



Refinamento de Consultas

Prof. Luciano Barbosa

(Parte do material retirado dos slides dos livros adotados)



UNIVERSIDADE
FEDERAL
DE PERNAMBUCO



Consultas com Keywords

- Simples: qualquer pessoa é capaz de usar
- Em geral: engenhos de busca não tratam consultas em linguagem natural
- Curtas:
 - Tamanho médio das consultas: 2,3 palavras (30 palavras em CQA)



Dificuldade do Usuário para Formular Consultas

- Ex: considere uma consulta q : “carro” e um documento d contendo “automóvel” mas não contendo “carro”
 - Um sistema simples de RI não retorna d como resultado de q
 - Mesmo que d seja o documento mais relevante para q
- Solução:
 - Sugerir consultas relacionadas
 - Recuperar documentos ainda se não há termos que casam com a consulta -> melhora a cobertura (recall)



Estatégias de Refinamento

- Spell checking: recomendação para erros de ortografia
- Query expansion: termos relacionados com a consulta é usada para expandi-la
- Relevance feedback: usuários provêem informação sobre documentos relevantes para uma consulta



Spell Checking

- 10-15% de todas consultas têm erro de ortografia
- Vários tipos de erros:

poiner sisters
brimingham news
catamarn sailing
hair extenssions
marshmellow world
miniture golf courses
psychics
home doceration

realstateisting.bc.com
akia 1080i manunal
ultimatwarcade
mainsourcebank
dellottitouche



Spell Checking

- Abordagem básica: sugerir correções para palavras não encontradas no dicionário
- Sugestões: palavras mais similares no dicionário
- Medida de similaridade de string: edit distance
 - Número de operações necessárias para transformar uma palavra na outra



Edit Distance: Damerau-Levenshtein

- Calcula o número mínimo de inserções, deleções e substituições de caracteres necessários para transformar uma string em outra
- Ex: distância 1

extenssions → extensions (insertion error)

poiner → pointer (deletion error)

marshmellow → marshmallow (substitution error)

brimingham → birmingham (transposition error)

- Ex: distância 2

doceration → deceration

deceration → decoration



Edit Distance

- Problema: muito custoso verificar para todo dicionário
- Técnicas usadas para aumentar velocidade do cálculo
 - Começando com o mesmo caracter
 - Mesmo ou tamanho próximo
 - Pronúncia parecida
- Podem existir vários strings no dicionário próximos à consulta

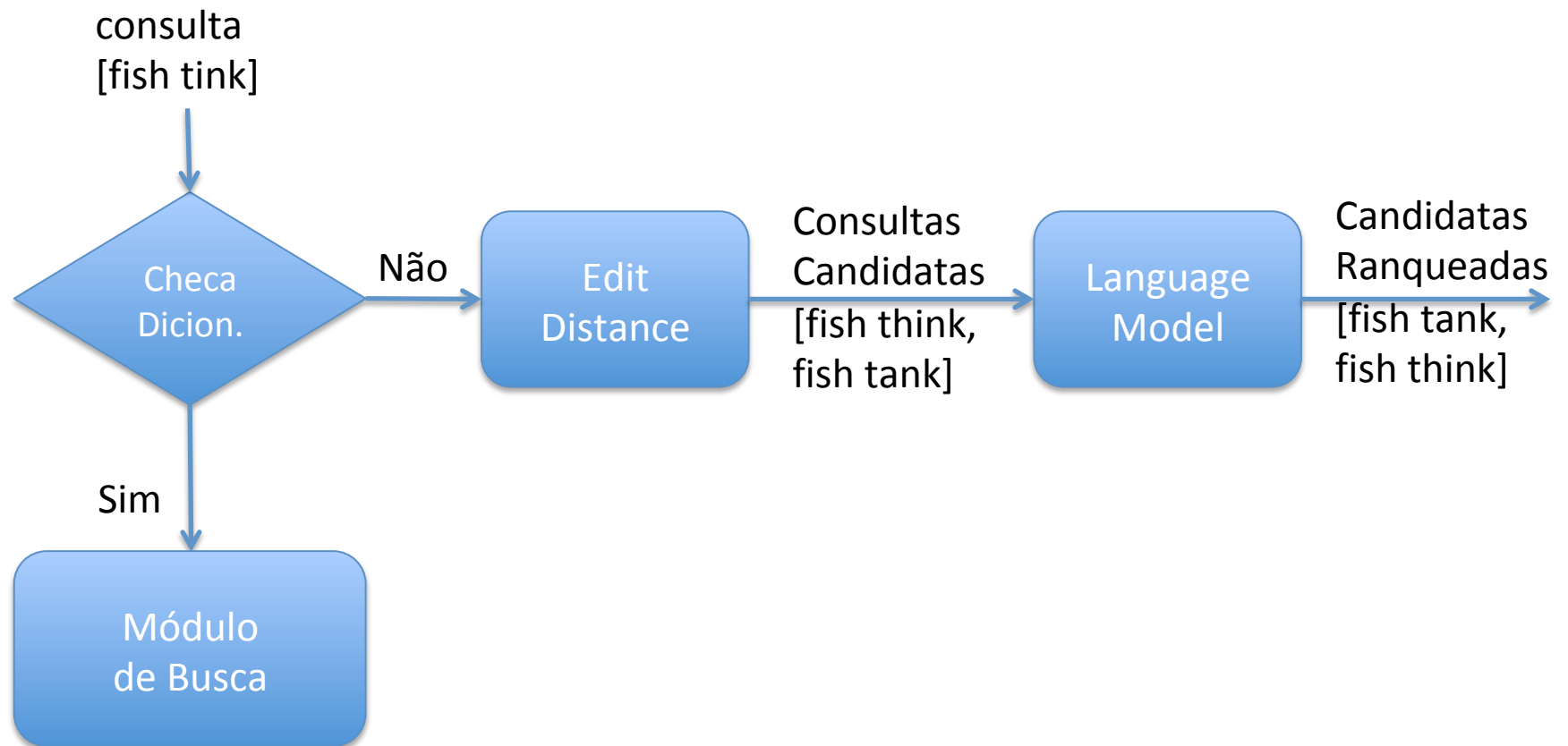


Escolher Melhor Sugestão

- Depende do contexto
 - Ex: lawers -> lowers, lawyers, layers, lasers, lagers
 - Ex: trial lawers -> trial lawyers
- Language model
 - Objetivo: ranquear sugestões
 - Estima probabilidade de uma consulta candidata ser bem formada
 - Usa corpus de texto e log de consulta
 - Ex: consulta: fish tink
 - Possíveis correções: tank ou think
 - $P(\text{tank}|\text{fish}) > P(\text{think}|\text{fish})$



Spell Checking: Funcionamento





Query Expansion

- Refinamento baseado em algum recurso global
- Recurso: palavras associadas (ex. sinônimos)
- Mais usado: thesaurus



Thesaurus: Tipos

- Manual: mantido por editores

MeSH Heading	Neck Pain
Tree Number	C10.597.617.576
Tree Number	C23.888.592.612.553
Tree Number	C23.888.646.501
Entry Term	Cervical Pain
Entry Term	Neckache
Entry Term	Anterior Cervical Pain
Entry Term	Anterior Neck Pain
Entry Term	Cervicalgia
Entry Term	Cervicodynia
Entry Term	Neck Ache
Entry Term	Posterior Cervical Pain
Entry Term	Posterior Neck Pain

- Automático:
 - Baseado em estatísticas



Thesaurus Manual

- Expandir a consulta com palavras semanticamente relacionadas a ela segundo o thesaurus
 - Ex: neck pain -> cervical pain
- Geralmente aumenta cobertura
- Pode diminuir precisão
- Usado em engenhos de busca especializados (ex: ciência e engenharia)
 - Ex: <https://www.ncbi.nlm.nih.gov/mesh>
- Muito caro para criar e manter



Construção Automática de Thesaurus

- Analisar a distribuição das palavras em documentos
- Dois tipos:
 - Duas palavras são similares se elas ocorrem em contextos semelhantes
 - “carro” \approx “motocicleta” por terem como contexto palavras como: “estrada”, “gasolina” etc
 - Duas palavras são similares se elas co-ocorrem
 - Ex: “honda” co-ocorre com “carro”
- Co-ocorrência é mais fácil de extrair



Mutual Information

- Mede quanto duas palavras estão associadas baseado em co-ocorrência em documento

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

- Se os termos forem independentes: $P(x, y) = P(x) * P(y)$
- Se tiver associação:
 - $I(x, y) \gg 0$

Table 3. Some interesting Associations with “Doctor” in the 1987 AP Corpus (N = 15 million)

$I(x, y)$	$f(x, y)$	$f(x)$	x	$f(y)$	y
11.3	12	111	<i>honorary</i>	621	<i>doctor</i>
11.3	8	1105	<i>doctors</i>	44	<i>dentists</i>
10.7	30	1105	<i>doctors</i>	241	<i>nurses</i>
9.4	8	1105	<i>doctors</i>	154	<i>treating</i>
9.0	6	275	<i>examined</i>	621	<i>doctor</i>
8.9	11	1105	<i>doctors</i>	317	<i>treat</i>
8.7	25	621	<i>doctor</i>	1407	<i>bills</i>
8.7	6	621	<i>doctor</i>	350	<i>visits</i>
8.6	19	1105	<i>doctors</i>	676	<i>hospitals</i>
8.4	6	241	<i>nurses</i>	1105	<i>doctors</i>

Some Uninteresting Associations with “Doctor”

0.96	6	621	<i>doctor</i>	73785	<i>with</i>
0.95	41	284690	<i>a</i>	1105	<i>doctors</i>
0.93	12	84716	<i>is</i>	1105	<i>doctors</i>

Fonte: <http://www.aclweb.org/anthology/J90-1003>



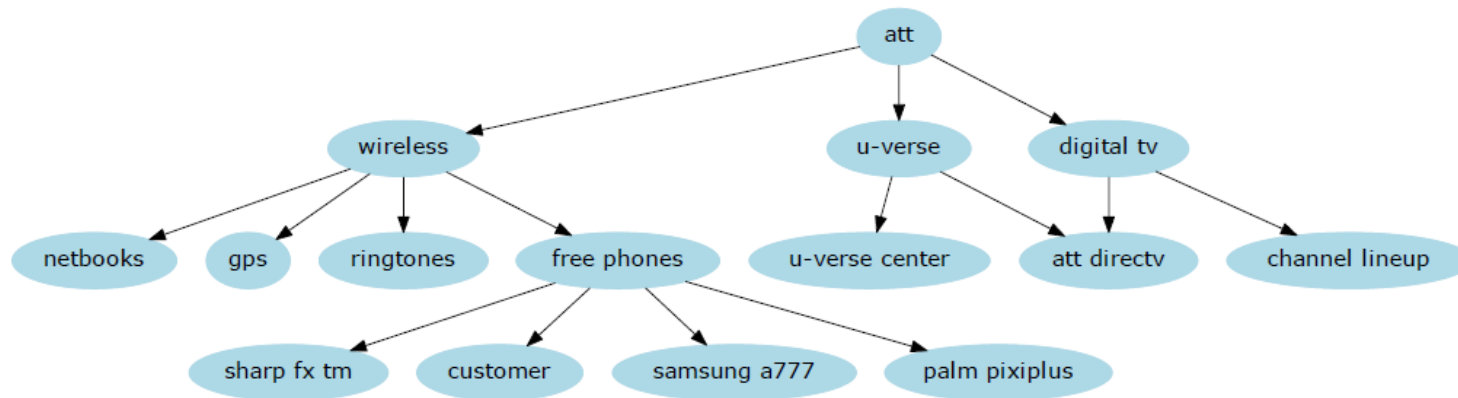
Thesaurus baseado em Co-Ocorrências

Word	Nearest neighbors
absolutely bottomed captivating doghouse makeup mediating keeping lithographs pathogens senses	absurd whatsoever totally exactly nothing dip copper drops topped slide trimmed shimmer stunningly superbly plucky witty dog porch crawling beside downstairs repellent lotion glossy sunscreen skin gel reconciliation negotiate case conciliation hoping bring wiping could some would drawings Picasso Dali sculptures Gauguin toxins bacteria organisms bacterial parasite grasp psyche truly clumsy naive innate

WordSpace demo on web

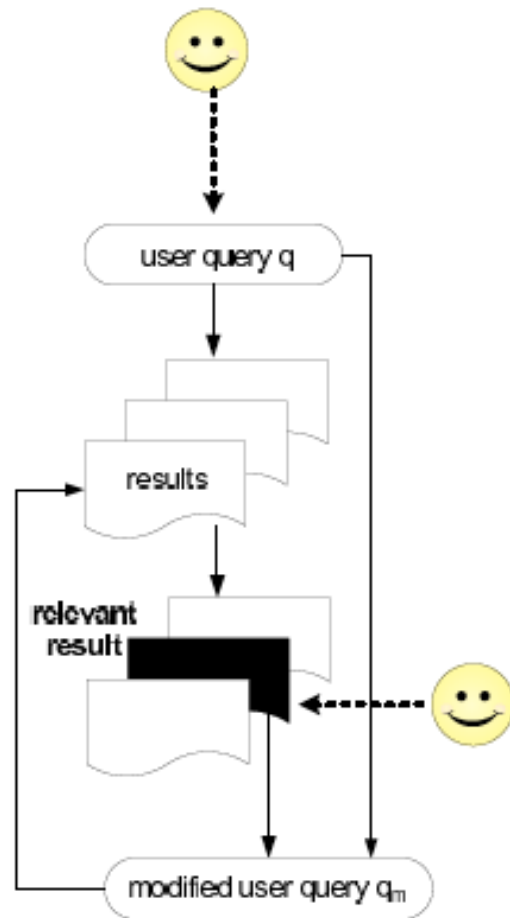


Thesaurus baseado em Hiperlinks



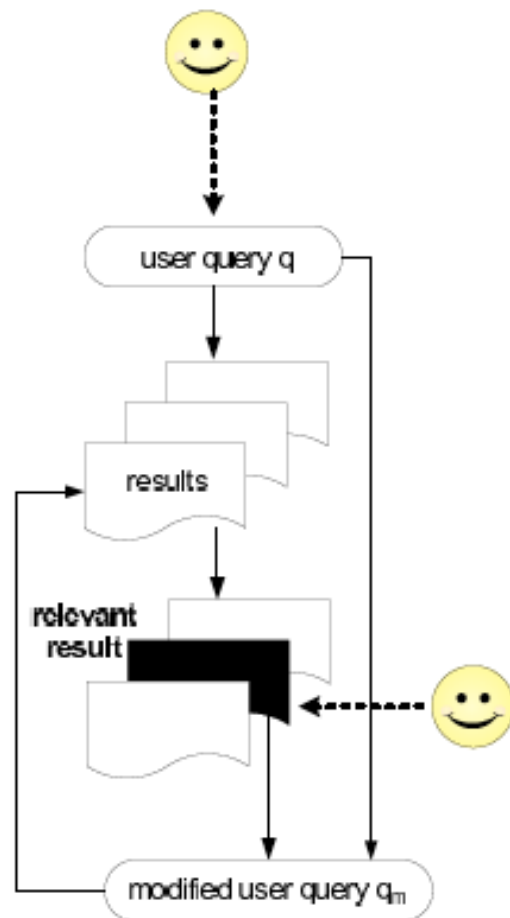


Relevance Feedback





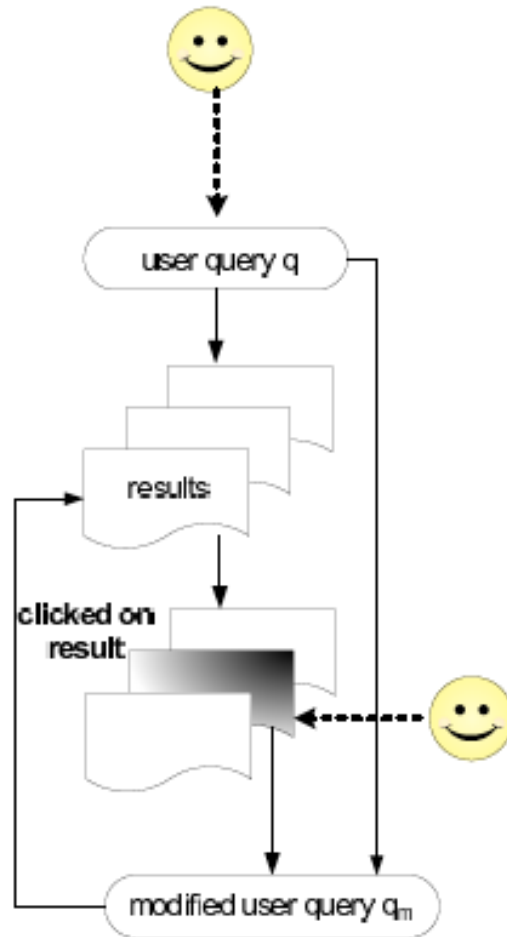
Relevance Feedback



- Usuário ajuda no processo de refinamento
- Com feedback explícito
- Feedback é custoso pro usuário



Relevance Feedback



- Com clique de usuário
- Clique indica interesse do usuário, não necessariamente que o documento é relevante

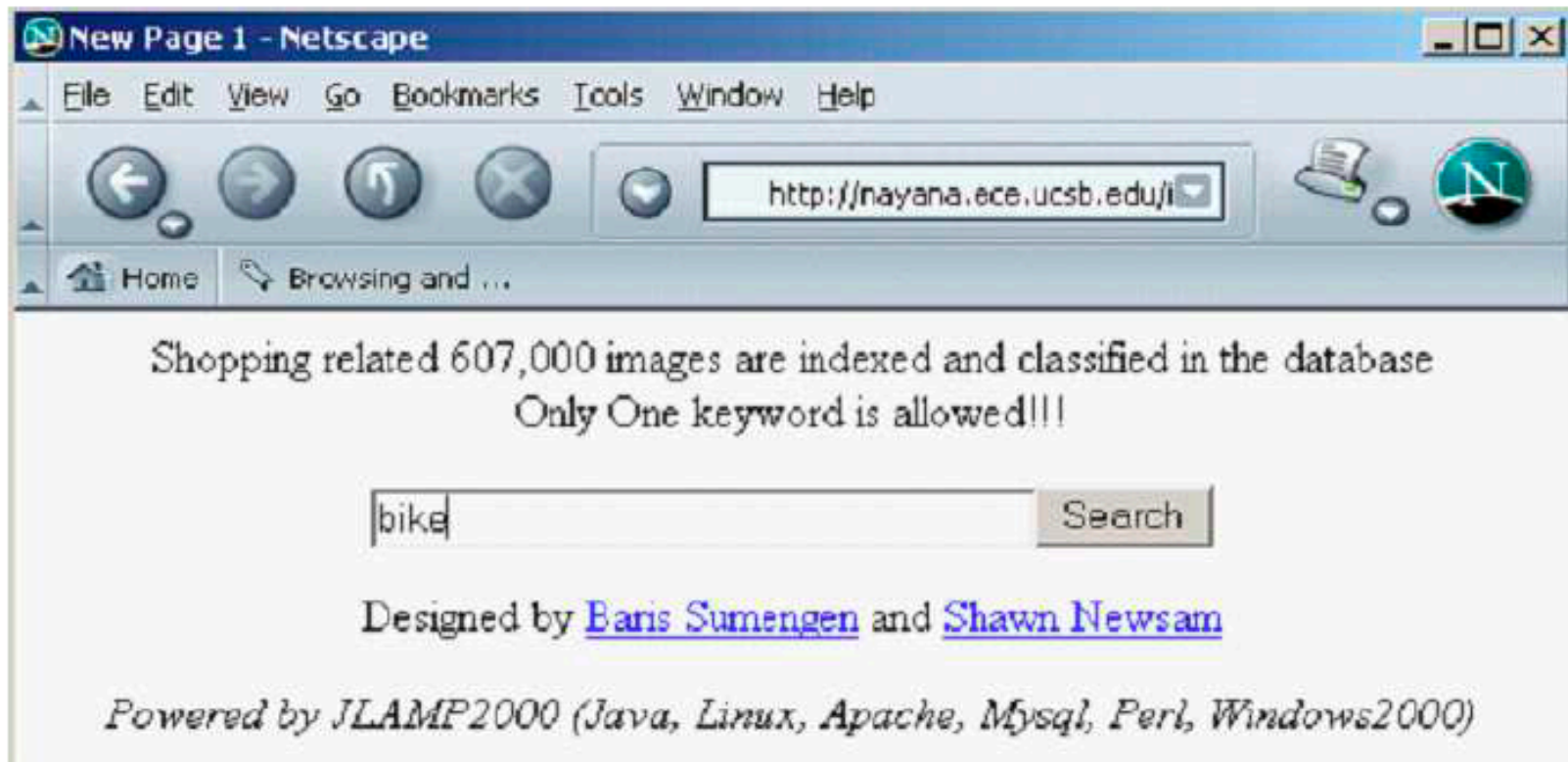


Relevance Feedback

- Principal ideia
 - Selecionar termos importantes em documentos relevantes
 - Aumentar a importância desses termos em uma nova consulta
- Efeito esperado: a nova consulta vai se mover em direção aos documentos relevantes e mais distante dos não relevantes
- Esconde do usuário detalhes do refinamento















Relevance Feedback: Exemplo1





Resultado para a Consulta Inicial

Interface for a search results page showing a grid of images related to bicycles and motorcycles. The interface includes navigation buttons at the top: Browse, Search, Prev, Next, and Random.













					
(144473, 16458)	(144457, 252140)	(144456, 262851)	(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, 204044)	(144483, 265133)	(144518, 237752)	(144538, 525937)	(144450, 249611)	(144450, 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



Feedback do Usuário: Selecciona o que é Relevante

Interface for user feedback on relevant items, displaying a grid of images and their associated coordinates and scores.










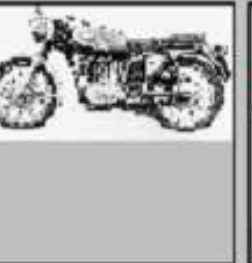


Buttons: Browse, Search, Prev, Next, Random

Image	Coordinates	Scores
	(144473, 16450)	0.0 0.0 0.0
	(144457, 252149)	0.0 0.0 0.0
	(144456, 262057)	0.0 0.0 0.0
	(144455, 262063)	0.0 0.0 0.0
	(144457, 252134)	0.0 0.0 0.0
	(144403, 265154)	0.0 0.0 0.0
	(144403, 264644)	0.0 0.0 0.0
	(144403, 265153)	0.0 0.0 0.0
	(144510, 257752)	0.0 0.0 0.0
	(144539, 525037)	0.0 0.0 0.0
	(144456, 249611)	0.0 0.0 0.0
	(144456, 250064)	0.0 0.0 0.0



Resultado depois do Feedback

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

 (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) 0.64301 0.351395 0.293615	 (144456, 253568) 0.650275 0.411745 0.23853	 (144538, 523799) 0.66709197 0.358033 0.309059
 (144473, 16249) 0.6721 0.393022 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



Exemplo em Texto

- Consulta inicial: new space satellite applications

	<i>r</i>		
+	1	0.539	NASA Hasn't Scrapped Imaging Spectrometer
+	2	0.533	NASA Scratches Environment Gear From Satellite Plan
	3	0.528	Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
	4	0.526	A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget
	5	0.525	Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
	6	0.524	Report Provides Support for the Critics Of Using Big Satellites to Study Climate
	7	0.516	Arianespace Receives Satellite Launch Pact From Telesat Canada
+	8	0.509	Telecommunications Tale of Two Companies



Consulta Expandida depois do Relevance Feedback

2.074	new	15.106	space
30.816	satellite	5.660	application
5.991	nasa	5.196	eos
4.196	launch	3.972	aster
3.516	instrument	3.446	arianespace
3.004	bundespost	2.806	ss
2.790	rocket	2.053	scientist
2.003	broadcast	1.172	earth
0.836	oil	0.646	measure

- Consulta inicial: new space satellite applications



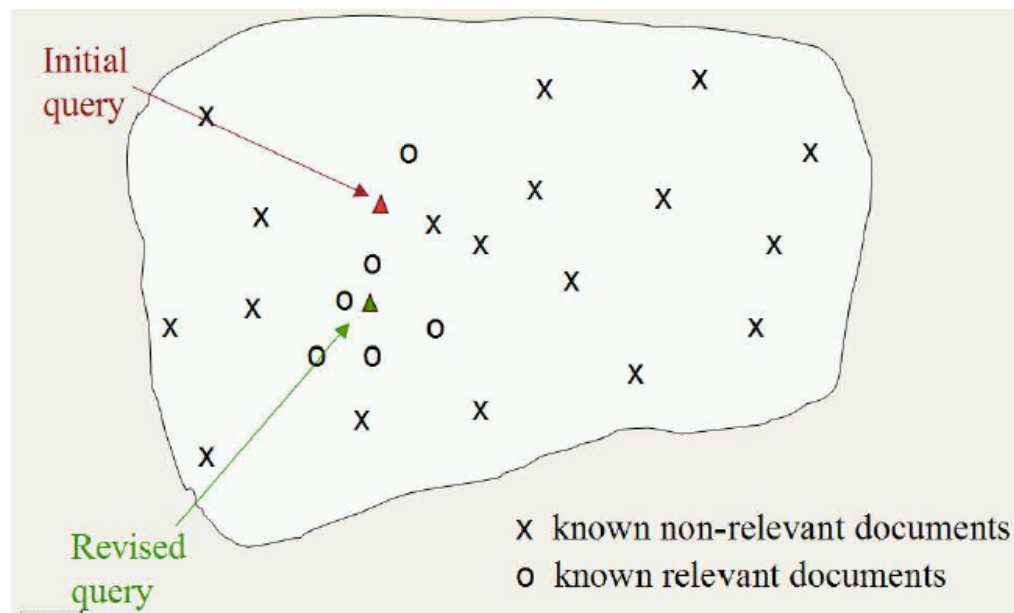
Resultados para a Consulta Expandida

<i>r</i>			
*	1 (2)	0.513	NASA Scratches Environment Gear From Satellite Plan
*	2 (1)	0.500	NASA Hasn't Scrapped Imaging Spectrometer
	3	0.493	When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
	4	0.493	NASA Uses 'Warm' Superconductors For Fast Circuit
*	5 (8)	0.492	Telecommunications Tale of Two Companies
	6	0.491	Soviets May Adapt Parts of SS-20 Missile For Commercial Use
	7	0.490	Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers
	8	0.490	Rescue of Satellite By Space Agency To Cost \$90 Million



Algoritmo Rocchio

- Implementa relevance feedback no vector space model
- Assume: vetores dos documentos relevantes são similares e não relevantes não são similares
- Ideia: reformular a consulta para que
 - Se aproxime dos documentos relevantes
 - Se afaste dos não relevantes





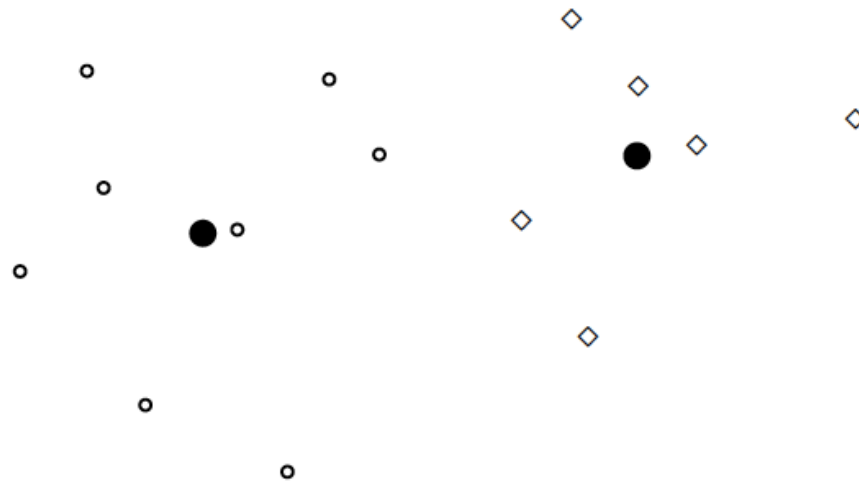
Algoritmo Rocchio

- Formula:
$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

Centroide
dos relevantes

Centroide
dos não relevantes

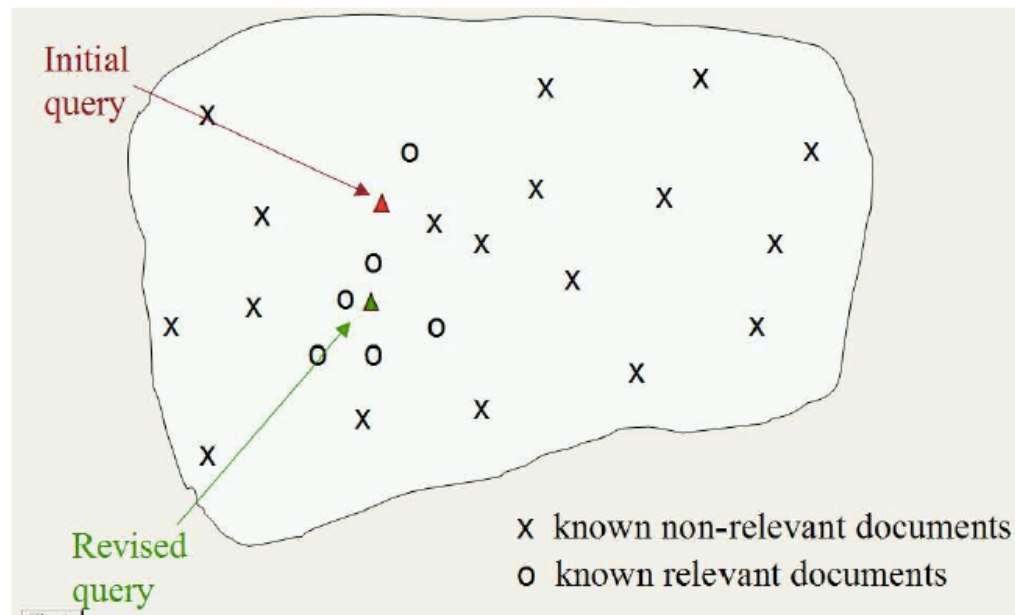
- Centroide: vetor médio





Algoritmo Rocchio

- Formula:
$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$
- Novas consultas se movem em direção aos documentos relevantes e se distanciam dos não relevantes





Algoritmo Rocchio

- Formula: $\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$
- Novas consultas movem em direção aos documentos relevantes e longe do não relevantes
- Coloca pesos nos termos negativos
- Feedback positivo é mais valioso
- Por exemplo: . $\beta=0.75$ e $\gamma=0.25$ dá maior peso ao feedback positivo
- A maioria dos sistemas usa apenas o feedback positivo



Relevance Feedback: Limitações

- Suposição: usuário sabe os termos na coleção para criar a consulta inicial
 - Violação: usuário usa vocabulário diferente do da coleção
 - Ex: macaxeira / aipim
- Problema
 - Cria consultas longas
 - Usuários são relutantes de prover feedback explícito



Pseudo-relevance Feedback

- Automatiza a parte manual da avaliação dos resultados
- Algoritmo
 - Recupera uma lista de resultados ranqueados
 - Assume que os top k resultados são relevantes
 - Executa relevance feedback
- Funciona bem na média
- Problema: top k de baixa qualidade
 - Query drift: saindo do tópico
 - Após várias iterações pode haver query drift



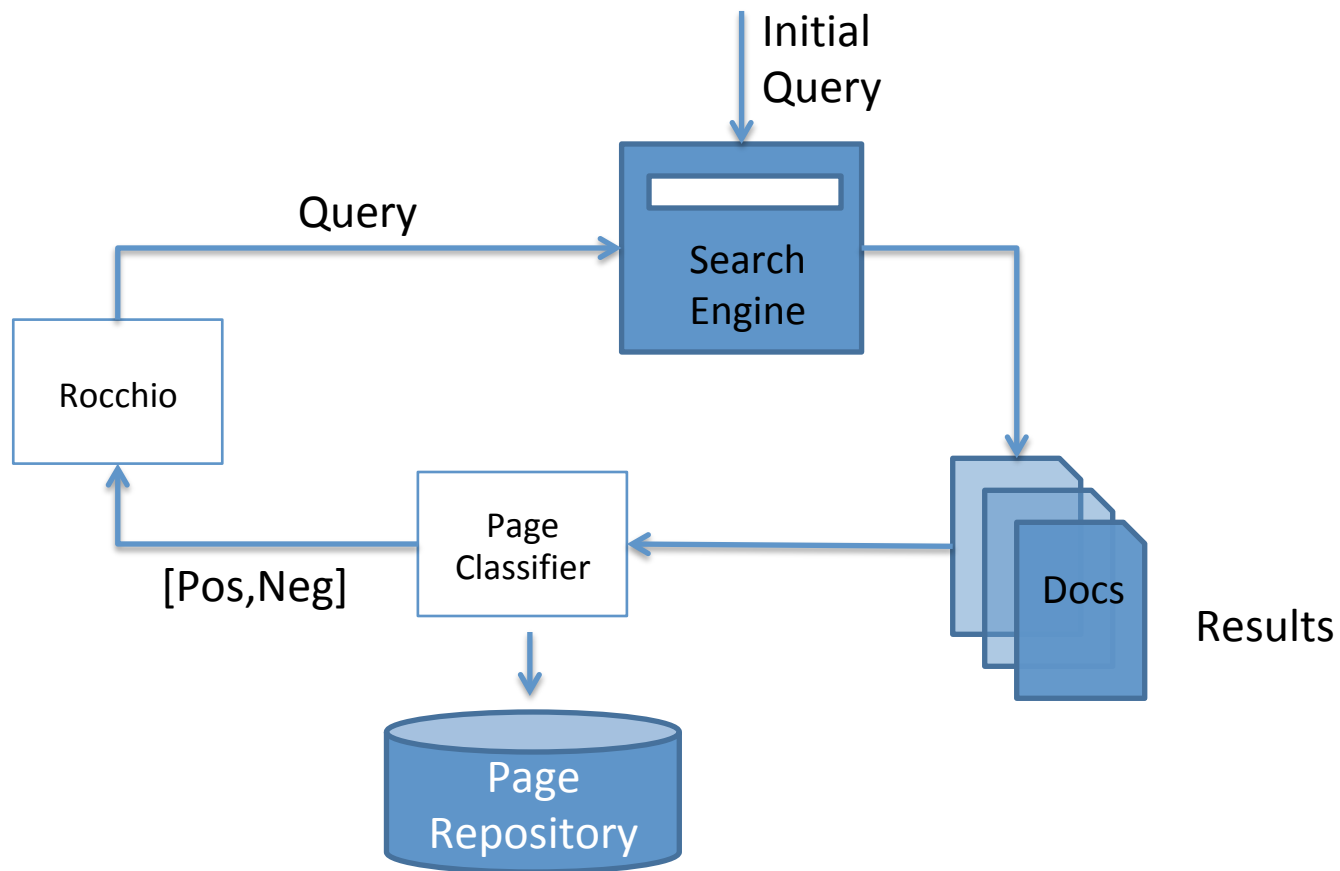
Pseudo-relevance Feedback na TREC4

- Cornell SMART system
- Número de documentos relevantes do top 50

Term weighting	Precision at $k = 50$	
	no RF	pseudo RF
Inc.ltc	64.2%	72.7%
Lnu.ltu	74.2%	87.0%



Pseudo-Relevance Feedback para Focused Crawlers





Refinamento baseado em Log de Consultas

YAHOO! SEARCH

Web | [Images](#) | [Video](#) | [Audio](#) | [Directory](#) | [Local](#) | [News](#) | [Shopping](#) | [More »](#)

[Answers](#) | [My Web](#) | [Search Services](#) | [Advanced Search](#) | [Preferences](#)

Search Results 1 - 10 of about 160,000,000 for **palm** - 0.07 sec. ([About this page](#))

Also try: [palm springs](#), [palm pilot](#), [palm trees](#), [palm reading](#) [More...](#)

SPONSOR RESULTS

- [Official Palm Store](#)
[store.palm.com](#) Free shipping on all handhelds and more at the official **Palm** store.
- [Palms Hotel - Best Rate Guarantee](#)
[www.vegas.com](#) Book the **Palms** Hotel Casino with our best rate guarantee at VEGAS.com, the official Vegas travel site.


SPONSOR RESULTS

[Palm Memory](#)
Memory Giant is fast and easy. Guaranteed compatible memory. Great...
[www.memorygiant.com](#)

[The Palms, Turks and Caicos Islands](#)
Resort/Condo photos, rates, availability and reservations....
[www.worldwidereservationsystems.c](#)

[The Palms Casino Resort, Las Vegas](#)
Low price guarantee at the **Palms** Casino resort in Las Vegas. Book...
[lasvegas.hotelscorp.com](#)

Y! [Palm Pilots](#) - [Palm Downloads](#)
[Yahoo! Shortcut](#) - [About](#)

1. [Palm, Inc.](#) 
Maker of handheld PDA devices that allow mobile users to manage schedules, contacts, and other personal and business information. Category: [B2B > Personal Digital Assistants \(PDAs\)](#)
[www.palm.com](#) - 20k - [Cached](#) - [More from this site](#) - [Save](#)



Baseado em Log de Consultas

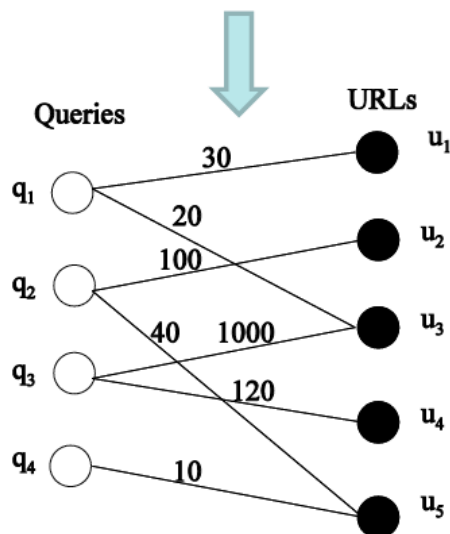
- Exemplo 1: após procurar por “ervas”, usuários frequentemente procuram por “ervas medicinais”
- Exemplo 2:
 - Usuários buscando por “flower pix” frequentemente clicam na URL: photobucket.com/flower
 - Usuários buscando por “flower clipart” também



Cliques em Grafo Bipartido

- Consulta representada por um vetor de pesos de acordo com as URLs

User ID	Time Stamp	Event Type	Event Value
User 1	20071205110843	QUERY	KDD 2008
User 2	20071205110845	CLICK	www.aaa.com
User 1	20071205110848	CLICK	www.kdd2008.com
...



$$\vec{q}_i[j] = \begin{cases} \frac{w_{ij}}{\sqrt{\sum_{\forall e_{ik}} w_{ik}^2}} & \text{if } e_{ik} \text{ exists} \\ 0 & \text{otherwise} \end{cases}$$

$$Q_1 = [1, 0, 0.019, 0, 0]$$

$$Q_3 = [0, 0, 0.981, 1, 0]$$

$$\text{dist}(q_i, q_j) = \sqrt{\sum_{u_k} (\vec{q}_i[k] - \vec{q}_j[k])^2}$$