Capítulo 11

Novos rumos para a recuperação de informação geográfica em português

Nuno Cardoso

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 11, p. 71–85, 2008

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: motores de busca na rede, pesquisa de correio electrónico ou 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 consulta de uma forma simples e funcional. Contudo, a incapacidade de interpretação do significado dos textos das consultas e dos documentos tem sido uma das principais limitações das ferramentas de RI, que encontram assim algumas dificuldades em encontrar documentos que satisfaçam algumas necessidades de informação mais elaboradas. Allan et al. (2003) 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 a perguntas, com o intuito de compreender os tópicos subjacentes às consultas do utilizador, e utilizar esse conhecimento no processo de recuperação de documentos.

Segundo Belkin (2008), os novos desafios em RI passam por dar uma maior atenção às necessidades de cada utilizador, personalizando os resultados de acordo com o seu perfil e o contexto da sua pesquisa. A pesquisa de informação deverá aplicar técnicas de tradução automática, de forma a incluir documentos escritos em várias línguas (RI multilingue) e fazer com que a língua não seja obstáculo para o acesso à informação desejada. O utilizador terá controle sobre o método de pesquisa, como por exemplo a ordenação dos resultados de acordo com uma determinada área geográfica de interesse (pesquisas com âmbito geográfico), ou a escolha do tipo de resposta pretendido (em forma de lista de documentos, resumos gerados automaticamente, ou somente a resposta exacta). Finalmente, os resultados deverão ser apresentados de acordo com o contexto da pesquisa, combinando documentos textuais, imagens, sons, vídeos ou mapas sempre que forem relevantes para ilustrar a informação pretendida.

Singhal (2008) resume esta nova fase da 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, na compreensão dos tópicos abordados nas suas línguas específicas, e no uso de novas aproximações semânticas na recuperação de documentos de forma a fornecer resultados que se adequem às características de cada pesquisa.

Neste artigo apresento a minha perspectiva sobre os novos rumos de recuperação de in-

formação, com base na investigação realizada até agora no âmbito do meu doutoramento. O meu trabalho foca a área de sistemas de recuperação de informação geográfica (RIG) para o português, nomeadamente os problemas da modelação do conhecimento geográfico, o tratamento dos textos em português para a extracção automática de pistas geográficas no texto, e a correcta interpretação e reformulação das consultas dos utilizadores com restrições geográficas. A secção 11.1 descreve a técnica de reformulação automática de consultas e a sua aplicação em RIG. A secção 11.2 caracteriza as fontes de informação que irei explorar para criar uma rede de conhecimento que permite dotar os diversos módulos desenvolvidos da informação necessária para raciocinar sobre o domínio geográfico. A secção 11.3 descreve o modelo RIG adoptado e detalha os respectivos módulos QuerCol, REMBRANDT, MG4J e RENOIR, e a secção 11.4 refere as participações em avaliações conjuntas internacionais realizadas até agora.

11.1 Compreendendo as consultas dos utilizadores

Os utilizadores interagem tipicamente com as ferramentas de RI com o intuito de realizar *pesquisas* e satisfazer uma determinada necessidade de informação. As pesquisas são compostas por uma ou mais *consultas*, ou seja, linhas de texto contendo normalmente termos-chave que procuram descrever a informação pretendida. Para cada consulta enviada, a ferramenta RI devolve uma lista de documentos ordenados de acordo com a sua pertinência em relação à consulta.

Muitas vezes o utilizador não consegue descrever convenientemente a sua necessidade de informação numa consulta. Nestes casos, ele opta por realizar consultas pequenas, cujos termos são vagos e/ou ambíguos, 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 que evita que certos documentos relevantes sejam recuperados, só porque certos conceitos são descritos através de termos diferentes.

11.1.1 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 referida anteriormente. A RAC procura reformular a consulta inicial 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 (Efthimiadis, 1996). No final, a consulta reformulada deverá ser mais precisa e fiel à necessidade de informação real do utilizador, e mais robusta em relação às diferenças

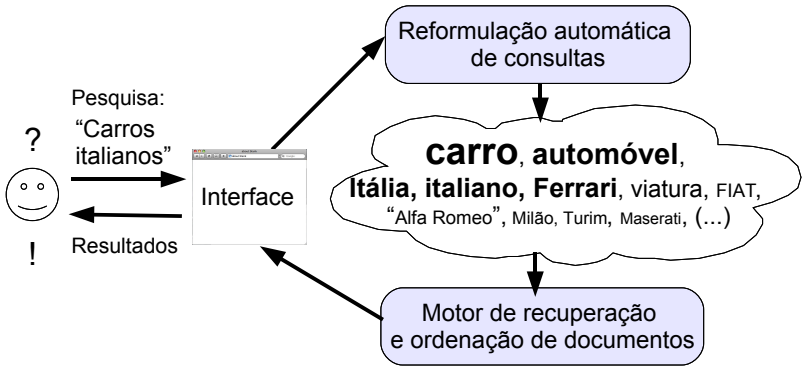

Figura 11.1: Esquema de funcionamento da reformulação automática de consultas (RAC).

de vocabulário patente entre documentos e consultas. A actuação da RAC está esquematizada na figura 11.1.

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" (Mika, 2006, 2004), onde é normal associar uma nuvem de termos para catalogar um determinado documento, imagem ou vídeo. A nuvem de termos pode ser criada por diversos utilizadores que possuem diferentes perspectivas do documento em questão, e como tal, é frequente que as nuvens tenham bastantes termos, e inclusivé oriundos de diversas línguas.

11.1.2 Consultas de âmbito geográfico

Existe uma percentagem considerável de consultas realizadas a motores de busca que dizem respeito a determinados tópicos de interesse confinados a uma área geográfica específica (Kohler, 2003). As dificuldades nas pesquisas com âmbitos geográficos estão muitas vezes relacionadas com o facto de os nomes de locais usados 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 em português com âmbito geográfico, de forma a desambiguar o significado real dos nomes geográficos nas consultas e realizar a

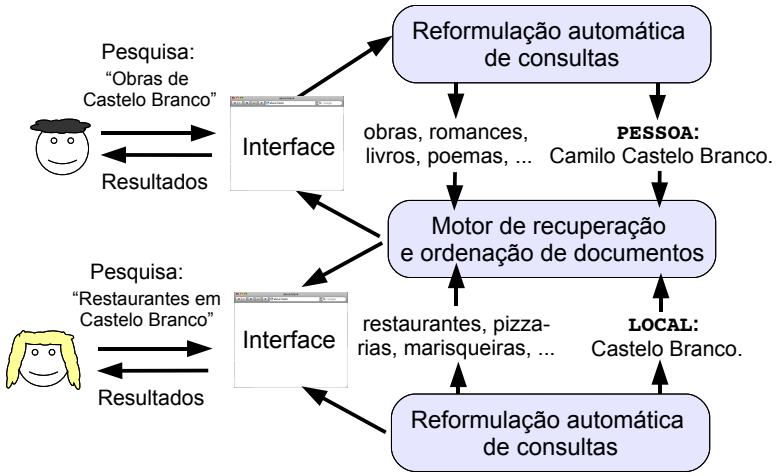

Figura 11.2: Reformulação automática de consultas para pesquisas diferentes.

reformulação de acordo com a verdadeira intenção do utilizador, fornecendo resultados de acordo com a sua área geográfica de interesse.

Um exemplo prático da aplicação do trabalho da minha tese está ilustrado na figura 11.2, onde podemos observar dois utilizadores com necessidades de informação diferentes, que formularam duas consultas diferentes nas suas pesquisas, "Obras de Castelo Branco" e "Restaurantes em Castelo Branco". Assumindo que o primeiro utilizador está interessado nas obras literárias do romancista português, e o segundo em restaurantes na cidade portuguesa[1], cabe ao sistema RIG interpretar correctamente a intenção subjacente nas duas pesquisas, e interpretar correctamente o significado de "Castelo Branco" em cada uma das consultas. O módulo de RAC deverá reajustar o seu mecanismo de reformulação de maneira a gerar consultas mais fiéis sobre a verdadeira semântica da pesquisa, em especial a consulta com âmbito geográfico na cidade de Castelo Branco. Desta forma, a recuperação de documentos terá atenção às diferenças semânticas entre as duas pesquisas, fornecendo os resultados mais relevantes para cada um dos utilizadores.

11.2 Rede de conhecimento

No contexto do meu trabalho, estou a investigar novas formas de realizar a RAC em português, aproveitando o conhecimento da língua e do significado dos termos para melhor entender as consultas. Para tal, estou a construir uma *rede de conhecimento* em português, com o objectivo de fornecer a informação necessária para que a RAC interprete convenien-

[1] Para efeitos deste exemplo, vamos considerar que estas são as reais intenções dos utilizadores, e que não estão nem interessados em obras artísticas sobre a cidade, nem sobre restaurantes relacionados de alguma forma com o romancista.

temente os conceitos envolvidos na consulta, raciocine sobre a melhor estratégia a aplicar na consulta, e obtenha consequentemente novos termos relevantes. Defino a rede de conhecimento como sendo uma rede semântica composta por diversas fontes de informação de onde é possível extrair conhecimento de uma forma objectiva e automática.

11.2.1 Fontes de informação

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

i. Ontologias geográficas

A Geo-Net-PT01 é uma ontologia geográfica detalhada sobre o território português, e é usada como fonte de informação primordial para operações básicas de raciocínio geográfico (Chaves et al., 2005b). As ontologias geográficas representam o conhecimento humano sobre o domínio geográfico de uma forma hierárquica e inteligível, permitindo o acesso a conhecimento geográfico complexo, como por exemplo saber que cidades estão contidas numa região, ou quais os países atravessados por um determinado rio.

ii.Recolhas da rede

A WPT 03 é uma recolha da rede portuguesa realizada em 2003, e permite extrair informação sobre os sítios, os URL, os títulos e os resumos mais relevantes para as pesquisas realizadas pela comunidade portuguesa (Cardoso et al., 2007). Esta informação pode ser usada, por exemplo, para gerar um grafo da rede Arasu et al. (2001) e estimar a importância de cada sítio na rede, de forma a determinar se a consulta é do tipo transaccional, navegacional ou informativo (Broder, 2002), para auxiliar na detecção de consultas de cariz geográfica, ou para determinar se a consulta é vaga ou precisa,

. 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 no seu trabalho sobre a classificação dos resultados de busca na rede portuguesa (Aires, 2005).

iii. Wikipédia

A porção portuguesa da Wikipédia, que conta em 2008 com mais de 400.000 artigos, é usada como fonte de conhecimento sobre diversos tópicos de interesse, auxiliando a interpretação das consultas dos utilizadores portugueses. Esta enciclopédia electrónica é uma referência incontornável na Internet, reunindo descrições detalhadas e bem documentadas sobre um grande número de tópicos, beneficiando das contribuições e validações de muitos 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 propriedades do local numa caixa de informação (*infobox*), como por exemplo as áreas, populações ou coordenadas respectivas, podendo ser aproveitadas para extrair conhecimento geográfico adicional para o módulo de RAC.

iv. Diários dos servidores de motores de busca

Os diários dos servidores do motor de busca *tumba!* registam as interacções entre os utilizadores e o *tumba!* (Silva, 2003). Estes diários permitem determinar as necessidades de informação mais típicas do utilizador, analisar o tipo de consultas formuladas, 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 diários podem ser explorados de maneira a encontrar termos importantes a serem adicionados na RAC, ao identificar necessidades de informação semelhantes mas com consultas diferentes, ou até inferir certos focos de interesse sobre determinados tópicos a partir de determinados locais (por exemplo, pesquisas sobre o surto de determinada doença podem ser originadas 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 11.3 ilustra uma forma de aplicar a rede de conhecimento formada com base nas fontes de informação apresentadas, para extrair mais conhecimento sobre o conceito "Lisboa". Um grafo da WPT 03 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", do ponto de vista dos utilizadores do *tumba!*. 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.

11.2.2 Características das fontes de informação

A tabela 11.1 resume as características de cada uma das fontes de informação mencionadas, 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 rede é mais restritivo 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 diários dos servidores do motor de busca *tumba!*.

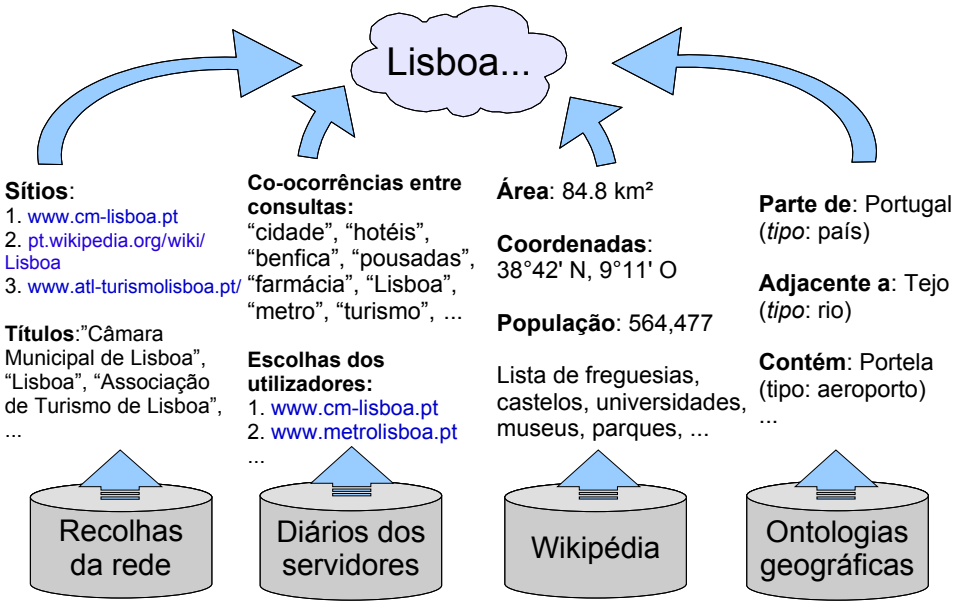

Figura 11.3: Uso da rede de conhecimento sobre o conceito "Lisboa".

No caso das ontologias geográficas, a Geo-Net-PT01 está disponível gratuitamente em `xldb.di.fc.ul.pt/geonetpt`.

A informação contida nas ontologias possui altos níveis de credibilidade, uma vez que estas são cuidadosamente revistas e validadas. A Wikipédia e a sua vasta comunidade que actualiza e verifica os seus conteúdos fazem com que seja um recurso com elevada credibilidade na sua informação. A rede, por sua vez, não possui restrições quanto à informação publicada, pelo que a sua credibilidade normalmente é estimada indirectamente através do sítio onde foi publicado, por exemplo.

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 rede e os diários dos servidores são o oposto, incluindo uma grande variedade de assuntos. A Wikipédia representa um meio termo interessante, permitindo uma organização hierárquica dos assuntos através de um leque de categorias, restringindo apenas a diversidade de assuntos com base numa política de relevância para os propósitos de uma enciclopédia da rede (ver `en.wikipedia.org/wiki/Wikipedia:List_of_policies`).

Em relação à inteligibilidade de formatos, as ontologias são o recurso mais fácil de ser usado pelos sistemas, uma vez que já vêem num formato estruturado, próprio para processamento computacional (normalmente o formato OWL/RDF). A estrutura da Wikipédia também é bastante amigável para ser analisada automaticamente, enquanto que a rede coloca bastantes desafios quanto à sua limpeza de dados. Os diários dos servidores, apesar

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

Tabela 11.1: Características das fontes de informação.

de terem uma formatação típica com campos separados por tabulações, não possuem uma formatação padrão no que diz respeito à representação da informação sobre as interacções dos utilizadores. Os diários do *tumba!* incluem bastante informação adicional a esse nível, permitindo extrair informação sobre os hábitos de pesquisa dos utilizadores, como por exemplo estimar o tempo médio que os utilizadores dispendem nas suas pesquisas, ou agregar as várias consultas usadas para cada pesquisa (Seco e Cardoso, 2006).

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 rede estar sempre actualizada, é preciso dispender algum tempo para realizar a recolha de documentos na rede, pelo que poderá haver alguma desactualização dos conteúdos. 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.

11.3 Trabalho desenvolvido até ao momento

A figura 11.4 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 necessária para o desempenho das suas tarefas. O trabalho realizado até agora tem focado os seguintes três pontos:

i. Reformulação automática de consultas

A abordagem de RAC adoptada possui uma atenção especial na reformulação dos termos geográficos com a ajuda da ontologia geográfica Geo-Net-PT01. O QuerCol é um módulo desenvolvido com o propósito de investigar as melhores práticas para extrair a "geogra-

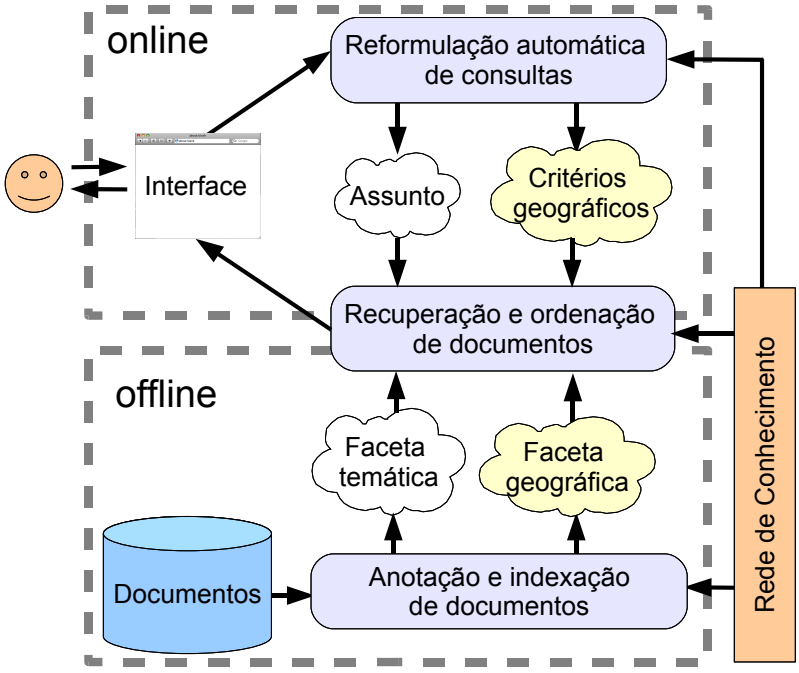

Figura 11.4: Arquitectura do sistema de RIG.

ficidade" 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) (Cardoso e Silva, 2007).

ii. Anotação dos documentos

Os documentos em português são analisados automaticamente, com o intuito de extrair conteúdos de relevância geográfica e 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 porção portuguesa da Wikipédia como fonte de informação para poder identificar e classificar as entidades mencionadas que estão presentes no texto em português (Cardoso, 2008).

iii. Ordenação de documentos por critério geográfico

Na fase de recuperação e ordenação de documentos, procura-se 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 (Boldi e Vigna, 2005) ao modelo de RIG.

11.3.1 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 cego (em inglês, *blind relevance feedback*, BRF) a todos os termos da consulta inicial (Rocchio Jr, 1971), e ii) realiza uma expansão de termos geográficos ao associar os nomes geográficos na consulta às respectivas entidades geográficas, e explorando as suas relações ontológicas com outros locais para obter mais nomes geográficos

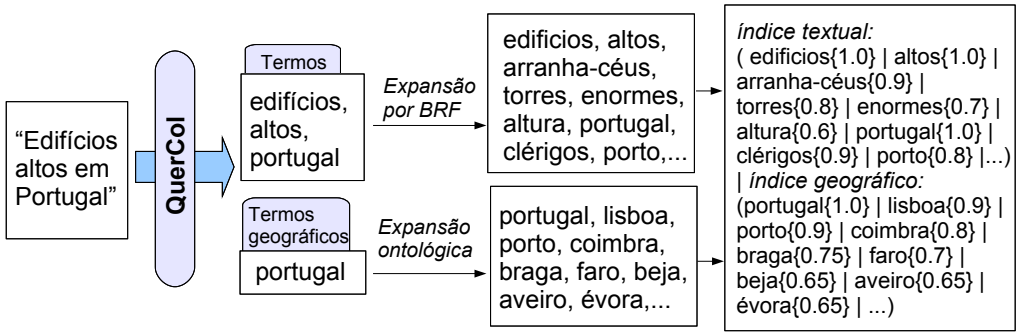

Figura 11.5: Funcionamento do QuerCol, um módulo de RAC.

A figura 11.5 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, utilizando o algoritmo $w_t(p_t\text{-}q_t)$ para atribuir pesos numa escala normalizada de [0,1]. (Efthimiadis, 1993) Os termos expandidos, como é exemplo "arranha-céus", são concatenados à consulta inicial 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 geográficos 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 usados para conduzir a procura por mais nomes geográficos relevantes (Cardoso e Silva, 2007). Final-

mente, são atribuídos pesos aos termos geográficos, e são etiquetados para serem usados num índice geográfico.

11.3.2 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, o 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. Usemos como exemplo o termo "Porto", que pode ser utilizado num contexto não-geográfico, como em "António da Silva Porto". 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, como é ilustrado na figura 11.6.

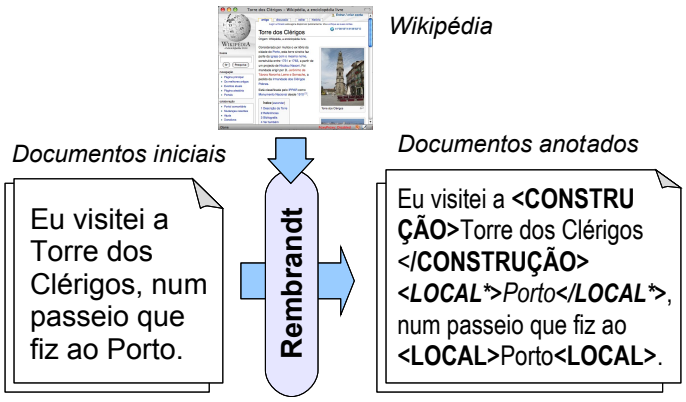

Figura 11.6: Acção do REMBRANDT na anotação de textos. Os asteriscos assinalam os locais inferidos a partir do texto.

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 (Santos et al., 2006, 2008b). 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" pode ser considerada 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 com a sua página respectiva na Wikipédia, e na análise da sua estrutura, ligações e categorias para obter mais conhecimento sobre ela. O REMBRANDT também depende de regras manuais para capturar pistas internas e externas para textos em português, tal como é descrito por McDonald (1996). As regras são usadas tanto para classificar EM 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, extraindo assim os nomes geográficos das categorias e permitindo a extracção de informação geográfica *implícita* para cada EM, como é ilustrado na figura 11.6 e descrito mais detalhadamente em Cardoso et al. (2008b).

11.3.3 MG4J

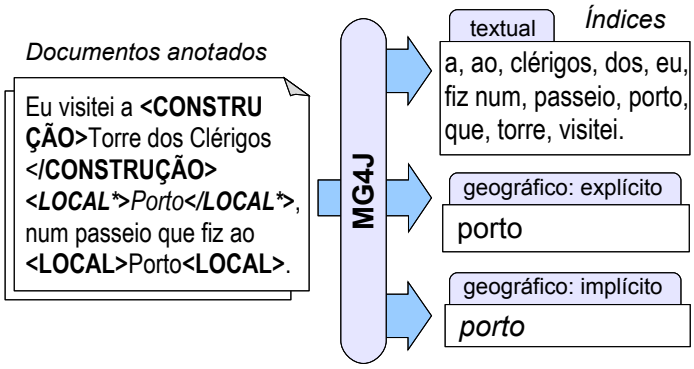

Figura 11.7: Indexação dos documentos anotados pelo MG4J. Os asteriscos assinalam os termos que serão indexados no índice geográfico implícito.

O MG4J é o módulo responsável pela indexação e ordenação dos documentos. A figura 11.7 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 ilustrado pela figura 11.7, 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.

11.3.4 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 (Auer et al., 2007), 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 consultas com os termos "Castelo Branco.". Tal como foi referido anteriormente, uma pesquisa por "Obras de Castelo Branco" muito provavelmente 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 e, como tal, é uma consulta de cariz geográfico.

Com o RENOIR, procura-se investigar novas formas de enriquecer as consultas de forma a introduzir etiquetas semânticas de um modo manual, supervisionado ou automático. Nos exemplos anteriores, as consultas 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 fornecendo documentos de acordo com o contexto correcto de "Castelo Branco".

11.4 Avaliação do desempenho dos sistemas

As avaliações conjuntas 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.

O trabalho desenvolvido no âmbito do meu 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. A participação nas pistas de avaliação é feita para as tarefas em língua portuguesa.

i. GeoCLEF

O GeoCLEF é uma pista de avaliação específica para sistemas de RIG, e que inclui o português como uma das línguas usadas nas suas tarefas de avaliação (Mandl et al., 2008). 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 (Cardoso et al., 2008a). O estado actual dos módulos e a linha de investigação até agora seguida têm sido constantemente aperfeiçoados mediante uma análise detalhada dos resultados da avaliação, sendo que, na edição de 2008 do GeoCLEF, obtivémos resultados bastante encorajadores (Cardoso et al., 2008c).

ii. 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.

iii. 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 (Santos e Cardoso, 2008; Santos et al., 2008a). O GikiP usou na sua tarefa de avaliação as porções portuguesa, inglesa e alemã de uma recolha da Wikipédia de 2006.

O RENOIR participou no GikiP de uma forma supervisionada, utilizando a Wikipédia e o REMBRANDT como fontes de informação e de extracção de conhecimento para assistir a sua 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".