Automatic Query Reformulation using Contextual Information

Por: Nuno Cardoso

Orientadores:
Diana Santos e Mário J. Silva

Seminário Doutoral da Linguateca
3 e 4 de Outubro de 2006

Faculdade de Ciências
Universidade de Lisboa

Introdução

- Na última década, a Internet e RI colocaram o mundo à distância de um *click* de rato.
- Quais os desafios a RI para a próxima década?

Introdução

- Na última década, a Internet e RI colocaram o mundo à distância de um *click* de rato.
- Quais os desafios a RI para a próxima década?
  - RI para outras aplicações (pesquisas em clientes mails, pesquisas no PC,...)
  - RI Multimédia, RI Geográfico, IR Multilingue,...
  - Novas estratégias para melhorar os resultados das pesquisas

O utilizador típico de sistemas de RI...

- ...usa uma média de 2 termos por pesquisa, e espera resultados bons e imediatos.
- ...tem dificuldade, por vezes, de descrever uma necessidade de informação (NI) em lógica booleana.
- ...escolhe mal os termos, obtendo resultados irrelevantes.
- ...reformula a pesquisa (ou a NI) até ficar satisfeito... ou desistir.

Como ajudar o utilizador

- Os sistemas de RI actuais podem usar **informação contextual** para:
  - Assistir o utilizador na descrição da sua necessidade de informação.
  - Capturar a NI latente nos termos escolhidos.
  - Aprender com casos de pesquisas semelhantes.
  - Reformular automaticamente as pesquisas.
  - Restringir os âmbitos das pesquisas.

Hipótese

Automatic Query Reformulation using Contextual Information

"By **researching** new data-mining and NLP methods on alternative data sources, queries can be **reformulated** to best fit users' information needs, **improving** IR results."

Reformulação de Pesquisas (Query Reformulation)

Query Reformulation (QR) =

Query Expansion (QE)+Term re-weighting(TR)

(selecciona e adiciona termos) (pesa os termos)

Reformulação de Pesquisas (Query Reformulation)

Query Reformulation (QR) =

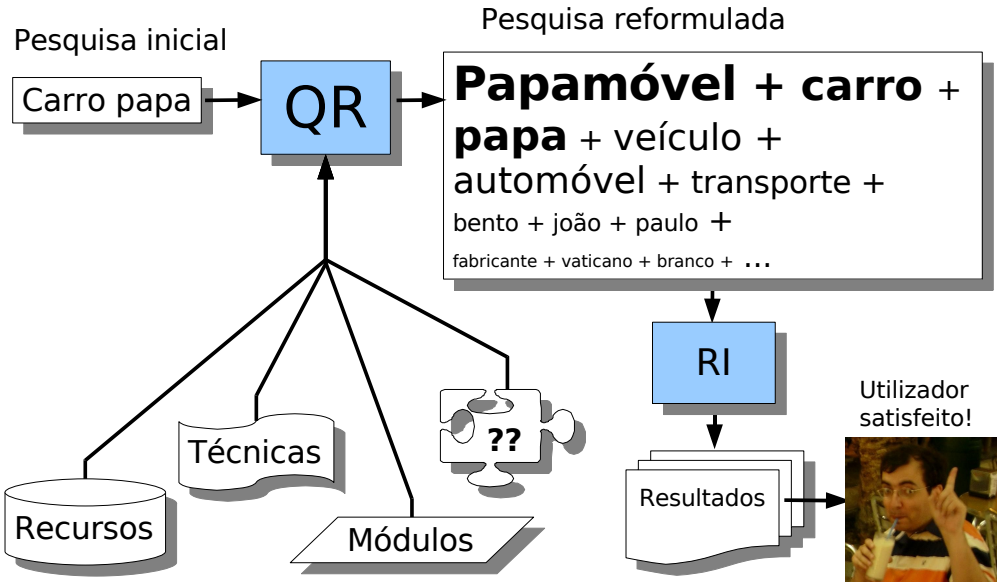
Query Expansion (QE)+Term re-weighting(TR)

(selecciona e adiciona termos) (pesa os termos)

- Ao adicionar mais termos relacionados, aumenta-se as probabilidades de encontrar termos da pesquisa em documentos relevantes.
- Ao pesar os termos, dá-se um critério de importância adicional para a ordenação dos documentos.

Exemplo de Query Reformulation

Query Reformulation (QR)



Vantagens / Desvantagens de QR

Vantagens:

- Está provado que QR melhora os resultados de sistemas de RI.
- Os melhores sistemas de RI do TREC e do CLEF usam módulos de QR.

Porque não se usa QR em motores de busca?

Vantagens / Desvantagens de QR

Vantagens:

- Está provado que QR melhora os resultados de sistemas de RI.
- Os melhores sistemas de RI do TREC e do CLEF usam módulos de QR.

Porque não se usa QR em motores de busca?

Desvantagens:

- Aumenta os tempos de resposta.
- "*Query Drift*"
- Prejudica pesquisas de páginas nomeadas.

Estratégias para QR

1) Recursos para selecção de termos:

- **Local** – documents recuperados sobre a pesquisa inicial.
- **Global** – O corpus completo, outros recursos como redes semânticas, tesauros, ontologias, etc.

Xu & Croft (1996) referem que a combinação das duas estratégias produz os melhores resultados.

Estratégias para QR

2) Aproximações para selecção de termos:

- ***Manual relevance feedback***: O utilizador selecciona um grupo de resultados relevantes. QR usa essa informação para reformular a pesquisa.
- ***Automatic relevance feedback***: os resultados relevantes são escolhidos automaticamente.
  - *Blind-relevance feedback*: usa um **nível de corte** para separar os documentos relevantes dos documentos irrelevantes.

Blind Relevance Feedback de Rocchio

Rocchio propôs em 1971:

$$Q_{i+1} = \frac{Q_i}{|Q_i|} + \frac{1}{|R|}\sum_{j=1}^{|R|}\frac{R_j}{|R_j|} - \frac{1}{|S|}\sum_{j=1}^{|S|}\frac{S_j}{|S_j|}$$

R: Documentos relevantes (*top-k docs*)
S: Documentos irrelevantes (*bottom-k docs*)

- Os termos são mais pesados se aparecerem em documentos relevantes, e são penalizados se aparecerem em documentos irrelevantes.

QR no contexto da tese

IR pode beneficiar com novas técnicas e recursos adicionais:

- Ideias a partir de sistemas de recomenda-ção ou da Web2.0.
- Prospecção dos *logs* do tumba! para encontrar *feedback* manual e comportamentos de utilizadores passados.
- Uso de ontologias de domínios específicos.
- Uso do conhecimento da língua portuguesa e/ou cultura.
- Aplicar técnicas de NLP.
- ...REM? QA?

Exemplo: *logs* do tumba!

- Os logs do tumba! são um recurso precioso:
  - Fonte de juízos de relevância de documentos em relação às pesquisas.
  - Contém os temas favoritos dos portugueses, e as estratégias de pesquisa usadas.
  - Uma grande variedade de termos de pesquisa, pesquisas reformuladas, pesquisas refinadas, etc para a mesma NI.
  - Fonte de sinónimos, relações semânticas entre termos, expressões multi-palavra, etc.

Exemplo: *logs* do tumba!

pesquisa?pag=**http://chinchilas.planetaclix.pt**&query_id=1064998254840&pos=2&terms=**chinchilas**&index=pt&lang=pt

- Pesquisa por: *chinchilas*.
- O utilizador escolheu o 3º resultado a partir da página de resultados (pos=2): http://chinchilas.planetaclix.pt

⇒ O utilizador julgou que o 3º resultado era mais relevante do que o 1º ou o 2º resultado, para a sua NI.

Exemplo: *logs* do tumba!

pesquisa?pag=**http://chinchilas.planetaclix.pt**&query_id=1064998254840&pos=2&terms=**chinchilas**&index=pt&lang=pt

Resultados podiam ser melhor ordenados.

- Porquê o 3º resultado?

  - URL / título / termos no snippet mais sugestivos? Quais?
  - Se sim, podem conter termos relacionados com a pesquisa inicial, e podem ser usados para QR.

Exemplo: *logs* do tumba!

- Outro utilizador pesquisou por “roedores” e escolheu, da lista de resultados, a página http://chinchilas.planetaclix.pt.

  • “chinchilas” e “roedores” são termos relacionados ao mesmo conceito, pois ambos pertencem a duas pesquisas que satisfazem a mesma NI.

  • Wen et al (2002) exploram a *query similarity* e o q*uery clustering* nos logs do Encarta, obtendo resultados promissores.

Exemplo: *logs* do tumba!

```
pesquisa?pag=http://www.terravista.pt/portosanto/3433/&query_id=1065115393712&pos=6&terms=diocese+viseu&index=sidra&lang=pt
```

- Pesquisa: “diocese viseu”
- A pesquisa tem âmbito geográfico, e deve ser interpretada como “**diocese@viseu**”, e não “**diocese+viseu**”.
- Uso de ontologias geográficas, *ranking* segundo critérios geográficos, etc.

Plano da tese

- 1º ano:
  - Trabalho relacionado / Literatura
  - Desenvolvimentos iniciais (logs do tumba!, p.ex)
  - Proposta para um protótipo de QR
- 2º ano:
  - Desenvolvimento de métodos para o módulo
  - Melhoramento do módulo com novas técnicas e recursos.
- 3º ano:
  - Implementação de QR no tumba!

Durante os 3 anos: Avaliação dos resultados do módulo de QR no CLEF

Automatic Query Reformulation using Contextual Information

Por: Nuno Cardoso

Orientadores:
Diana Santos e Mário J. Silva

Seminário Doutoral da Linguateca
3 e 4 de Outubro de 2006

Faculdade de Ciências
Universidade de Lisboa