Diz que é uma espécie de *survey* sobre Query Expansion

Nuno Cardoso

Orientadores:
Diana Santos e Mário J. Silva

Simpósio Doutoral da Linguateca
30 de Março de 2007

FCUL

O que é *Query Expansion* (QE)?

- Adição de novos termos aos termos iniciais do utilizador, para definir melhor os conceitos por detrás da sua necessidade de informação (NI).
- Aumenta a probabilidade de encontrar documentos relevantes com termos comuns (Xu & Croft, 1996).
- Diminuição do 'fosso' semântico entre a pesquisa e os documentos.

Tipos e fontes de QE

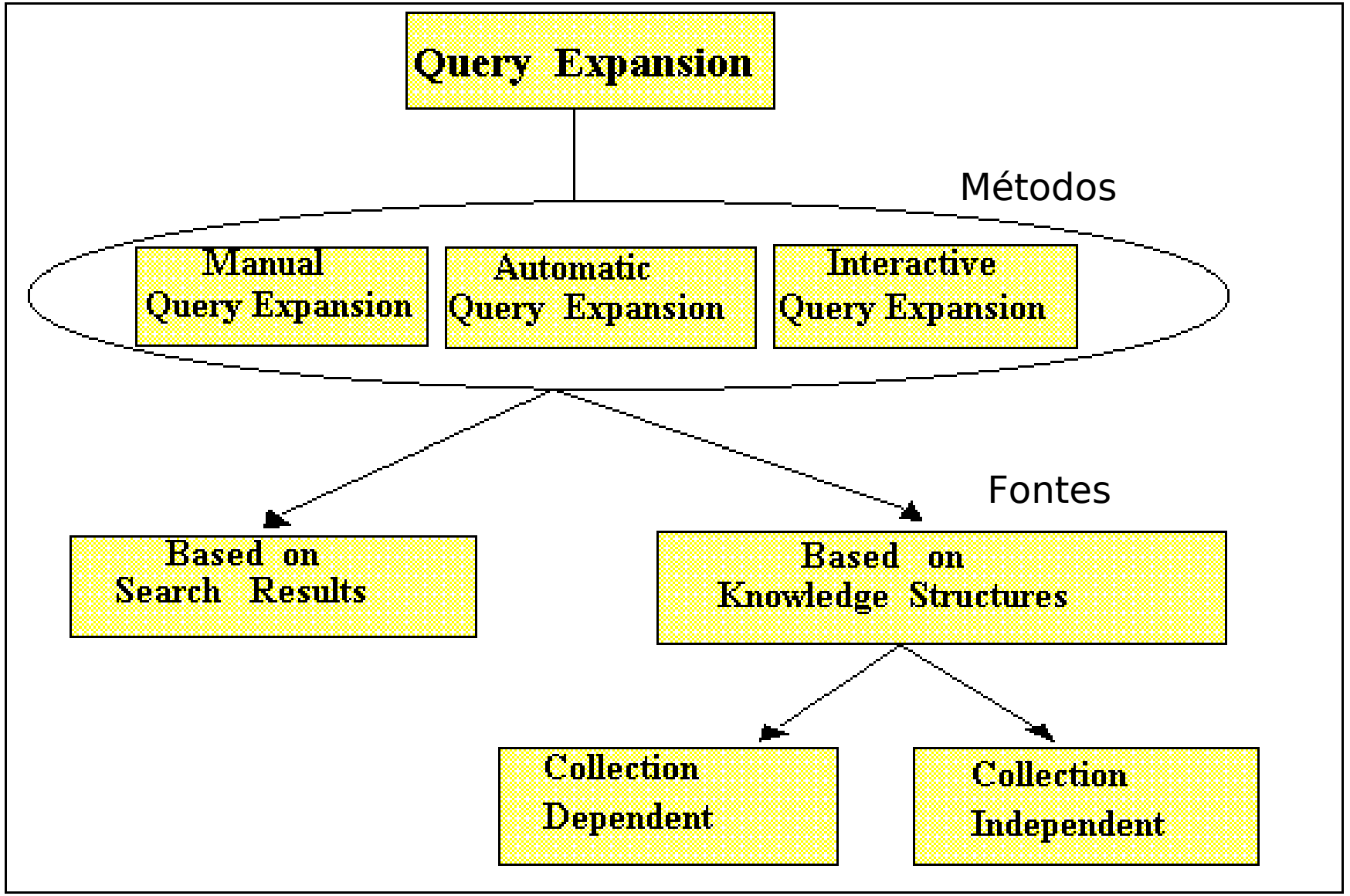
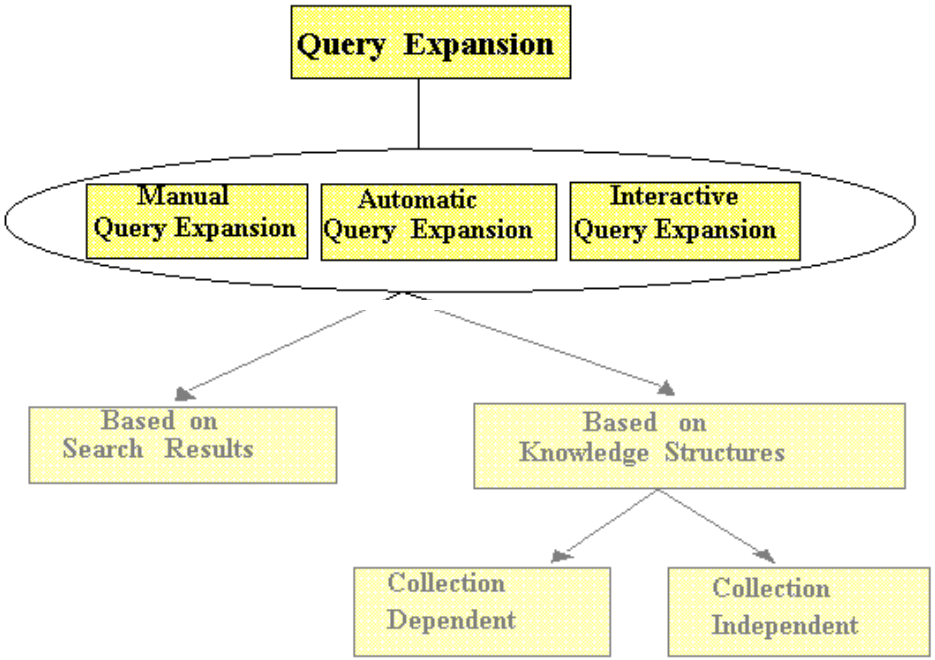


Imagem retirada de E. Efthimiadis, 'Query Expansion'

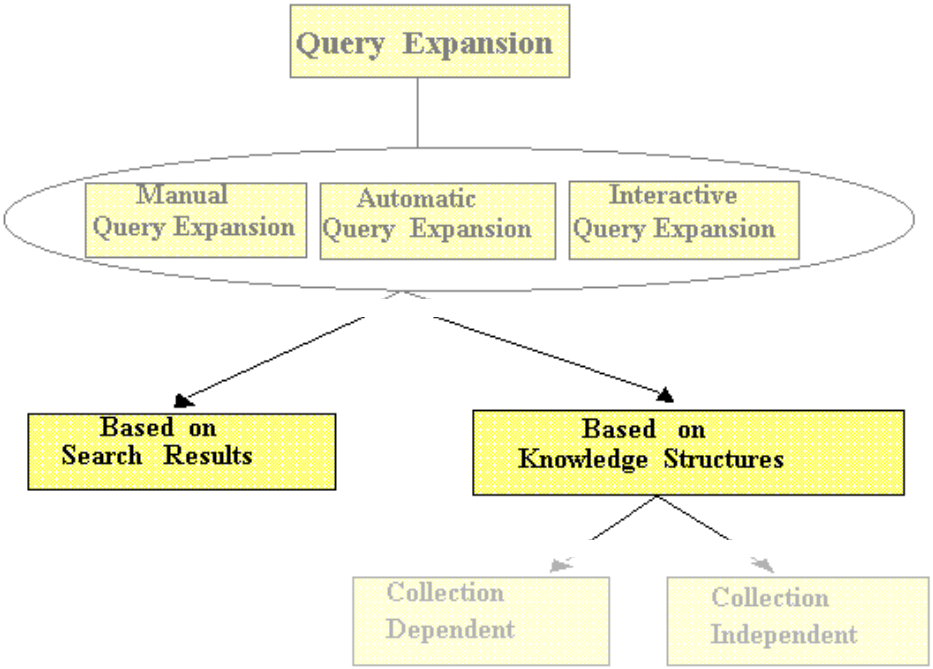
Métodos de QE

- Manual: feita pelo utilizador.
- Automática: feita pelo sistema.
- Interactiva:
  - Utilizador auxiliado pelo sistema.
  - Sistema auxiliado pelo utilizador.



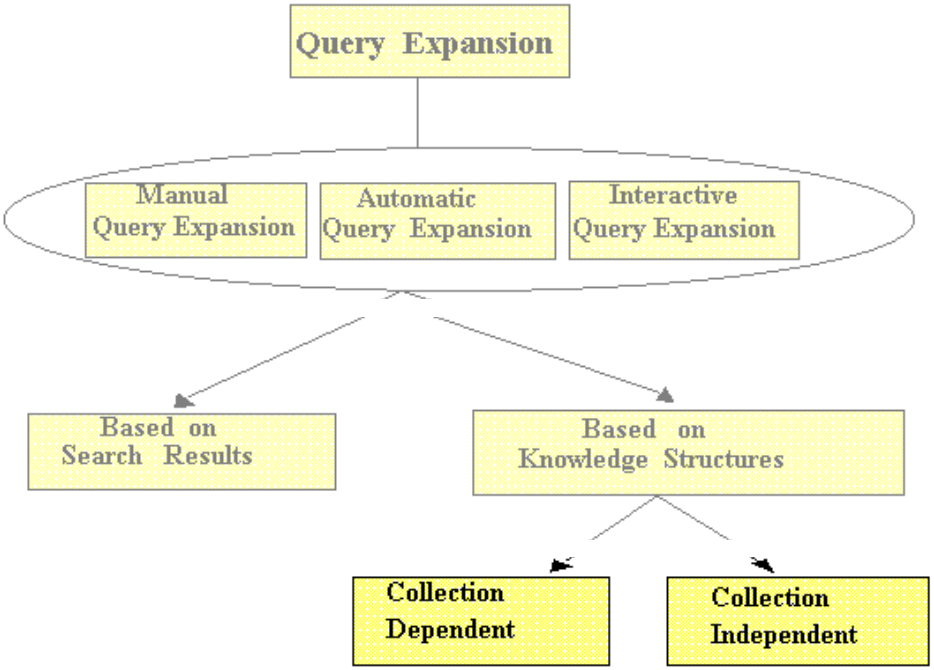
Fontes de QE

- Baseado nos resultados:
  - Processos de *Relevance Feedback.*
- Baseados em Estruturas de dados:
  - Outras fontes de informação. independentes do processo de consulta.



Fontes de QE

- Recursos baseados em Colecções independentes:
  - Tesauros genéricos (WordNet).
  - Dicionários / léxicos.
  - Ontologias.
- Recursos baseados em Colecções dependentes.
  - Tesauros e outros recursos construídos a partir da colecção.
  - Colecções Web: diários de pesquisas.



Estado da Arte em QE:

- Tesauros em IR: "*Any data structure that defines semantic relatedness between words*" (Schutze & Pedersen, 1997; McGettrick)
- Trabalho remonta a Luhn, nos anos 50.
- A intenção é boa...
  - pintura -> quadro, tela, etc.
- Mas tesauros manuais são difíceis de desenvolver e manter; e podem não servir para as NI dos utilizadores da web.

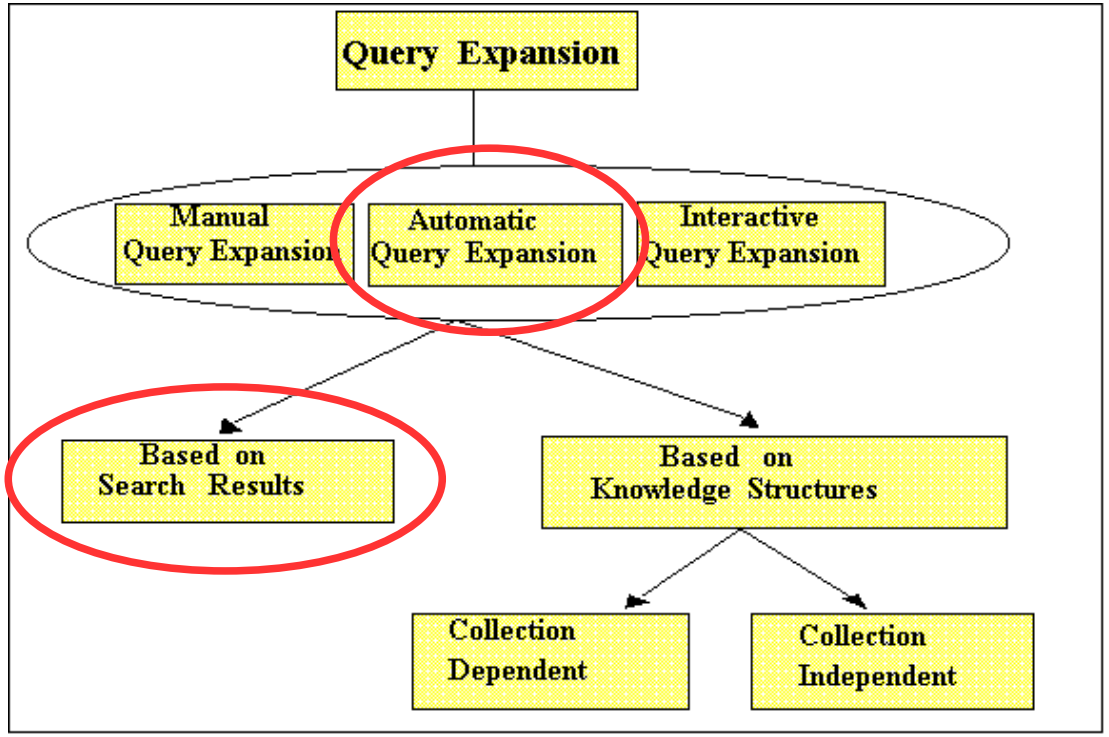
Estado da Arte em QE:

- Tesauros automáticos:
  - Muito trabalho publicado sobre extracção de co-ocorrências, cálculo de semelhanças entre termos, *clustering, latent semantic indexing*...
  - Muitos resultados encorajadores, como o de Qiu e Frei (1993) mas...
  - Xu e Croft (1996) mostram que QE a partir dos resultados de uma consulta inicial (*local analysis*) é mais eficiente do que QE que analise o corpus e que extraia relações entre termos (*global analysis*) .
  - Misturando os dois (*local context analysis*) ainda é melhor...

Estado da Arte em QE:

Em resumo:

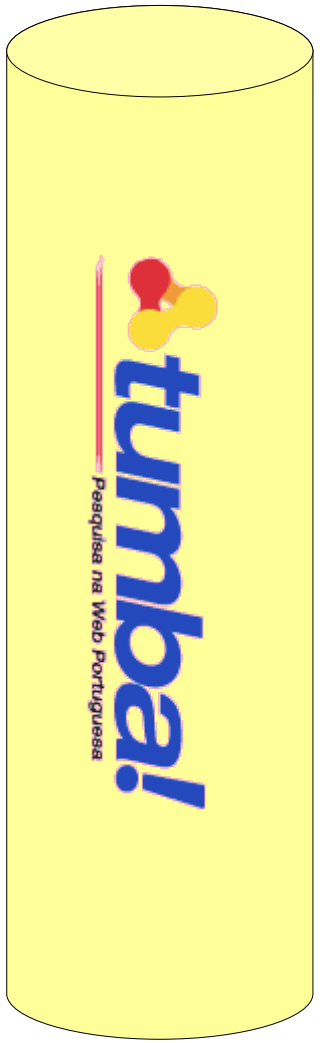
- O método mais usado: **QE automático**
- A fonte mais usada: **Resultados da consulta**



? Sistema típico sem QE

Utilizador
insatisfeito

pinturas italianas



Resultados finais

:(

Utilizador
menos insatisfeito

Sistema típico de QE automático

Utilizador insatisfeito

pinturas italianas

Relevantes

Irrelevantes

Resultados iniciais

desenhos, exposição, arte, século, artistas, museu, grande, desenhos, exposição, arte, século,...

pinturas italianas desenhos exposição arte século artistas museu grande desenhos exposição arte século...

Resultados finais

tumba! Pesquisa na Web Portuguesa

:D

Utilizador satisfeito!

Exemplo: XLDB @ CLEF 2006

- Tópico 303:

<top>
<num> C303 </num>
<PT-title> **Pinturas italianas** </PT-title>
<PT-desc> Encontrar informação sobre **locais** onde **pinturas ou desenhos italianos** estão expostos ao público. </PT-desc>
<PT-narr> Documentos relevantes devem mencionar os locais em qualquer parte do mundo onde estão expostas permanente ou temporariamente pinturas da escola italiana ou de autores italianos. A localização de pinturas ou desenhos por artistas italianos, expostos ao público em museus, galerias de arte, ou similares, também interessa. A informação deve ser suficiente para identificar o local exacto, ou seja, o nome da cidade ou do país não é suficiente. </PT-narr>
</top>

Exemplo: XLDB @ CLEF 2006

- Tópico 348:

<top>
<num> C348 </num>
<PT-title> **Assassinato de Yann Piat** </PT-title>
<PT-desc> Encontrar documentos discutindo o assassinato de Yann Piat, activista política de direita, em 1994.
</PT-desc>
<PT-narr> Documentos relevantes devem conter detalhes do homicídio de Yann Piat perto de sua casa em Toulon.
</PT-narr>
</top>

Exemplo: XLDB @ CLEF 2006

- Um utilizador típico de motores de busca, normalmente, usa dois termos [Spink et al, 2002]
- Consulta provável do utilizador:
  - pinturas italianas
  - assassinato Yann Piat
- Consultas iniciais do XLDB:
  - pinturas italianas OR pinturas italiana OR pintura italianas OR pintura italiana
  - assassinato Yann Piat

Exemplo: XLDB @ CLEF 2006

- Tópico 303:
  - [pinturas italianas] + [desenhos, exposição, arte, século, artistas, museu, grande, obras, pintor, esculturas, mestres, artista, artes, trabalhos, mostra, vida, fellini, desenho, historia, brasileiros, escultura, cor, parte, 50, livros, trabalho, contemporânea, américa, individual, cinema, pintores, galerias]
- Tópico 348:
  - [assassinato Yann Piat] + [deputada, hyeres, jijel, carpizo, drogas, legalizou, antidroga, toulon, molestamento, orfanato, renuncie, gloucester, governo, rosemary, pistoleiros, pais, viajavam, liderava, francesa, contra, pietro, duas, frederick, soltos, plebiscito, sinn, fein, afeganistao, doar, argelinos, horrores, viciados]

Exemplo: XLDB @ CLEF 2006

tópico 303	Sem QE	Com QE (32 termos)
Docs relevantes	50	50
Docs recuperados	228	1000
Docs rel e rec.	28	36
Precisão @10	30%	0%
Abrangência	56%	72%
MAP	**0,1515**	**0,0519**

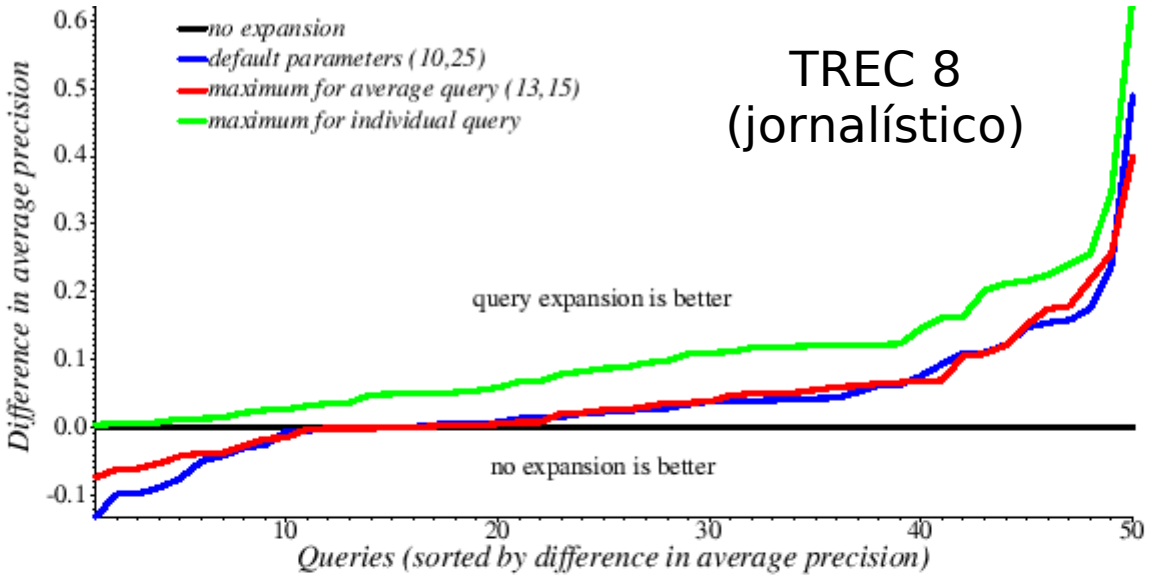
tópico 348	Sem QE	Com QE (32 termos)
Docs relevantes	8	8
Docs recuperados	2	1000
Docs rel e rec.	2	8
Precisão @10	20%	70%
Abrangência	25%	100%
MAP	**0,2500**	**0,9276**

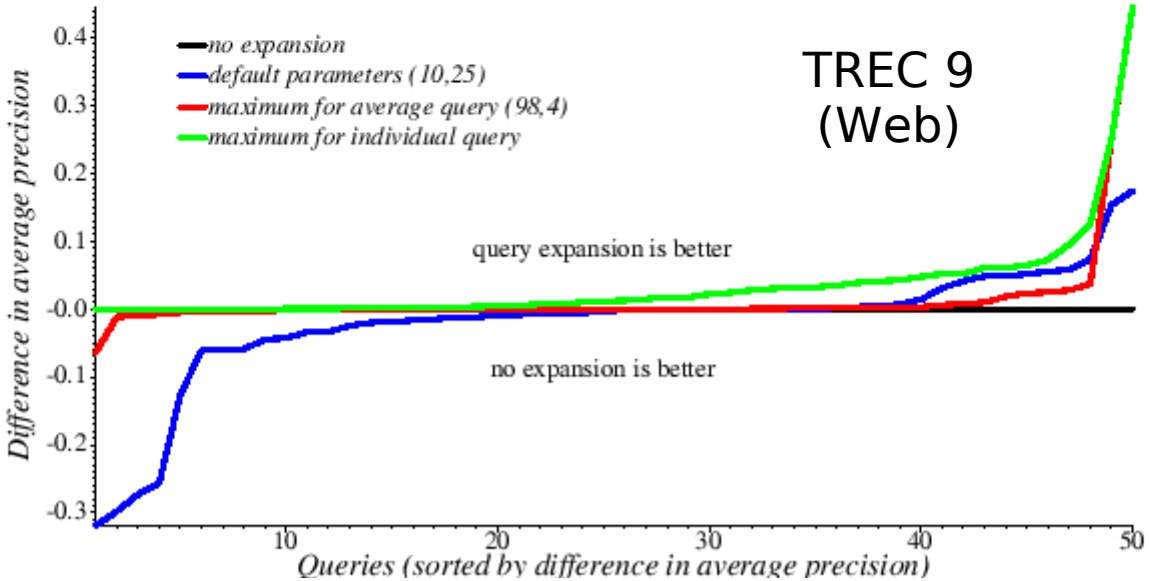
É bonito, mas...

- Há uma minoria de tópicos que são prejudicados pela QE (*query drift*).
- QE depende muito:
  - da colecção usada como fonte de informação
  - de um sistema RI que retorne bons documentos iniciais
  - de uma boa optimização dos parâmetros.
- ...e (digo eu) uma aproximação muito adaptada ao ambiente de avaliação usado.
- Quais os pontos críticos na Web?

É bonito, mas...

- QE automático funciona bem em avaliações **ad-hoc sobre textos jornalísticos;** para colecções web, os resultados não são brilhantes





Imagens retiradas de Billerbeck [2005]

Pontos críticos

Utilizador insatisfeito

pinturas italianas

Relevantes

Irrelevantes

Resultados iniciais

- Consultas ainda mais vagas, sobre tudo
- Impacientes e exigentes
-

desenhos, exposição, arte, século, artistas, museu, grande, desenhos, exposição, arte, século,...

pinturas italianas desenhos exposição arte século artistas museu grande desenhos exposição arte século...

Resultados finais

Utilizador satisfeito!

Pontos críticos

Utilizador insatisfeito

pinturas italianas

Colecções maiores
Restrições booleanas
Ordenação resultados

Relevantes

Irrelevantes

Resultados iniciais

desenhos, exposição, arte, século, artistas, museu, grande, desenhos, exposição, arte, século,...

pinturas italianas desenhos exposição arte século artistas museu grande desenhos exposição arte século...

tumba! Pesquisa na Web Portuguesa

Resultados finais

Utilizador satisfeito!

Pontos críticos

Utilizador insatisfeito

pinturas italianas

Relevantes

Irrelevantes

Resultados iniciais

desenhos, exposição, arte, século, artistas, museu, grande, desenhos, exposição, arte, século...

Escolha dos documentos "relevantes": quais? quantos? qual a fonte?

pinturas italianas desenhos exposição arte século artistas museu grande desenhos exposição arte século...

tumba! Pesquisa na Web Portuguesa

Resultados finais

Utilizador satisfeito!

Pontos críticos

Utilizador insatisfeito

pinturas italianas

Relevantes

Irrelevantes

Resultados iniciais

tumba! Pesquisa na Web Portuguesa

Escolha dos novos termos
- ordenação
- pesagem
- *stemming*
- EM / EMP
- fontes externas

desenhos, exposição, arte, século, artistas, museu, grande, desenhos, exposição, arte, século,...

pinturas italianas desenhos exposição arte século artistas museu grande desenhos exposição arte século...

Resultados finais

Utilizador satisfeito!

Pontos críticos

Utilizador insatisfeito

pinturas italianas

Relevantes

Irrelevantes

Resultados iniciais

tumba! Pesquisa na Web Portuguesa

Combinação dos termos

desenhos, exposição, arte, século, artistas, museu, grande, desenhos, exposição, arte, século,...

pinturas italianas desenhos exposição arte século artistas museu grande desenhos exposição arte século...

Resultados finais

Utilizador satisfeito!

Em resumo:

- QE para avaliações *ad-hoc* são um começo, mas é preciso repensar QE para uma colecção *web*:
  - explorar fontes de informação adicionais
    - diários de pesquisa
    - textos das âncoras (Wang & Tanaka, 2006)
    - ontologias geográficas
    - 'desktop' do utilizador
  - analisar o tipo de consulta inicial
    - consulta para uma página (ex: “FCUL”)
    - consulta geográfica (ex: “pinturas em Lisboa”)
    - consultas de tópicos (ex: “técnicas pintores”)

Diários de Pesquisas

- Billerbeck [2005], na sua tese, concluiu que:
  - um módulo QE automático consegue bons resultados para colecções jornalísticas (TREC 8), mas nada brilhantes para colecções web (TREC 9-10).
  - Ao usar os diários de pesquisa (*Query Association*), obtém-se melhores resultados (26-29%) numa colecção web (TREC-10) do que sem expansão
- É só a ponta do *iceberg*.

Notas soltas

- Um sistema de RI com base em aproximações estatísticas funciona bem num tópico, funciona mal no próximo.
- QE funciona ainda melhor num tópico, ainda pior noutro tópico.
- Um recurso / fonte de informação para QE não chega!
- É necessário escolher o recurso mais adequado para QE, a partir do tipo de consulta (ex: consultas geográficas)

Notas soltas

- QE para a Web precisa de ser útil, mas também rápido
  - QE eficaz e eficiente
  - Novos índices
- QE dá-se mal com o modelo booleano (Kekalaiken & Jarvelin, 1998)
- Conciliar o modelo booleano (Web) com os modelos probabilísticos (colecções jornalísticas) (Yoshioka & Haraguchi, 2005)
- E os modelos linguísticos?
- NLP: REM, Sumarização, EI, ...

Não sou só eu...

- Allan (2002), “Challenges in Information Retrieval and Language Modeling”
- NIST, em 2003, organizou um workshop – Reliable Information Access (RIA) para estudar os motivos de falha dos sistemas de RI actuais.
- SIGIR 2004 workshop: "Where can IR go from here?” (Harman & Buckley, 2004)
- Query Clarity Score [Cronen-Townsend et al, 2002]
- Pistas-piloto do NTCIR

Avaliação ~~HARE~~... QE!

- Como avaliar os passos intermédios de um módulo de QE?
- Como avaliar a 'utilidade' de cada fonte de informação?
- NTCIR-5 WEB Query Term Expansion pilot task (Yoshioka, 2005) (http://research.nii.ac.jp/ntcweb/cfp-ntcir5web-q-en.html)
  - sem descrição de tarefas nem critérios de avaliação, mas com algumas propostas interessantes
  - não está presente no NTCIR-6 e 7...

Avaliação ~~HARE~~... QE!

- avaliação *user-oriented*:
  - Pedir aos 'juízes' para avaliar termos expandidos interactivamente
  - Pedir aos 'juízes' para seleccionar termos a partir de documentos relevantes.
- avaliação a partir de um sistema RI de referência
  - Com base em informação estatística (ex: fórmula Robertson/Sparck-Jones no conjunto de documentos relevantes)
  - Termos 'orientados' para prec. ou para abr.
  - Desempenho do sistema

Referências

- Shao-Chi Wang* and Yuzuru Tanaka, Topic-Oriented Query Expansion for Web Search, WWW 2006, May 23-26, 2006, Edinburgh, Scotland. 2006
- Xu, J. and Croft, W. B. 1996. Query expansion using local and global document analysis. In Proceedings of the 19th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Zurich, Switzerland, August 18 - 22, 1996).
- A. Spink and B. Jansen, A Study of Web Search Trends, Webology, Volume 1, Number 2, December, 2004
- E. Efthimiadis, 'Query Expansion', ARIST, v31, pp. 121-187, 1996
- Bodo Billerbeck. PhD, "Efficient Query Expansion"., RMIT University, Melbourne, Austrália, 2005
- Smyth, B., Freyne, J., Coyle, M., Briggs, P., Balfe, E. (2003) I-SPY: Anonymous, Community-Based Personalization by Collaborative Web Search. Proceedings of the 23rd SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence. Oxford, UK.
- Sean McGettrick, 'Query Expansion', www.ist.psu.edu/faculty_pages/giles/IST497/presentations/McGettrick.ppt
- Schütze, H. and Pedersen, J. O. 1997. A cooccurrence-based thesaurus and two applications to information retrieval. Inf. Process. Manage. 33, 3 (May. 1997), 307-318.
- Qiu, Y. and Frei, H. 1993. Concept based query expansion. In Proceedings of the 16th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Pittsburgh, Pennsylvania, United States, June 27 - July 01, 1993). R. Korfhage, E. Rasmussen, and P. Willett, Eds. SIGIR '93. ACM Press, New York, NY, 160-169.
- Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pages 299--306. ACM Press, 2002.
- Masaharu Yoshioka. Introduction for Evaluation Results of the NTCIR-5 WEB Query Term Expansion Subtask. Proceedings of the NTCIR-5, 2005
- M. Yoshioka and M. Haraguchi, On a Combination of Probabilistic and Boolean IR Models for WWW Document Retrieval. ACM Transactions on Asian Language Information Processing, Vol. 4, No. 3, September 2005, Pages 340–356.
- J. Kekalainen and K. Jarvelin, The impact of query structure and query expansion on retrieval performance. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 130–137.

Diz que é uma espécie de *survey* sobre Query Expansion

Nuno Cardoso

Orientadores:
Diana Santos e Mário J. Silva

Simpósio Doutoral da Linguateca
30 de Março de 2007

FCUL

Um caso de estudo: I-Spy

- I-Spy [Smyth, 2003] é um meta-motor de busca comunitária, que regista as interacções entre os utilizadores e os resultados. http://ispy.ucd.ie

I-Spy: 'pinturas italianas'

Oops..


Communities
About I-Spy
the smarter way to search
pinturas italianas
Search
default: Your Search for pinturas italianas returned 39 Results | Displaying 1 - 20 | Result Page: 1 2 Next
Related Information
I-Spy Recommends
Recent Queries
VIEW ALL
1. sonicstage 2.0 full dow...
2. myspace
3. caroline murphy
4. meta-search
5. sonicstage 2.0 download
Recent Web Pages
VIEW ALL
1. Re: Download SonicStage...
2. MySpace.com
3. MySpace
4. AskMen.com - Carolyn Mu...
5. Metacrawlers and Metase...
Popular Queries
VIEW ALL
Addio Gallery
Exhibits several paintings by famous Renaissance artists.
http://www.mcs.csuhayward.edu/~malek/Addio.html
Lotto, pinturas italianas del alto renacimiento, frescos de iglesi
Obras de Lorenzo Lotto, artista del renacimiento italiano, pintor de frescos famosos y breve historia de su vida.
http://pintoresfamosos.juegofanatico.cl/lotto.htm
museo/pinturas
PINTURAS ITALIANAS Y ESPAÑOLAS. La Capilla Real guarda un significativo elenco de obras salidas de la paleta