	1	1	0	4										
			<hr/>												
			<table><tr><td>-1</td><td>6</td><td>3</td><td>7</td><td>0</td><td>-3</td></tr></table>	-1	6	3	7	0	-3						
-1	6	3	7	0	-3										
			<table><tr><td>0</td><td>6</td><td>3</td><td>7</td><td>0</td><td>0</td></tr></table>	0	6	3	7	0	0						
0	6	3	7	0	0										

Typically $\beta > \gamma$,
since positive
feedback is more
meaningful.

Negative term
weights become 0.

Método de Rocchio - Exercício

Considere a seguinte coleção:

- $D1 = \{\text{good movie trailer shown}\}$
- $D2 = \{\text{trailer with good actor}\}$
- $D3 = \{\text{unseen movie}\}$

Considere também que o vocabulário seja formado pelas palavras movie, trailer e good. Suponha que o usuário considerou $D1$ e $D2$ relevantes para a consulta $Q = \{\text{movie trailer}\}$. Qual seria a consulta modificada usando o método de Rocchio?

Método de Rocchio - Vantagens

- Simplicidade – pesos modificados dos termos são computados diretamente a partir do conjunto de documentos recuperados;
- Bons resultados – vetor modificado da consulta reflete uma parte da semântica da consulta pretendida.

Método de Rocchio - Desvantagens

- Retorno de relevância é caro;
 - Retorno de relevância implica em consultas longas;
 - Consultas longas são caras.
- Usuário evitam fornecer retorno explícito.

Métodos de realimentação implícitos

Realimentação implícita

1. Análise global

2. Análise local

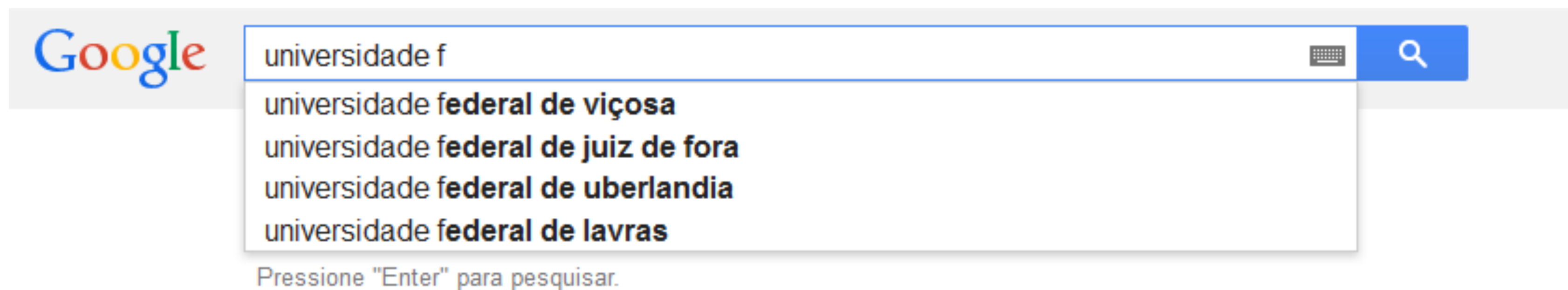
Realimentação implícita – Análise global

- Expandir a consulta usando informações de todo o conjunto de documentos da coleção;
- A consulta é modificada baseada em alguma característica **global** da coleção, isto é, um recurso que não depende da consulta.

Realimentação implícita – Análise global

- Informação principal utilizada: sinônimos (ou quase sinônimos);
- Uma publicação ou base de dados que lista esse tipo de sinônimo é conhecido como **tesauro**;
- Existem dois tipos de tesauro:
 - Criados manualmente;
 - Criados automaticamente.

Análise global - Exemplo



Análise global – Tipos de expansão de consultas

1. Tesouro construído manualmente (mantido por editores, exemplo UNESCO);
2. Tesouro construído automaticamente (baseado em estatísticas de coocorrências, por exemplo);
3. Equivalência de consulta baseada em mineração do histórico de consultas (muito comum na Web como no exemplo do slide anterior).


Expansão de consultas com tesauros

- Para cada termo t na consulta, a ideia consiste em expandi-la com palavras relacionadas ao termo t presentes no tesauro;
 - HOSPITAL – MÉDICO;
- Em geral aumenta a recuperação mas pode diminuir a precisão com a presença de termos ambíguos;
- Utilizado amplamente em motores de busca especializados para ciência e engenharia;
- É muito caro criar e manter um tesauro manualmente.

Exemplo de tesauro manual - UNESCO

- Tesauro utilizado para o controle e indexação de termos das áreas de Educação, Cultura, Ciências Naturais, Sociais e Humanas (<http://databases.unesco.org/thesaurus/>)

Exemplo de tesauro manual - UNESCO

Computer systems > Computer networks		
PREFERRED TERM	Computer networks	 Search in UNESDOC
BROADER CONCEPT	Computer systems	
NARROWER CONCEPTS	Computer interfaces Internet	
RELATED CONCEPTS	Computer applications Computers Computer terminals Telecommunications Telecommunications networks	
ENTRY TERMS	<i>Computer communications</i> <i>Data networks</i> <i>Electronic networking</i>	
BELONGS TO GROUP	Information and communication > Information technology (hardware)	
IN OTHER LANGUAGES	شبكات الحاسب مشابكة الكترونية شبكات البيانات ربط شبكي الكتروني اتصالات حاسوبية	Arabic
	Réseau informatique <i>Réseau d'ordinateurs</i> <i>RE</i> <i>Réseau local d'ordinateurs</i> <i>Réseau électronique</i> <i>RLE</i> <i>Travail en réseau</i>	French

Exemplo de tesouro manual - INEP

- Vocabulário controlado que reúne termos e conceitos extraídos de documentos analisados no Centro de Informação e Biblioteca em Educação (http://pergamum.inep.gov.br/pergamumweb/biblioteca/pesquisa_thesouro.php)

Geração automática de tesouro

- Como gerar de forma automática um tesouro?

Geração automática de tesouro

- Como gerar de forma automática um tesouro?
- Resposta: analisar a **distribuição** das palavras em documentos e usar o conceito de **similaridade** entre duas palavras;

Geração automática de tesouro

- Definição 1: Duas palavras são similares se elas co-ocorrem com as mesmas palavras:
 - “carro” é similar a “motocicleta” pois ambos ocorrem com “estrada”, “gasolina” e “carteira”;
- Definição 2: duas palavras são similares se elas ocorrem em uma relação gramatical com as mesmas palavras:
 - Maça é similar a pera pois em ambos os casos podemos colher, descascar, comer, preparar...

Realimentação implícita – Análise local

- Documentos recuperados para uma determinada consulta q são examinados para determinar os termos para a expansão da consulta;
- Semelhante a um ciclo de realimentação de relevância, mas feito sem o envolvimento do usuário.

Análise local – Abordagem 1 - Clusterings de associações

- Construir matrizes de associação que quantificam as correlações entre os termos dos documentos que aparecem no topo do ranking;
- Quanto maior o número de documentos em que os dois termos coocorrem, mais forte a correlação;
- Termos com tais características podem ser usados para criar a consulta modificada.

Análise local – Abordagem 1 - Clusterings de associações

$$\begin{array}{c}
 \begin{array}{cc} d_1 & d_2 \end{array} \\
 \begin{array}{c} k_1 \\ k_2 \\ k_3 \end{array} \begin{bmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \\ w_{3,1} & w_{3,2} \end{bmatrix} \\
 \mathbf{M}
 \end{array}
 \times
 \begin{array}{c}
 \begin{array}{ccc} k_1 & k_2 & k_3 \end{array} \\
 \begin{array}{c} d_1 \\ d_2 \end{array} \begin{bmatrix} w_{1,1} & w_{2,1} & w_{3,1} \\ w_{1,2} & w_{2,2} & w_{3,2} \end{bmatrix} \\
 \mathbf{M}^T
 \end{array}
 \end{array}$$

$\underbrace{\hspace{15em}}$

$$\begin{array}{c}
 \begin{array}{ccc} k_1 & k_2 & k_3 \end{array} \\
 \begin{array}{c} k_1 \\ k_2 \\ k_2 \end{array} \begin{bmatrix} w_{1,1}w_{1,1} + w_{1,2}w_{1,2} & w_{1,1}w_{2,1} + w_{1,2}w_{2,2} & w_{1,1}w_{3,1} + w_{1,2}w_{3,2} \\ w_{2,1}w_{1,1} + w_{2,2}w_{1,2} & w_{2,1}w_{2,1} + w_{2,2}w_{2,2} & w_{2,1}w_{3,1} + w_{2,2}w_{3,2} \\ w_{3,1}w_{1,1} + w_{3,2}w_{1,2} & w_{3,1}w_{2,1} + w_{3,2}w_{2,2} & w_{3,1}w_{3,1} + w_{3,2}w_{3,2} \end{bmatrix}
 \end{array}$$

matriz de termos por termos

Análise local – Abordagem 1 - Clusterings de associações

Seja $\mathbf{M} = [m_{ij}]$ uma matriz de termos por documentos com t linhas e N colunas onde $m_{ij} = w_{i,j}$, ou seja, cada célula da matriz é dada pelo peso associado ao par termo-documento (k_i, d_j) . Sendo \mathbf{M}^T a transposta de \mathbf{M} , a matriz $\mathbf{C} = \mathbf{M} \cdot \mathbf{M}^T$ é uma matriz de correlação entre termos. Cada elemento $c_{u,v} \in \mathbf{C}$ expressa uma correlação entre os termos k_u e k_v , dada por

$$c_{u,v} = \sum_{d_j} w_{u,j} \times w_{v,j}$$

Análise local – Abordagem 1 - Clusterings de associações

- A matriz C de correlação entre termos estabelece um relacionamento entre quaisquer dois termos baseado em suas coocorrências dentro de documentos da coleção;
- Quanto maior o número de documentos nos quais os termos k_u e k_v coocorram, maior será essa correlação;
- Para criar uma consulta modificada q_m , bastaria adicionar a q_m os termos com maior correlação aos termos da consulta original.

Análise local – Abordagem 1 - Exemplo

- Considere a matriz de correlação da coleção do livro;
- Se a consulta for “to do”, uma abordagem de análise local por clustering de associação poderia criar uma consulta modificada qm = “to be do da”;
- “be” foi inserido pois a correlação entre to e be é alta (12), ou seja, o número de documentos que eles coocorrem é alto! O mesmo vale para o termo “da”;
- As coisas ficam mais interessantes em coleções maiores...

Análise local – Abordagem 2

- Usa grupos de substantivos (único, dois ou três adjacentes no texto) selecionados a partir dos documentos no topo do ranking;
- Em vez de documentos, são usadas passagens de tamanho fixo do documento (por exemplo, 300 palavras) para a determinação das coocorrências entre os substantivos;
- Para cada grupo de substantivo calcula-se a similaridade entre ele e a consulta q – os grupos mais bem ranqueados são adicionados a consulta modificada q_m ;

Análise local – Abordagem 2

- Usa grupos de substantivos (único, dois ou três adjacentes no texto) selecionados a partir dos documentos no topo do ranking;
- Em vez de documentos, são usadas passagens de tamanho fixo do documento (por exemplo, 300 palavras) para a determinação das coocorrências entre os substantivos;
- Para cada grupo de substantivo calcula-se a similaridade entre ele e a consulta q – os grupos mais bem ranqueados são adicionados a consulta modificada q_m ;
- Método ajustado para dados específicos (TREC);
- Não funcionou bem para outras coleções.

Roteiro de estudo

Estudos

- Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca
 - Capítulo 4.1, 4.2, 4.3, 4.4, 4.5.1 e 4.6.1
- Lista 3 - Já disponível no Teams