Novos rumos na reformulação de consultas para RIG

Por: Nuno Cardoso

Orientadores:
Diana Santos e Mário J. Silva

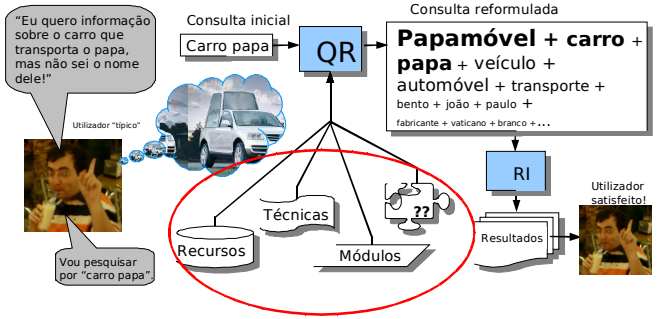
Simpósio Doutoral da Linguateca
4 de Outubro de 2007

Faculdade de Ciências
Universidade de Lisboa

Resumo da apresentação

- Há um ano atrás...
- mudanças na linha de investigação nos últimos 12 meses
- Participação no GeoCLEF 2007 e novas ideias que surgiram
- Primeiros esboços concretos de Motivação e Objectivos
- Plano de trabalhos até o SD de Abril de 2008

Há um ano atrás...



"(...) exploração de recursos alternativos, novas técnicas inspiradas em PLN, desenvolvimento de novos módulos para a reformulação automática de consultas"

Ideias em 2006

Foco na **reformulação automática de consultas** de todos os tipos.

- Uso de **recursos alternativos** e ainda pouco explorados
  - registos dos servidores *web*
  - ontologias específicas (ex. geográficas)
  - recursos Linguateca (BACO, RepEntiNo, etc)

Ideias em 2006 (cont.)

Foco na **reformulação automática de consultas** de todos os tipos.

- prospecção de texto, técnicas PLN
  - Detecção de termos compostos / EM
  - Correlações entre termos
  - Expansão por tesauros / sinónimos
  - extracção de lemas
- Aplicação de módulos no tumba!

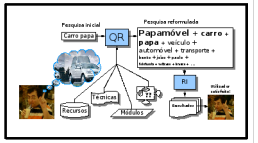
Nos últimos 12 meses...

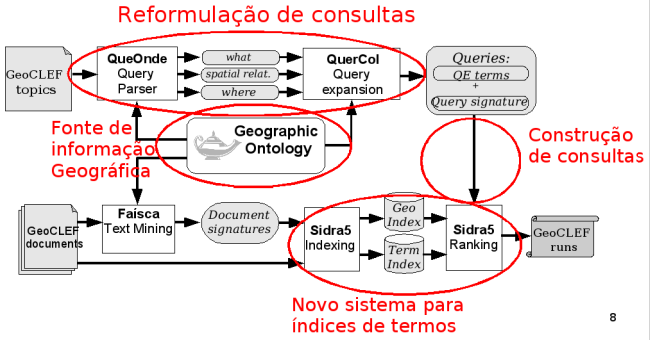
- Foco de investigação ainda em **reformulação de consultas**, mas num cenário mais particular: **consultas de cariz geográfico, num sistema de RI geográfico (RIG).**
- Melhor conhecimento do estado da arte no domínio de RIG e em reformulação de consultas.
- "Mãos na massa": GeoCLEF permitiu desenvolver um novo sistema RIG para testar novas técnicas e abordagens.

Nos últimos 12 meses (cont.)

- Novo sistema de RI baseado no MG4J (Boldi & Vigna, Universidade de Milão)
  - Permite construções de consultas mais elaboradas (AND, OR, distância de termos, *wildcards*)
  - Permite re-pesagem de termos de consulta
- Melhorias ao sistema de RIG (no XLDB, em colaboração com David Cruz)
  - Assinaturas geográficas
  - Novas **medidas de semelhança geográfica** (ainda por avaliar e melhorar)
  - **Novo motor de reformulação de consultas**

Sinopse do sistema RIG actual





QueOnde: Processamento de consultas

QueOnde = Query Engine for Ontology Defined Entities

- Lidar com as **diferentes formas de consultas** de cariz geográfico (nem tudo é *restaurantes em Lisboa*)
- Divisão em **<O quê, relação espacial, onde>**
- Reconhecimento de **"entidades geográficas" *(features)*** e de **"tipos de entidades geográficas" *(feature types)***

Exemplo: Tráfego marítimo nas ilhas portuguesas =

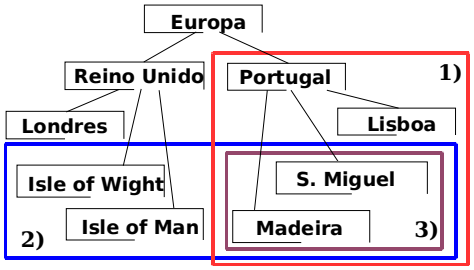
Reformulação de consultas

QuerCol: Query Collator, agora com dois eixos de expansão:

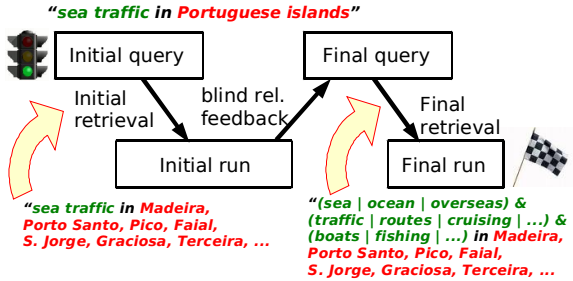
- Expansão geográfica
  - Orientada pelo tipo de consulta, e presença de relações espaciais, features e feature types
  - Baseada na informação da ontologia
- Expansão de termos
  - Até agora, só usei *blind relevance feedback* (retorno de relevância cego)

Expansão geográfica

1) *tráfego marítimo* **em Portugal**
2) *tráfego marítimo* **em ilhas**
3) *tráfego marítimo* **em ilhas portuguesas**



Expansão de termos: blind relevance feedback



- A esmagadora maioria dos grupos de investigação participantes no GeoCLEF (XLDB incluído) usam BRF. Não há nada mais "inteligente"?!

Utilidade da participação no GeoCLEF em 2007

- Primeiros testes com expansão geográfica de acordo com os critérios geográficos existentes
- Usar *feature types* de maneira mais inteligente (não só para desambiguar)
- Usar assinaturas geográficas nos documentos e nas consultas (ou seja, várias entidades geográficos como parte do âmbito)
- Ou seja, dar os primeiros passos em **reformulação de consultas mais inteligente**

Os resultados no GeoCLEF

- A maioria dos sistemas não consegue grandes ganhos de desempenho, ao adicionar processamento geográfico.
- Inclusivé, muitos sistemas RI puros obtêm resultados melhores do que RIGs!
- Porém, este ano:

			IR	GIR		IR/GIR
		GeoScore	Terms only	Geo. QE before RF	Geo. QE after RF	Terms/GIR
PT	Initial run		**0.210**	0.126	0.084	**0.210**
	Final Run	Maximum	**0,233**	0.125	0.104	0.205
		Mean		0.022	0.021	0.048
		Boolean		**0.135**	**0.125**	**0.268**
		Null		0.115	0.093	0.021
	a) Results for the Portuguese monolingual subtask.					
EN	Initial run		**0,175**	0.086	0.089	**0.175**
	Final Run	Maximum	**0.166**	0.093	0.104	**0.218**
		Mean		0.043	0.044	0.044
		Boolean		**0.131**	**0.135**	0.204
		Null		0.081	0.087	0.208
	b) Results for the English monolingual subtask.					

Esmiuçando os resultados

Parecia intuitivo que:

- separar a parte temática da parte geográfica ao início...
- ...expandir separadamente a parte temática com expansão de termos, e a parte geográfica com expansão guiada pela ontologia...
- ...usar dois índices separados (temático e geográfico)...

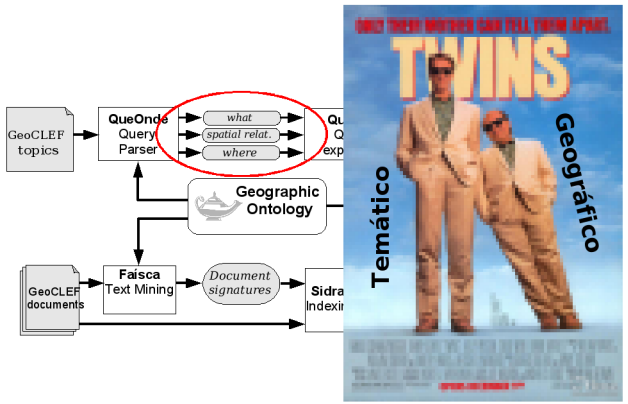
...produzisse melhores resultados, mas não.

Termo temático vs termo geográfico

Aviso: "consuma" os seguintes exemplos com moderação.

- "Actividades piscatórias em Portugal"
  - piscatória pode dar boas indicações de que os locais de interesse são zonas costeiras
- "Naufrágios de barcos portugueses"
  - Naufrágios sugere zonas aquáticas, perto de costas, ilhas, etc.
  - termos geográficos (*portugueses*) também podem ser bons termos de expansão (ex: nomes de ilhas)

Separar à nascença? Sim ou não?

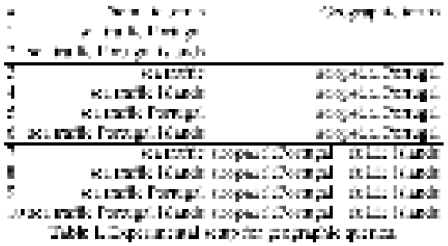


Rescaldo: novas experiências na forja

- *Rollback* completo nos sistemas GIR.
  - Analisar cada módulo ao detalhe
  - Avaliar, avaliar, avaliar (que é o grande problema da área de GIR, na minha opinião: se calhar, avançamos demasiado rápido)
  - Só depois de medir o módulo *n*, passar ao módulo *n+1*, e analisar a propagação dos erros
- Ou seja, de volta ao QueOnde...

Rescaldo: novas experiências na forja (cont.)

- Analisar os resultados com "misturas" entre parte geográfica e parte temática



Rescaldo: novas experiências na forja (cont.)

- QuerCol: mais experiências e avaliações com a reformulação de consultas
  - Adição de termos: deve haver bem melhor do que BRF
  - Re-pesagem de termos: ainda não usado
  - Termos compostos: usar índices próprios?
  - EM: útil para consultas tipo "James Bond"
  - Construção de consultas com os operadores disponíveis
  - Avaliar[avaliar]

Avaliação (em cima do joelho) do QueOnde

- "Corpus": 75 tópicos dos três anos de GeoCLEF (só títulos)
- Avaliação sistemática por módulo, ao longo do doutoramento.
- E não esquecer de fazer com consultas reais!

Ano dos tópicos do GeoCLEF	2005			2006			2007			todos os anos			
Tipo de consultas	total	😊	😞	total	😊	😞	total	😊	😞	total	😊	😞	
WHAT, REL, WHERE (**feat**)	20	18	2	19	14	5	13	13		52	45	7	87%
WHAT, REL, WHERE (**feat + feat. type**)	3	1	2	1		1	7	6	1	11	7	4	64%
WHAT, REL, WHERE (**feat. type**)							4	4		4	4	0	100%
REL, WHERE (**feat, feat. type**)	1	1		4	3	1	1	1		6	5	1	83%
REL, WHERE (**feat. type**)				1		1				1	0	1	0%

Avaliação (em cima do joelho) do QueOnde

Ano dos tópicos do GeoCLEF		2005			2006			2007			todos os anos			
Tipos de *features*		total	😊	😞	total	😊	😞	total	😊	😞	total	😊	😞	
Continente	(ex: Europa)	9	9	0	1	1		1,5	1,5		11,5	11,5	0	100%
Subconjunto	(Ex: Sudeste Asiático)				5	4	1	5	5		10	9	1	90%
País/es	(ex: Portugal)	9	7	2	2,5	2,5		2,5	2,5		14	12	2	86%
Região	(ex: Grande Lisboa)	2,5	0,5	2	8,5	5,5	3	5	5		16	11	5	69%
Cidade	(ex: Évora)				2	2	0	3	3		5	5	0	100%
Mares/Ocean	(ex: Mediterrâneo)	2,5	2,5	0	3	3		1	1		6,5	6,5	0	100%
Ilhas	(ex: ilhas escocesas)	1	1	0				2	2		3	3	0	100%
Rios/Lagos	(ex: Douro)				2	2	0				2	2	0	100%
montanhas	(ex: Himalaias)							1	1		1	1	0	100%
Sem *features*					1	0	1	4	4		5	4	1	80%

Ano dos tópicos do GeoCLEF	2005			2006			2007			todos os anos			
Tipos de *relações espaciais*	total	😊	😞	total	😊	😞	total	😊	😞	total	😊	😞	
Parte de (na, em, etc)	23	22	1	20	18	2	15,5	15,5		58,5	55,5	3	95%
Parte de + próximo	1	0	1	1	0	1				2	0	2	0%
Adjacente (beira, ao longo)				2	2	0				2	2	0	100%
Distância numerada (a X km de)				1	0	1				1	0	1	0%
Distância não especificada (perto)				1	0	1	2	2		3	2	1	67%
Entre							1	1		1	1	0	100%
Not (excluíndo)							0,5	0,5		0,5	0,5	0	100%
nas costas	1	1	0				1	1		2	2	0	100%
Sem relação (ao, de, com)							5	5		5	5	0	100%

Rescaldo: novas experiências na forja (cont.)

- Wikipedia: o recurso da moda :)
  - Rico (digo eu) em referências geográficas
  - Já há trabalhos (Overell et al.) de desambiguação e mapeamento de ref. geográficas na Wikipedia na ontologia TGN.
- Proposto para o GeoCLEF 2008, a ver se é usado.
  - Assuntos novos, permite consultas novas
  - bom recurso para complementar falhas de ontologia (plano B)

Porquê Wikipedia no GeoCLEF?

Dois tópicos concretos do GeoCLEF:

- "Regiões vinícolas à beira de rios na Europa"
- "Eleições livres em África"

- Intituitivamente, parece que a Wikipedia é melhor para o 1º tópico, e o CHAVE para o 2º tópico
- Por outro lado, a Wikipédia pode permitir tópicos geográficos com outra granularidade e morfologia, bem mais "desafiadores" para os sistemas RIG

Uma boa motivação

- Qual é realmente a melhor abordagem de tratamento dos termos das consultas num sistema RIG?
- Qual a melhor técnica de reformulação de consultas? Não há melhor que BRF? E a expansão geográfica já realizada carece de uma avaliação detalhada
- Que outros recursos valiosos se pode explorar, que melhorem significativamente os resultados?

Uma boa motivação (cont.)

- Agora que tenho um sistema RIG pronto para realizar experiências, posso analisar:
  - novas técnicas de reformulação temática (e comparar com *blind relevance feedback*)
  - aperfeiçoar a reformulação geográfica
  - talvez usar novos recursos (tesauros, dicionários de co-ocorrências, extracções referências Wikipedia) para as expansões
  - Verificar o peso de cada termo no processamento temático e no geográfico.

Espero, até Abril de 2008:

- Ter já uma ideia bem concreta de qual a melhor prática de utilização dos termos das consultas.
- Ter propostas concretas de novas técnicas de:
  - reformulação temática, com contribuição do raciocínio geográfico oriundo da ontologia e de outros recursos
  - reformulação geográfica, com a maturação de análise de tipos de restrições geográficas, o seu tratamento e o uso de índices temáticos

Espero, até Abril de 2008:

- Ter uma ideia mais concreta dos melhores recursos a “atacar”
  - avaliar a utilidade da Wikipedia, tesauros, diários de registos e outros recursos para o sistema RIG
- Prazo do GeoCLEF 2008: provavelmente, 30 de Maio. Ou seja, já ter um plano traçado para a participação, e já identificadas as experiências a avaliar.

Novas ideias para reformulação de consultas para RIG

Por: Nuno Cardoso

Orientadores:
Diana Santos e Mário J. Silva

Seminário Doutoral da Linguateca
4 de Outubro de 2007

Faculdade de Ciências
Universidade de Lisboa