Novos rumos para a recuperação de informação em português

Nuno Cardoso
ncardoso@xldb.di.fc.ul.pt

1 Introdução

A recuperação de informação (RI) tem sido uma área em franco crescimento nos últimos tempos, devido ao aumento exponencial de documentos e de serviços disponíveis através da Internet. As ferramentas de pesquisa de informação já fazem parte da nossa vida quotidiana, sendo usadas sobretudo para a procura de documentos concretos e de informação contida em documentos: desde motores de busca na Web, a pesquisa de correio electrónico ou as ferramentas de pesquisa de documentos no computador, todas estas aplicações têm como base os conceitos fundamentais de RI.

As ferramentas de RI baseiam-se na sua maioria em modelos estatísticos de termos, que estimam a relevância dos documentos para cada pesquisa de uma forma simples e funcional. Contudo, a incapacidade de interpretação das mensagens presentes nas linhas de consulta e nos documentos tem sido uma das principais limitações das ferramentas de RI, que encontram algumas dificuldades em encontrar documentos que satisfaçam necessidades de informação mais elaboradas. Allan et al [2] prevêem a exaustão dos actuais modelos de RI num futuro próximo, e referem que as novas tendências de RI passarão por uma contribuição decisiva de outras áreas de investigação mais afectas ao processamento de linguagem natural, como é o exemplo da extracção de informação, sumarização de textos ou a resposta automática de perguntas, com o intuito de compreender os tópicos subjacentes às consultas do utilizador, e utilizar esse conhecimento no processo de pesquisa.

Segundo Belkin [4], os novos desafios de RI passam por dar uma maior atenção às necessidades de cada utilizador, como são exemplos a personalização dos resultados de acordo com o perfil de pesquisa do utilizador e o contexto da sua pesquisa, a diversificação do tipo de documentos a recuperar e a apresentar ao utilizador (combinando documentos textuais com imagens, sons e vídeos), a pesquisa de informação em documentos escritos em várias línguas (RI multilingue) com possibilidades de usar tradução automática para que a língua não seja obstáculo para o acesso à informação desejada, a escolha do tipo de apresentação dos resultados da pesquisa (em forma de lista de documentos, resumos gerados automaticamente, ou somente a resposta pretendida), ou a ordenação dos resultados de acordo com uma determinada área geográfica de interesse (pesquisas com âmbito geográfico).

Singhal [24] resume esta nova fase de RI como uma mudança do ponto de vista do utilizador em relação à pesquisa de informação, onde este usa os sistemas de RI numa atitude de "Dá-me o que eu quero" em vez de "Dá-me o que eu disse". O futuro da investigação em RI passa inquestionavelmente pela compreensão das necessidades do utilizador e do contexto das suas pesquisas, e utilizando novas aproximações semânticas na recuperação de documentos de forma a fornecer resultados que se adequem às características da pesquisa de cada utilizador.

1.1 Sinopse

Este artigo descreve de forma sucinta o trabalho realizado até agora no âmbito do meu doutoramento, que está intimamente relacionado com os novos rumos de RI descritos na secção anterior. O trabalho de doutoramento foca a área de sistemas de recuperação de informação geográfica (RIG), nomeadamente os problemas da modelação do conhecimento geográfico, extracção e tratamento

automático de pistas geográficas no texto, e a correcta interpretação e reformulação das pesquisas dos utilizadores com restrições geográficas.

Neste artigo far-se-á uma descrição das directrizes que norteiam a minha investigação, seguido de uma apresentação detalhada do trabalho já realizado e dos módulos desenvolvidos no âmbito do doutoramento, e terminando com um resumo dos resultados obtidos em avaliações realizadas até ao momento. A secção 2 descreve a reformulação automática de consultas, e a sua aplicação para RIG, e a secção 3 caracteriza as fontes de informação que iremos explorar para criar uma rede de conhecimento que permita dotar os diversos módulos desenvolvidos da capacidade de raciocinar sobre o domínio geográfico. A secção 4 descreve o modelo RIG adoptado e detalha os respectivos módulos QuerCol, Rembrandt, MG4J e Renoir, e a secção 5 resume os resultados obtidos em avaliações conjuntas internacionais.

2 Compreendendo as consultas dos utilizadores

A interacção típica entre os utilizadores e as ferramentas de RI resume-se à criação de uma linha de consulta com termos chave que descreve a informação pretendida, e à consequente exibição de uma lista de documentos, ordenados de acordo com a sua pertinência em relação à informação pretendida.

Muitas vezes o utilizador não consegue descrever convenientemente a sua necessidade de informação numa lista de termos. Nestes casos, o utilizador opta por redigir linhas de consultas pequenas, cujos termos são vagas e/ou ambíguas, o que dificultará a tarefa do sistema de RI. Adicionalmente, o vocabulário usado pelo utilizador e pelos autores dos documentos para descrever os diversos assuntos pode ser diferente, existindo então uma barreira terminológica nas pesquisas que evita que certos documentos relevantes sejam recuperados, só porque certos conceitos são descritos através de termos diferentes.

2.1 Reformulação automática de consultas

A reformulação automática de consultas (RAC) é uma técnica frequentemente usada para lidar com certas limitações dos modelos tradicionais de RI, nomeadamente a barreira terminológica descrita na secção anterior. A RAC procura reformular a consulta inicial do utilizador de forma automática, adicionando termos fortemente relacionados com a pesquisa, removendo termos irrelevantes ou geradores de ruído, e atribuindo pesos de importância a cada termo [11]. No final, a linha de consulta reformulada será mais precisa e fiel à necessidade de informação real do utilizador, e mais robusta em relação às diferenças de vocabulário patente entre documentos e consultas. A actuação da RAC está esquematizada na Figura 1.

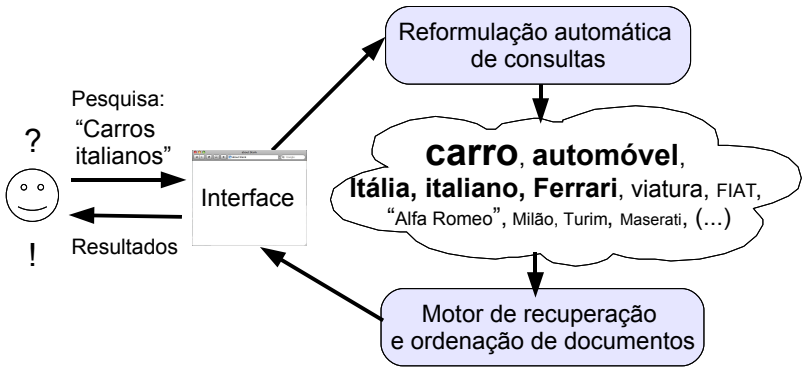

Figura 1: Esquema de funcionamento da reformulação automática de consultas (RAC).

A aplicação de RAC nas pesquisas tem como objectivo representar melhor os conceitos chave através das suas várias formas textuais, algo também subjacente à filosofia das "folksonomias" [17, 16], onde é normal associar uma nuvem de termos para catalogar um determinado documento, imagem ou vídeo, podendo essa nuvem de termos ser criada por diversos utilizadores que possuem diferentes perspectivas do documento em questão, e como tal, etiquetam-no com diferentes termos.

2.2 Pesquisas de âmbito geográfico

Existe uma percentagem considerável de pesquisas realizadas a motores de busca que dizem respeito a determinados tópicos de interesse confinados a uma área geográfica específica [13]. As dificuldades nas pesquisas com âmbitos geográficos estão muitas vezes relacionadas com o facto de os nomes de locais serem ambíguos, e poderem designar várias entidades distintas, como é o exemplo de nomes de pessoas ("Camilo Castelo Branco") ou de nomes de empresas ("France Press"). Mesmo quando os nomes geográficos se referem a locais, podemos encontrar vários locais com o mesmo nome (por exemplo, "Cuba" refere-se a um país e a uma cidade de Portugal), ou até ser um nome usado de forma metonímica (por exemplo, usando "Bruxelas" para mencionar as instituições da União Europeia).

O objectivo da minha tese de doutoramento é a investigação de novos métodos de RAC aplicados à recuperação de informação com âmbito geográfico, de forma a desambiguar o significado real dos nomes geográficos nas consultas e realizar a reformulação de acordo com a verdadeira intenção do utilizador, retornando resultados de acordo com a sua área geográfica de interesse. Um exemplo prático da aplicação do trabalho da minha tese está ilustrado na figura 2, onde podemos observar dois utilizadores com necessidades de informação diferentes, e que formulam as consultas "Obras de Castelo Branco" e "Restaurantes em Castelo Branco". Assumindo que o primeiro utilizador está interessado nas obras do romancista português, e o segundo em restaurantes na cidade portuguesa[1], cabe ao sistema RIG interpretar correctamente o significado de "Castelo Branco" destas duas consultas, tendo o módulo de RAC a responsabilidade de reajustar o seu mecanismo de reformulação para gerar linhas de consulta mais fiéis sobre a verdadeira semântica da consulta, em especial a consulta com âmbito na cidade de Castelo Branco. Desta forma, a recuperação de documentos terá atenção às diferenças semânticas entre as duas pesquisas, fornecendo resultados mais relevantes a cada um dos utilizadores.

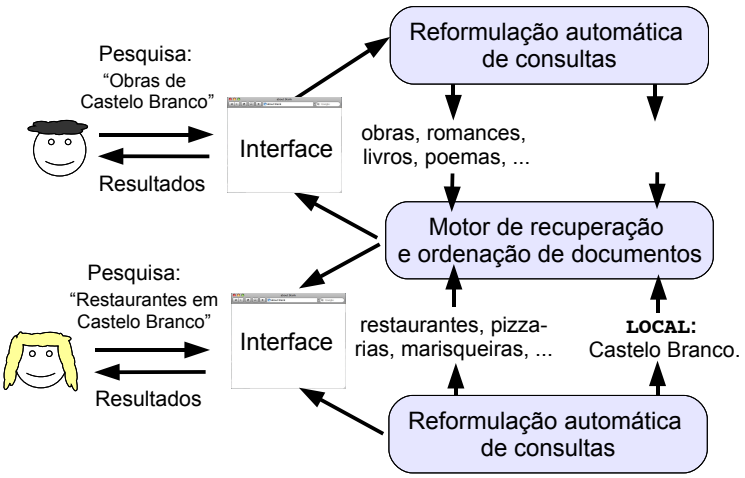

Figura 2: Reformulação automática de consultas com âmbitos geográficos.

[1]Para efeitos ilustrativos, vamos considerar que estas são as reais intenções dos utilizadores, e que não estão nem interessados em obras literárias sobre a cidade, nem sobre restaurantes onde o romancista terá alguma relação

3 Rede de conhecimento

Os métodos de RAC mais usados baseiam-se em algoritmos estatísticos, e usam a própria colecção de documentos como fonte de termos adicionais [26]. No contexto do nosso trabalho, procuramos novas formas de realizar a RAC que aproveitam a semântica dos termos para melhor entender as mensagens. Assim sendo, estou a construir uma *rede de conhecimento*, com o objectivo de fornecer a informação necessária para que o RAC interprete convenientemente os conceitos envolvidos na consulta, para que possa raciocinar sobre a melhor estratégia de reformulação a aplicar na consulta, e para que obtenha novos termos relevantes para serem usados na reformulação das consultas.

Defino a rede de conhecimento como sendo uma rede semântica composta por diversas fontes de informação, tais como enciclopédias electrónicas e ontologias, de onde é possível extrair conhecimento de uma forma objectiva e compatível para os sistemas inteligentes.

3.1 Fontes de informação

No âmbito do trabalho do doutoramento, estamos a explorar quatro fontes de informação particularmente relevantes para a extracção de conhecimento geográfico.

Ontologias geográficas

As ontologias geográficas representam o conhecimento humano sobre o domínio geográfico de uma forma hierárquica e inteligível para sistemas inteligentes. As ontologias permitem que os sistemas possam realizar operações de raciocínio geográfico básicas, tais como saber que cidades estão contidas numa região, ou quais os países atravessados por um determinado rio.

World-Wide Web

A partir de recolhas da Web é possível extrair informação sobre os sítios, os URL, títulos e resumos mais relevantes para cada pesquisa. Esta informação pode ser usada para caracterizar a consulta, como é o exemplo de detecção de consultas de cariz geográfica, se é uma consulta vaga ou precisa, ou se é uma consulta do tipo transaccional, navegacional ou informativo [5]. A caracterização das consultas é um passo importante para que seja possível ajustar a acção do módulo de RAC à pesquisa concreta, tal como evidencia Aires [1] no seu trabalho sobre a classificação dos resultados de busca na *web* portuguesa.

Wikipédia

A enciclopédia electrónica Wikipédia é uma referência incontornável na Internet. A Wikipédia reúne descrições detalhadas e bem documentadas sobre praticamente todos os tópicos, beneficiando das contribuições e validações de milhões de utilizadores de modo a garantir a fidelidade e a organização da informação a um nível sem precedentes. As páginas da Wikipédia referentes a locais (como por exemplo rios, países ou cidades), normalmente possuem informação adicional sobre as propridades do local numa *infobox*, como é exemplo as áreas, populações ou coordenadas desses locais. As propriedades desses locais podem ser aproveitadas para gerar conhecimento geográfico adicional para o módulo de RAC. A estrutura da Wikipédia, com as suas ligações, categorias e páginas de redireccionamento, tornam-na num recurso apetecível para áreas de investigação relacionadas com a extracção de informação e processamento de linguagem natural.

Diários dos servidores de motores de busca

Os diários dos servidores *web* registam as interacções entre os utilizadores e o motor de busca. Os diários permitem determinar as necessidades de informação mais típicas do utilizador, analisar o tipo de consultas formuladas ao motor de busca, estudar quais as páginas visitadas ao longo da pesquisa, e analisar as estratégias de reformulação manual das consultas, até o utilizador ficar satisfeito com a pesquisa, ou desistir sem conseguir obter a informação pretendida. Os

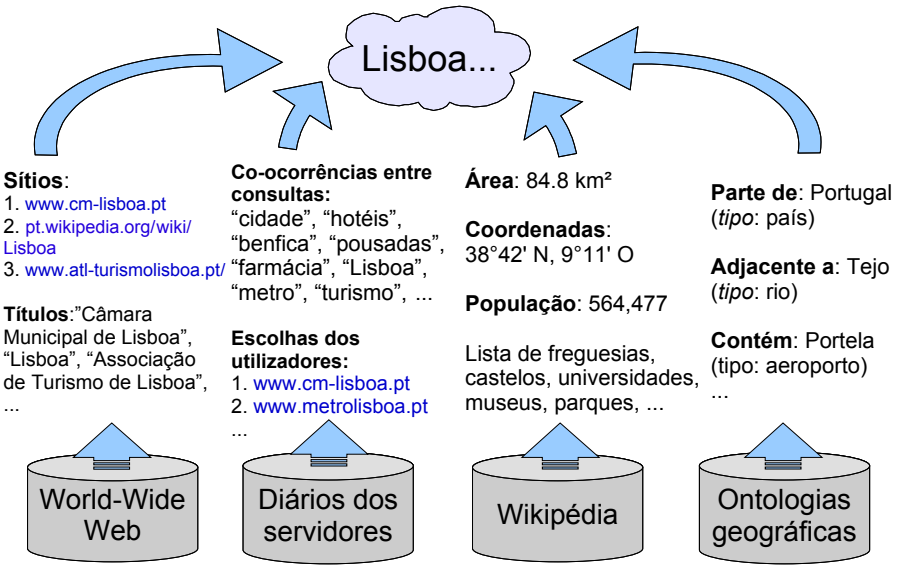

Figura 3: Uso da rede de conhecimento sobre o conceito "Lisboa".

	Ontologias geográficas	World-Wide Web	Wikipédia	Diários dos servidores
Acessibilidade	-	++	++	++
Credibilidade da informação	++	-	+	-
Diversidade de assuntos	-	++	++	+
Especificidade do domínio	++	-	+	- -
Inteligibilidade do formato	++	-	+	-
Actualização da informação	-	+	++	-
Conteúdos de utilizadores	- -	-	- -	++

Tabela 1: Características das fontes de informação.

diários dos servidores podem ser explorados de maneira a encontrar termos importantes a serem adicionados na RAC, ao identificar necessidades de informação semelhantes mas com linhas de consulta diferentes, ou até inferir certos focos de interesse sobre determinados tópicos a partir de determinados locais (por exemplo, pesquisas sobre um determinado surto pode ser originada a partir de um determinado local), e estudar o padrão de visualização de documentos para analisar a importância desses documentos para a respectiva área geográfica dos utilizadores.

A Figura 3 ilustra uma forma de aplicar a rede de conhecimento formada com base nas fontes de informação descritas acima, para extrair mais conhecimento sobre o conceito "Lisboa". A recolha da web fornece uma lista de sítios mais relevantes sobre Lisboa, e em conjunto com os diários de registos, podem fornecer um conjunto de termos normalmente correlacionados com "Lisboa", de um ponto de vista dos utilizadores da web. A Wikipédia pode fornecer informação importante sobre a cidade, e juntamente com a ontologia geográfica, é possível determinar a semelhança de Lisboa com outras entidades geográficas (tais como freguesias, monumentos ou aeroportos), e usar essa informação para o cálculo da relevância geográfica.

3.2 Características das fontes de informação

A tabela 1 resume as características de cada uma das fontes de informação, e refere as suas principais contribuições para a rede de conhecimento. O acesso aos conteúdos da Wikipédia em formato compactado é livre, enquanto que o acesso a recolhas da web é mais restritiva para

fins não-académicos. O público geral normalmente não tem acesso aos diários dos servidores, por causa dos problemas relacionados com a privacidade dos utilizadores do motor de busca. Contudo, para este trabalho de investigação, é possível usar os registos dos servidores do motor de busca tumba! [23].

As ontologias são cuidadosamente revistas e validadas, e consequentemente a sua informação possui altos níveis de credibilidade, seguindo-se a Wikipédia e a sua comunidade associada para actualizar e verificar os seus conteúdos. As ontologias são a escolha típica para a representação fidedigna de um determinado domínio, e como tal, estão confinadas ao domínio ao qual foram projectadas. A WWW e os diários dos servidores são o oposto, incluindo uma variedade vasta de assuntos. A Wikipédia representa um compromisso interessante, permitindo uma organização hierárquica dos assuntos através de um leque de categorias, sem restringir a diversidade de assuntos.

Em relação à inteligibilidade de formatos, as ontologias são o recurso mais fácil de ser usado pelos sistemas, que normalmente usam o formato OWL/RDF para a representação dos seus dados. A estrutura da Wikipédia também é bastante amigável para ser analisada automaticamente, enquanto que a WWW coloca bastantes desafios quanto à sua limpeza de dados. Os diários dos servidores não possuem uma estrutura pré-definida.

A Wikipédia gera periodicamente ficheiros compactados com o seu conteúdo, em formato XML ou em SQL, e como tal, a actualização da sua informação é elevada. Apesar de teoricamente a WWW estar sempre actualizada, é preciso despender algum tempo para realizar a recolha de documentos na web, pelo que poderá haver alguma desactualização, consoante o nível de actualização pretendido. Por outro lado, as ontologias são actualizadas com baixa frequência, uma vez que requerem a revisão e validação cuidadosa dos novos dados através de humanos peritos no domínio da ontologia. Finalmente, a característica mais atraente dos diários dos servidores é que possuem informação sobre os tópicos de interesse dos utilizadores, enquanto que os outros recursos não possuem dados sobre os utilizadores.

4 Trabalho desenvolvido até ao momento

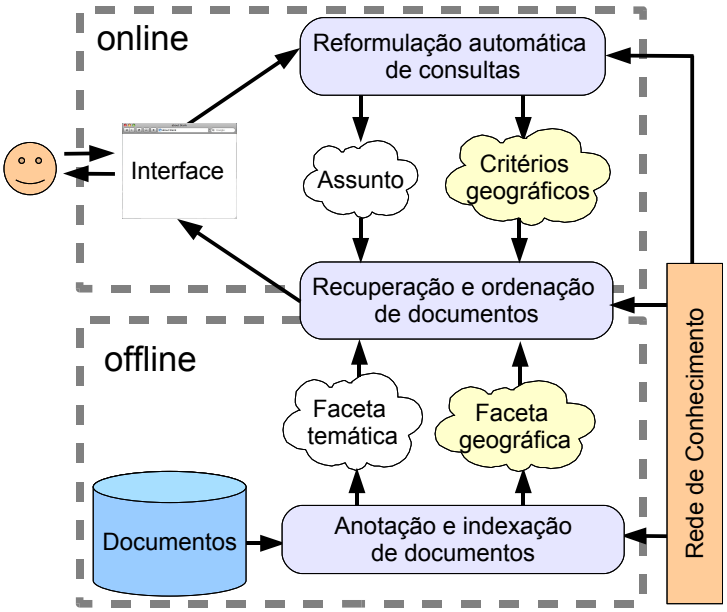

Figura 4: Arquitectura de um sistema RIG.

A Figura 4 esquematiza o modelo de RIG adoptado no meu trabalho. Podemos observar que a rede de conhecimento desempenha um papel crucial, assistindo os diversos módulos com informação geográfica essencial para o desempenho das suas tarefas. O trabalho realizado até agora tem focado os seguintes três pontos:

Reformulação automática de consultas, com particular ênfase na reformulação dos termos geográficos com a ajuda de ontologias geográficas. O QuerCol é um módulo desenvolvido com o propósito de investigar as melhores práticas para extrair a "geograficidade" das consultas, e de realizar a reformulação dos termos geográficos (expandindo "Ilhas Portuguesas" para os respectivos nomes, por exemplo), ou como lidar com relações espaciais nas consultas (por exemplo, "ao largo da costa portuguesa" torna locais como Peniche relevante, mas Évora não) [7].

Anotação dos documentos, onde se analisa automaticamente os documentos e procura-se extrair conteúdos de relevância geográfica, de maneira a encontrar pistas que possam indicar as áreas de interesse de cada documento. O trabalho desenvolvido neste ponto está patente no REMBRANDT, um sistema de reconhecimento de entidades mencionadas vocacionado para textos em português, e que utiliza principalmente a Wikipédia como fonte de informação para poder identificar e classificar as entidades mencionadas que estão presentes no texto [6].

Ordenação de documentos por critério geográfico, onde se procura conciliar os dois eixos de relevância (o assunto e a área geográfica de interesse) de forma a apresentar uma lista final de resultados com documentos relevantes e que correspondam às expectativas do utilizador. O trabalho realizado tem focado a adaptação do MG4J [25] ao nosso modelo de RIG.

4.1 QuerCol

O QuerCol é um módulo de RAC que possui duas formas de actuação: i) aplica uma técnica básica de expansão de termos intitulada de retorno de relevância cega (em inglês, *blind relevance feedback*, BRF) a todos os termos da consulta inicial [18], e ii) realiza uma expansão de termos geográficos ao associar os nomes geográficos na consulta às respectivas entidades geográficas na consulta, e explorando as suas relações ontológicas com outros locais para obter mais nomes geográficos

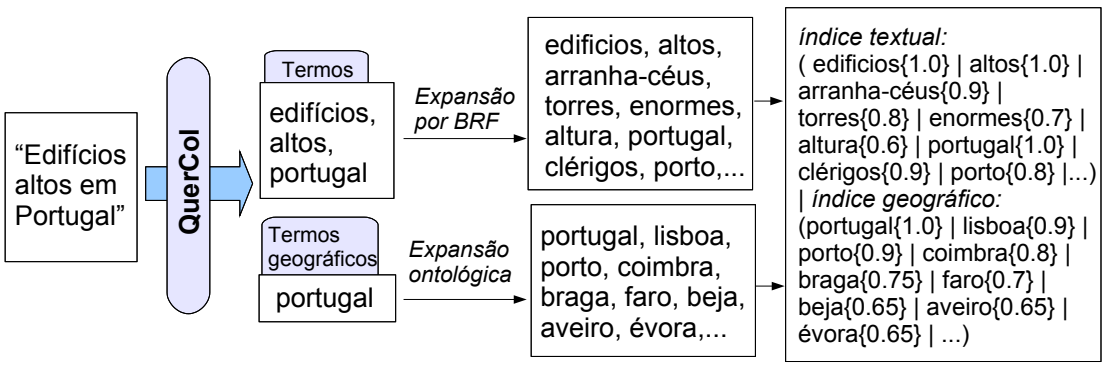

Figura 5: Esquema de funcionamento do módulo de RAC, QuerCol.

A figura 5 ilustra o procedimento usado pelo QuerCol para reformular a consulta "Edifícios altos em Portugal". Primeiro, o QuerCol remove palavras muito frequentes da consulta (como é o caso de "em"), e reconhece "Portugal" como sendo um termo potencialmente geográfico, com a ajuda do REMBRANDT. Os termos *edifícios*, *altos* e *portugal* são enviados ao processo de BRF, e utilizando o algoritmo de $w_t(p_t\text{-}q_t)$ para atribuir pesos numa escala normalizada de [0,1]. [12] Os termos expandidos, como é o exemplo de "arranha-céus", são concatenados à linha

inicial de consulta através de operadores lógicos OU (|), e etiquetados de forma a serem usados posteriormente num índice textual.

Por outro lado, o termo geográfico "Portugal" é emparelhado com o conceito geográfico de 'Portugal (país)'. A expansão ontológica procura outros conceitos georgáficos que estejam contidos dentro do território português, devido à relação espacial "em". As relações espaciais (por exemplo, "perto de" ou "nas costas de") e os tipos de entidades geográficas especificados (por exemplo, "praias", "montanhas" ou "universidades") são usadas para conduzir a procura por mais nomes geográficos relevantes [7]. Finalmente, são atribuídos pesos aos termos geográficos, e são etiquetados como sendo termos para serem usados num índice geográfico.

4.2 REMBRANDT

O REMBRANDT (**R**econhecimento de **E**ntidades **M**encionadas **B**aseado em **R**elações e **AN**álise **D**etalhada do **T**exto, `xldb.di.fc.ul.pt/Rembrandt`) é um sistema de reconhecimento de entidades mencionadas (REM) que utiliza a Wikipédia como fonte de informação, e que explora a sua estrutura rica em categorias, ligações e redirecionamentos para classificar todo o tipo de entidades presentes no texto. Desta forma, REMBRANDT tem acesso a conhecimento adicional sobre cada entidade mencionada (EM), o que se pode revelar útil para compreender o contexto da mensagem, detectar relações com outras EM, e usar essa informação para contextualizar e classificar EM vizinhas. Um exemplo pode ser o termo "Porto", que pode ser usado num contexto não-geográfico, como em "Porto de abrigo". Contudo, a presença da EM "Torre de Clérigos" na mesma frase pode reforçar a confiança em que "Porto" de facto seja uma EM relativa à cidade portuguesa, devido à sua ligação com a cidade que pode ser extraída a partir da informação na sua respectiva página da Wikipédia. A figura 6 exemplifica a actuação do REMBRANDT.

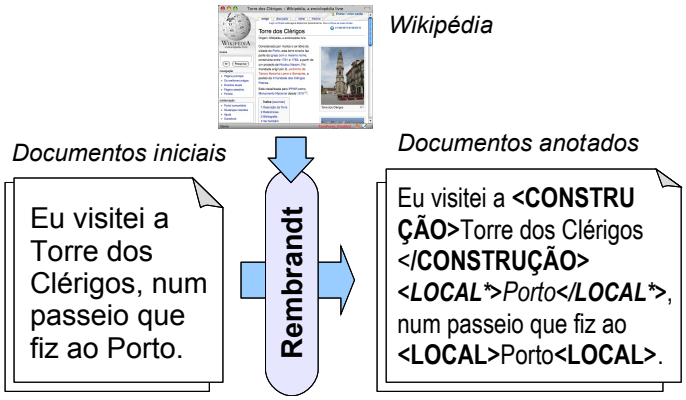

Figura 6: Acção do REMBRANDT na anotação de textos.

O REMBRANDT classifica as EM de acordo com as nove categorias e as 47 sub-categorias definidas pelo Segundo HAREM, uma avaliação conjunta para sistemas de REM para textos em português [20, 22]. As categorias principais são: `PESSOA`, `ORGANIZAÇÃO`, `LOCAL`, `TEMPO`, `VALOR`, `ABSTRACÇÃO`, `ACONTECIMENTO`, `COISA` e `OBRA`. O REMBRANDT lida perfeitamente com a vagueza intrínseca em algumas EM, ao classificá-las com mais de uma categoria ou sub-categoria. Por exemplo, a EM "Bombeiros Voluntários" podem ser considerados tanto uma organização ou um grupo de pessoas, consoante o contexto; se o contexto não permitir destrinçar o seu verdadeiro significado, o REMBRANDT atribui as duas classificações à EM.

A estratégia do REMBRANDT baseia-se no emparelhamento de cada EM à sua página respectiva na Wikipédia, e na análise da sua estrutura, ligações e categorias para obter mais conhecimento sobre a EM. REMBRANDT também depende de regras manuais para capturar pistas internas e externas para textos em português e inglês, tal como é descrito por McDonald [15]. As regras são usadas tanto para classificar EMs que não têm correspondência na Wikipédia ou correspondem

a páginas com informação insuficiente, como para corrigir o significado das EM de acordo com o contexto (por exemplo, "Rua de Portugal" designa uma rua, não um país). Adicionalmente, o REMBRANDT trata as categorias da Wikipédia como se fosse texto corrente, extraíndo assim os nomes geográficos das categorias e permitindo a extracção de informação geográfica *implícita* para cada EM [9].

4.3 MG4J

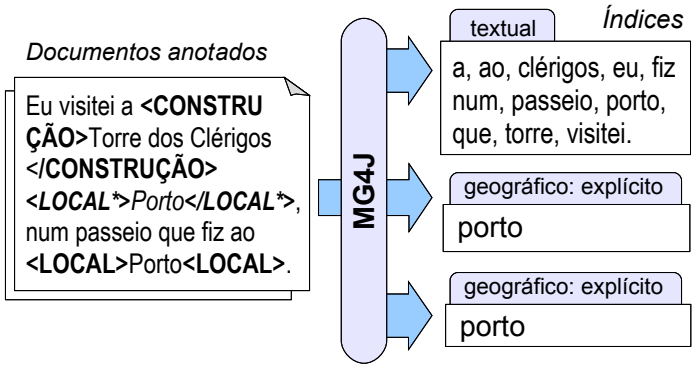

Figura 7: Indexação dos documentos anotados pelo MG4J.

O MG4J é o módulo responsável pela indexação e ordenação dos documentos. A figura 7 exemplifica a indexação selectiva que o MG4J faz aos textos anotados pelo REMBRANDT. Os termos não-geográficos são indexados num índice textual, enquanto que os termos geográficos são indexados em dois índices geográficos: um índice `geográfico explícito`, que inclui EM classificadas como sendo locais geográficos, e um índice `geográfico implícito`, para os locais associados a EM que não são explicitamente locais geográficos. No caso ilustrativo da figura 7, podemos observar que o termo "Porto" representa o local geográfico implícito da EM "Torre dos Clérigos", e como tal é indexado no índice destinado a termos geográficos implícitos.

4.4 RENOIR

Outro módulo que está a ser desenvolvido é o RENOIR (**R**EMBRANDT's **E**xtended **N**ER **O**n **I**nteractive **R**etrievals, `xldb.di.fc.ul.pt/Renoir`). O RENOIR pode ser visto como uma maneira de incorporar algumas técnicas interessantes aplicadas na área de resposta automática a perguntas (RAP), explorando não só a rede de conhecimento criada no âmbito do trabalho deste doutoramento, como também outras redes de conhecimento já extraídas e disponibilizadas, como é o caso da DBpedia [3], com o objectivo de adequar a pesquisa a um processo de interpretação das consultas e recuperando documentos com a informação pretendida.

Um exemplo que ilustra bem as motivações que norteiam o desenvolvimento do RENOIR é a realização de pesquisas com os termos "Castelo Branco.". Tal como foi referido anteriormente, uma pesquisa por "Obras de Castelo Branco" indicia que o utilizador está à procura de documentos sobre trabalhos do romancista português. Contudo, a consulta "Restaurantes de Castelo Branco" é mais direccionada para RIG, pois Castelo Branco refere-se à cidade portuguesa.

Com o RENOIR, procura-se investigar novas formas de enriquecer a sintaxe das consultas de forma a introduzir etiquetas semânticas de um modo manual, supervisionado ou automático. Nos exemplos anteriores, as linhas de consulta poderiam ser reformuladas para reflectir o contexto das pesquisas, como por exemplo, "Obras de `PESSOA`:{Castelo Branco}", e "Restaurantes `LOCAL`:{Castelo Branco}". Desta forma, o sistema RIG pode adaptar a sua actuação consoante a semântica da consulta, destrinçando os significados de "Castelo Branco" nos documentos (graças

às anotações do REMBRANDT) e retornando documentos de acordo com o contexto correcto de Castelo Branco.

5 Avaliação do desempenho dos sistemas

O trabalho desenvolvido no âmbito do doutoramento tem sido objecto de avaliação periódica, de maneira a aferir o desempenho dos protótipos e dos seus módulos constituintes na realização das tarefas a que se propõem. As avaliações constituem uma componente fundamental no processo de construção e validação dos módulos, uma vez que permitem analisar os pontos fortes e as fraquezas de cada componente, em ambientes de avaliação controlados que procuram recriar situações de pesquisas reais para as quais o sistema deverá estar devidamente preparado.

GeoCLEF

O GeoCLEF é uma pista de avaliação específica para sistemas de RIG [14]. No decurso do trabalho de investigação, a participação no GeoCLEF tem fornecido resultados bastante reveladores das potencialidades e das limitações das estratégias adoptadas para cada módulo [8]. O estado actual dos módulos e a linha de investigação agora seguida têm sido constantemente aperfeiçoados mediante uma análise detalhada dos resultados da avaliação, e que neste ano culminaram na participação na edição de 2008 do GeoCLEF, no qual se obteve resultados bastante encorajadores [10].

HAREM

O REMBRANDT participou no segundo HAREM, com o propósito de reconhecer todo o tipo de EM no texto. Também participou na sub-tarefa ReRelEM, para a detecção de relações entre EM. O REMBRANDT obteve um valor de medida F de 0.567 para a tarefa genérica de REM, cotando-se como o segundo melhor sistema num total de 10, e foi o primeiro sistema classificado para o cenário de EM da categoria `LOCAL`, com uma medida F de 0.625. Na tarefa de ReRelEM, o REMBRANDT também obteve o melhor resultado entre três sistemas, com uma medida F de 0.103.

GikiP

O GikiP é uma pista piloto promovida pela Linguateca sob a chancela da pista GeoCLEF, propondo aos sistemas participantes uma tarefa de procura de artigos/entradas da Wikipédia que satisfazem uma dada necessidade de informação que exija algum raciocínio geográfico [21, 19]. O RENOIR participou no GikiP ainda de uma forma supervisionada, utilizando a Wikipédia e o REMBRANDT como fonte de informação e de extracção de conhecimento para assistir a sua nova estratégia de formulação de consultas. Apesar de o RENOIR ainda estar nos seus primeiros passos, a participação no GikiP permitiu ter uma primeira experiência de como a sua filosofia orientada a consultas semânticas poderá permitir responder a necessidades de informação elaboradas, como são os casos dos tópicos "Indique membros do círculo de Viena que nasceram fora do império austro-húngaro ou da Alemanha", ou "Locais onde Goethe viveu".

Referências

[1] Rachel Aires. *Uso de marcadores estilísticos para a busca na Web em português.* Tese de doutoramento, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Agosto de 2005.

[2] James Allan, Jay Aslam, Nicholas Belkin, Chris Buckley, Jamie Callan, Bruce Croft, Sue Dumais, Norbert Fuhr, Donna Harman, David J. Harper, Djoerd Hiemstra, Thomas Hofmann, Eduard Hovy, John Lafferty Wessel Kraaij, Victor Lavrenko, David Lewis, Liz Liddy, R. Manmatha, Andrew McCallum, Jay Ponte, John Prager, Dragomir Radev, Philip Resnik, Stephen Robertson, Roni Rosenfeld, Salim Roukos, Mark Sanderson, Rich Schwartz, Amit Singhal,

Alan Smeaton, Howard Turtle, Ellen Voorhees, Ralph Weischedel, Jinxi Xu e ChengXiang Zhai. Challenges in Information Retrieval and Language Modeling: Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, September 2002. *SIGIR Forum*, p. 31–47, 2003.

[3] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak e Zachary Ives. DBpedia: A Nucleus for a Web of Open Data, Em Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber e Philippe Cudré-Mauroux, editores, *6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11–15, 2007, Proceedings*, número 4825 em LNCS, p. 722–735, Springer, 2007.

[4] Nicholas J. Belkin. Some(what) Grand Challenges for Information Retrieval, Em Craig MacDonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven e Ryen W. White, editores, *Advances in Information Retrieval, 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*, volume 4956 de *LNCS*, p. 1, Springer, 2008.

[5] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[6] Nuno Cardoso. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e Análise Detalhada do Texto. Em Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: Actas do Encontro do Segundo HAREM*, Aveiro, Portugal, 11 de Setembro de 2008.

[7] Nuno Cardoso e Mário J. Silva. Query Expansion through Geographical Feature Types. Em *Proceedings of the 4th Workshop on Geographic Information Retrieval, GIR'07 (CIKM'2007 Workshop)*, Lisboa, Portugal, 9 de Novembro de 2007.

[8] Nuno Cardoso, David Cruz, Marcirio Chaves e Mário J. Silva. Using Geographic Signatures as Query and Document Scopes in Geographic IR, Em *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 de *LNCS*, p. 802–810, Springer, 2008.

[9] Nuno Cardoso, Mário J. Silva e Diana Santos. Handling Implicit Geographic Evidence for Geographic IR. Em *Proceedings of the 17th Conference on Information and Knowledge Management, CIKM'2008*, Napa Valley, CA, EUA, 27–29 de Outubro de 2008.

[10] Nuno Cardoso, Patrícia Sousa e Mário J. Silva. The University of Lisbon at GeoCLEF 2008. Em Francesca Borri, Alessandro Nardi e Carol Peters, editores, *Working notes of the 9th Workshop of the Cross-Language Evaluation Forum, CLEF'2008*, Aarhus, Dinamarca, 17–19 de Setembro de 2008.

[11] Efthimis N. Efthimiadis. Query expansion. *Annual Review of Information Systems and Technology, ARIST*, 31:121–187, 1996.

[12] Efthimis N. Efthimiadis. A user-centered evaluation of ranking algorithms for interactive query expansion. Em Robert Korfhage, Edie M. Rasmussen e Peter Willett, editores, *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, SIGIR'93*, Pitsburgo, PA, EUA, 27 de Junho a 1 de Julho de 1993. p. 146–159.

[13] Janet Kohler. Analysing Search Engine Queries for the Use of Geographic Terms. Tese de mestrado, Universidade de Sheffield, 2003.

[14] Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker e Xing Xie. GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview, Em Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivian Petras

e Diana Santos, editores, *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, volume 5251 de *Lecture notes on Computer Science*, Springer, 2007.

[15] D. McDonald. Internal and external evidence in the identification and semantic categorization of proper names. Em I. Boguraev e J. Pustejovsky, editores, *Corpus processing for lexical acquisition*. MIT Press, Cambridge, MA, EUA, 1996, capítulo 2, p. 21–39.

[16] Peter Mika. Ontologies Are Us: A Unified Model of Social Networks and Semantics, Em Yolanda Gil, Enrico Motta, V. Richard Benjamins e Mark A. Musen, editores, *The Semantic Web – ISWC 2005, 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, November 6–10, 2005, Proceedings*, volume 3729 de *LNCS*, p. 522–536, Springer, 2005.

[17] Peter Mika. Social Networks and the Semantic Web. Em *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, WI'04*, Pequim, China, 20–24 de Setembro de 2004. p. 285–291.

[18] J. J. Rocchio Jr. Relevance Feedback in Information Retrieval. Em Gerard Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, 1971. p. 313–323.

[19] Diana Santos e Nuno Cardoso. GikiP: Evaluating geographical answers from Wikipedia. Em *5th Workshop on Geographic Information Retrieval, GIR'08*, Napa Valley, CA, EUA, 30 de Outubro de 2008.

[20] Diana Santos, Nuno Seco, Nuno Cardoso e Rui Vilela. HAREM: An Advanced NER Evaluation Contest for Portuguese. Em Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik e Daniel Tapias, editores, *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC'2006*, Génova, Itália, 22–28 de Maio de 2006. p. 1986–1991.

[21] Diana Santos, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling e Yvonne Skalban. Getting geographical answers from Wikipedia: the GikiP pilot at CLEF. Em Francesca Borri, Alessandro Nardi e Carol Peters, editores, *Working notes of the 9th Workshop of the Cross-Language Evaluation Forum, CLEF'2008*, Aarhus, Dinamarca, 17–19 de Setembro de 2008.

[22] Diana Santos, Paula Carvalho, Hugo Oliveira e Cláudia Freitas. Second HAREM: new challenges and old wisdom. Em *International Conference on Computational Processing of Portuguese Language, PROPOR'2008*, Aveiro, Portugal, 8-10 de Setembro de 2008.

[23] Mário J. Silva. The Case for a Portuguese Web Search Engine. Em *Proceedings of the 2003 IADIS International Confere nce on WWW Internet, ICWI-03*, Faro, Portugal, 2003. p. 411–418.

[24] Amit Singhal. Web Search: Challenges and Directions, Em Craig MacDonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven e Ryen W. White, editores, *Advances in Information Retrieval, 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*, volume 4956 de *LNCS*, Springer, 2008.

[25] Sebastiano Vigna e Paolo Boldi. MG4J: Managing Gigabytes for Java™. `http://mg4j.dsi.unimi.it/`. Dezembro de 2007.

[26] Jinxi Xu e Bruce Croft. Query Expansion Using Local and Global Document Analysis. Em Hans-Peter Frei, Donna Harman, Peter Schäuble e Ross Wilkinson, editores, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96*, Zurique, Suíça, 18-22 de Agosto de 1996. p. 4–11.