SUPeRB - Gerindo referências de autores de língua portuguesa

Luís Miguel Cabral, Diana Santos, e Luís Fernando Costa
Linguateca, Oslo node, SINTEF ICT
PB 124, Blindern NO-0314
Oslo, Norway
{luis.m.cabral, Diana.Santos, luis.costa}@sintef.no

ABSTRACT

In this paper we describe SUPeRB, a digital librarian helper, which has two specific goals: update and maintain specific publication repositories; and assist in the publishing of publication records, for institutions and individual actors. It does this by gathering bibliographic data from web pages and documents in order to build a local repository of bibliographic data on a specific subject. Also, by collecting information from these resources, SUPeRB assists in building a bibliographic database with specific domain intervenients such as authors, conferences and scientific journals.

Resumo

Este artigo descreve o SUPeRB, um sistema para procurar e tratar referências bibliográficas na Web, que possui dois objectivos: actualizar e manter repositórios de publicações numa área específica; e assistir na publicação de dados bibliográficos de instituições ou de investigadores individuais. Para tal, o SUPeRB recolhe informação de páginas e documentos electrónicos, construindo um repositório local de referências bibliográficas da área. Ao construir estes recursos, o SUPeRB cria ainda uma base de conhecimento num determinado domínio, contendo autores, conferências e revistas científicas.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process, Selection process

General Terms

Extracção de informação, Gestão de informação, Referências Bibliográficas

1. INTRODUÇÃO

Desde 1999 que a Linguateca disponibiliza um portal dedicado ao processamento computacional do português com o objectivo de fornecer uma boa panorâmica a todos os interessados nesta área. O nosso objectivo foi desde o início garantir a existência de um local que permita aos investigadores e programadores seguirem o trabalho feito nesta área, de forma a evitar repetição de esforços e potenciando, ao invés, colaborações entre diferentes instituições realizando esforços complementares. Um dos recursos que mantemos é um catálogo de publicações relacionadas com o processamento computacional do português. Entre 1999 e 2003, recolhemos manualmente cerca de 750 entradas, incluindo, quando disponíveis, as suas versões electrónicas. Embora a nossa equipa acompanhe as listas de discussão e de artigos aceites em conferências relevantes para a área, chegámos à conclusão que não era fácil manter este recurso actualizado. É particularmente difícil encontrar a informação completa sobre artigos e outras publicações científicas, dado que muitos investigadores não actualizam as suas páginas de publicações frequentemente. Para além disto, é comum encontrarmos outras dificuldades para obter e processar esta informação, tais como:

- Referências incompletas, onde se omitem por exemplo os nomes completos das conferências, os editores dos volumes, as edições das conferências ou a sua localização;
- Vários estilos bibliográficos usam as iniciais dos autores, o que complica a tarefa de os identificar automaticamente;
- As versões electrónicas não são exactamente iguais às versões publicadas (pelo menos no que diz respeito à formatação).

É também de referir que quase nenhum dos autores com trabalhos no nosso catálogo usa meta-informação ou qualquer tipo de categorização dos seus trabalhos que permitisse a sua classificação automática. Normalmente as suas listas de publicações são páginas Web incluindo apenas as referências codificadas em texto e, em alguns casos, sem ligações para as versões electrónicas.

A pouca informação disponível dificulta a tarefa de decidir, apenas pelo título, se uma determinada publicação deve ser considerada relevante para o domínio que pretendemos cobrir. Adicionalmente, os utilizadores do catálogo raramente têm a motivação necessária para nos ajudar a catalogar mais publicações, sugerindo as suas próprias publicações ou outras que pensem ser relevantes.

Para além do referido anteriormente, com o enorme crescimento da informação na Web, é consensual que são necessários métodos automáticos para ajudar a organização e disponibilização deste tipo de informação.

Nesse sentido, procurámos com o desenvolvimento do SUPeRB responder à necessidade de um assistente automático para a procura e recolha de informação bibliográfica de documentos electrónicos, bem como para a avaliação da relevância desta informação dado um determinado catálogo. O nosso objectivo nunca foi criar um sistema completamente automático, mas antes desenvolver um método supervisionado para suportar a criação e manutenção de listas de publicações coerentes sobre determinadas áreas de conhecimento, sendo essa manutenção auxiliada por contribuições da comunidade trabalhando nessas àreas. O nosso objectivo era portanto semelhante ao do Feitelson [5], e de forma nenhuma uma tentativa de competir com CiteSeer [8], por exemplo.

2. SUPERB, UM ASSISTENTE PARA A GESTÃO DE PUBLICAÇÕES

O SUPeRB é portanto um sistema semi-automático com o objectivo de procurar e processar referências bibliográficas com características específicas na Web. Auxilia também especialistas na construção de colecções de meta-informação bibliográfica a partir da informação coligida por vários utilizadores. O SUPeRB é uma ferramenta que permite recolher informação a partir dos dados na Web e validar e integrar esta informação bibliográfica num catálogo de publicações. Os utilizadores do sistema podem iniciar a sua interacção com o sistema introduzindo:

- uma referência em formato de texto;
- um conjunto de palavras-chave ou expressões que são usadas para procurar páginas Web com conteúdo bibliográfico relevante;
- um URL apontando para uma página contendo uma ou mais referências bibliográficas.

Um utilizador, pode por exemplo, colocar o nome de um autor e o título de uma publicação (completo ou parcial), sendo neste caso o objectivo do SUPeRB completar a referência bibliográfica usando recursos na Web e armazená-la no catálogo de publicações, incluindo ligações para os documentos na Web quando possível.

Para concretizar estas tarefas, o SUPeRB foi desenhado como um conjunto de módulos independentes, cada um responsável por uma tarefa específica. Esta estrutura modular permite que os módulos interajam entre si ou trabalhem independentemente, possibilitando também a usa integração em aplicações externas. Pretendemos automatizar as seguintes tarefas:

- Pesquisa por palavras-chave na Web, para obter conteúdo relevante, no domínio bibliográfico;
- Extracção de texto a partir de vários formatos de documentos;
- Extracção de referências bibliográficas a partir de texto;
- Decomposição de referências bibliográficas nos seus elementos bibliográficos (título, autor, canal de publicação, título da revista científica, etc.) focando particularmente as normas de referência bibliográfica brasileiras e portuguesas [1, 6];
- Conversão de referências entre diferentes estilos bibliográficos tais como os formatos BibTeX ou EndNote;
- Comparação e fusão de referências incompletas denotando a mesma publicação;
- Melhoria de ontologias bibliográficas contendo informação sobre editoras, conferências, autores, canais de publicação, etc.;
- Armazenamento e manutenção de toda a informação recolhida.

Adicionalmente, para potenciar a utilidade dos repositórios criados, o SUPeRB permite que os utilizadores atribuam etiquetas aos elementos dos catálogos.

Tendo em conta o espírito do catálogo de publicações da Linguateca que está disponível na Web, foi também construída uma interface na Web para interagir com os módulos do SUPeRB.

O SUPeRB foi pensado fundamentalmente para dois tipos de utilizadores:

1. **Utilizadores do repositório**, que podem usar o SUPeRB para procurar e classificar referências bibliográficas relacionadas com os seus conhecimentos e interesses. Dado que cada um dos módulos produz informação intermédia, esta informação pode ser validada e/ou corrigida pelos utilizadores que podem remover informação irrelevante ou editar resultados incorrectos. É esta informação validada e/ou corrigida que é fornecida aos módulos seguintes, permitindo desta forma um melhor desempenho do sistema.
2. **Administradores do repositório**, que decidem o que fica de facto armazenado no repositório ao validar ou não as propostas de adição/edição ao catálogo feitas pelos **utilizadores do repositório**.

3. SUPERB EM PORMENOR

O SUPeRB é composto por vários módulos, cada um dos quais capaz de providenciar resultados, de forma a serem utilizados pelo sistema em geral e de forma a serem armazenados. A figura 1 apresenta a arquitectura deste sistema, descrevendo a ligação entre os vários módulos e que será descrito nesta secção.

3.1 *PersquisaWeb*

Este módulo usa termos e expressões dados pelo utilizador, de forma similar a [2], capaz de gerar pesquisas que são depois fornecidas a serviços Web, como as API do Google ou do Yahoo. O resultado deste módulo é uma lista de URL, ordenada de acordo com os resultados obtidos pelos serviços

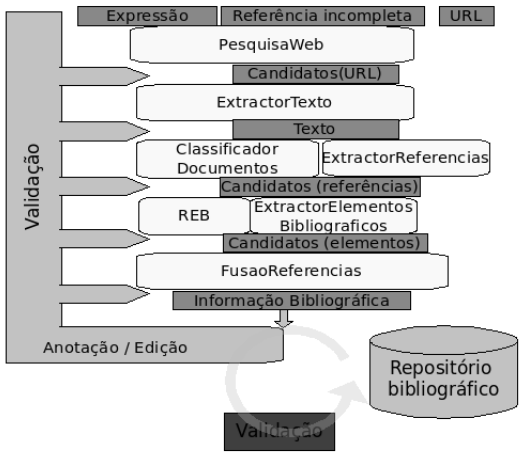

Figure 1: A arquitectura do SUPeRB

Web, tendo em consideração ainda a ocorrência de múltiplas páginas do mesmo sítio, aumentando o peso destas.

$$PontuacaoD = \Sigma_i^N (D_i) + \Sigma_i^N \frac{S_i}{N_S} \quad (1)$$

Onde D_i é o peso dado pelo serviço Web ao documento e S_i é o peso de outro documento encontrado no mesmo sítio do documento D. N_S é o número de ocorrências do sítio S.

Os dados fornecidos pelo utilizador podem ainda ser estruturados num contexto bibliográfico, onde elementos como o autor, o título ou o ano podem ser especificados, permitindo gerar uma combinação de pesquisas que contêm os elementos mais relevantes.

3.2 *ExtractorTexto*

Este módulo extrai o conteúdo a partir de vários tipos de formatos de documentos (HTML, Microsoft Word e PowerPoint, Open Office Word e apresentações, Adobe PDF, PostScript, Rich Text Format), para um formato comum, texto. Este módulo tem como entrada o nome de um documento, local ou remoto, e devolvendo como resultado o seu conteúdo em texto não formatado. Tentamos garantir neste módulo a simplicidade de configurar outros programas para facilitar a conversão de outros formatos para texto.

3.3 *ExtractorReferencias*

Este módulo trata o texto, produzido por *ExtractorTexto*, extraindo uma lista de candidatos a referências bibliográficas. Esta tarefa é ainda assistida por um outro módulo, *ClassificadorDocumentos*, um módulo que possui um conjunto de heurísticas tentar identificar a estrutura do documento original, seja ele um artigo académico, lista de referências ou mesmo uma apresentação. Pode ainda identificar na estrutura de documentos irrelevantes, como por exemplo um blogue. Esta tarefa baseia-se na ocorrências de palavras específicas, no tamanho do documento e das frases, e na comparação com estruturas pré-definidas.

Este módulo reconhece ainda diversos formatos bibliográficos estruturados que podem ser tratados pelo módulo *ConversorReferencias*, que será descrito mais à frente, podendo ser tratado directamente por esse mesmo módulo.

Após a identificação da estrutura do documento original, os blocos de texto com maior probabilidade de conterem informação bibliográfica são analisados. Estes podem incluir:

1. Processar o início do documento para obter autores, título, resumo ou outra informação bibliográfica, de um documento académico;
2. Processar o final de documentos académicos para obter as referências citadas num documento académico.
3. Processar um documento completo, uma lista de referências.

Uma bateria de métodos heurísticos é usada nas áreas candidatas, procurando marcas comuns de referências, tal como enumerações, quebras de linha e marcas de estilos bibliográficos.

Numa segunda fase, os candidatos são filtrados, garantindo que não são demasiados pequenos ou grandes, verificar a sua posição no texto, para garantir que não á sobreposição dos candidatos.

3.4 *ExtractorElementosBibliograficos*

Este módulo processa referências bibliográficas individuais. Recebe uma referência em formato texto e devolve como resultado os seus elementos bibliográficos de forma estruturada. O formato de saída é consistente com o formato de armazenamento do catálogo de publicações, descrito em [9, 3]. Este módulo recorre a vários métodos, usando um pacote Perl descrito por Jewell em [7], juntamente com um conjunto de heurísticas para atomização do texto e ainda um repositório de elementos bibliográficos, o *REB*. O *REB* é um módulo composto por uma ontologia de autores, editores, locais e conferências, construído a partir de dados recolhidos do actual catálogo de publicações da Linguateca. Actualmente o *REB* contém 1993 autores e editores, 548 nomes de conferências e acrónimos de conferências, 185 editoras e ainda 132 localizações. O *REB* proporciona ainda métodos para validar novos dados (identificando duplicados e erros ortográficos) e para aplicar essses dados. Esta informação é ainda usada pela interface Web, providenciando sugestões na inserção de dados, para diversos campos dos formulários (autor, editores, conferências, etc.)

3.5 *FusaoReferencias*

Uma vez que vários documentos distintos podem ser analisados, a mesma referência pode ser obtida mais do que uma vez, pelo que existe a necessidade de identificar duplicados ou equivalentes. É por esta razão que este módulo foi desenvolvido, permitindo ainda combinar ou omitir os diferentes elementos bibliográficos dos diferentes duplicados. Como este módulo trata referências bibliográficas devidamente estruturadas, ele requer a combinação de diferentes campos bibliográficos para considerar duas referências duplicadas ou referencestes À mesma publicação. Para além disto, as referências bibliográficas armazenadas no catálogo são também tidas em consideração, permitindo assim a sua actualização.

3.6 Outros módulos e opções de desenho do sistema

Existem ainda outros módulos que fornecem métodos importantes na realização de diversas tarefas:

1. O*AnotadorReferencias*, é uma interface que permite aos utilizadores associar palavras-chave à informação já recolhida, proporcionando informação útil para a pesquisa e organização dentro do catálogo;
2. O *ConversorReferencias* é um módulo que permite a conversão entre diversos formatos bibliográficos conhecidos (incluindo o formato interno de armazenamento do catálogo): BibTeX, RIS, EndNote e Refer. Isto facilita a exportação de referências bibliográficas no formato mais conveniente, bem como utilização de diversos módulos no formato mais familiar ao utilizador.

É ainda de notar que no desenho do SUPeRB, foi contemplado o factor temporal de manutenção: É frequente que diversos elementos bibliográficos, por exemplo as páginas, ou o link onde um artigo foi publicado, possam ser omitidos no primeiro registo da referência no sistema. Outro caso a considerar é quando um artigo possa ser republicado, onde relações/links necessitam ser adicionados. Na internet, os links mudam frequentemente e é necessário um sistema automático, que periodicamente verifique a disponibilidade e exactidão dos links. Assim, tomamos em consideração a criação de uma funcionalidade que permita ao SUPeRB actualizações periódicas (ou marcadas para uma data específica).

Por fim, tem sido considerado a disponibilização do sistema num ambiente multilingue, que permita obter citações da mesma referência bibliográfica em português, inglês ou outro contexto linguístico, o que requer a necessidade de gerar equivalentes linguísticos para localizações, editoras e data. Actualmente o português e inglês são suportados, estando em consideração extender esta funcionalidade para outras línguas.

4. AVALIAÇÃO DOS COMPONENTES DO SUPERB

O módulo *ExtractorReferencias* foi anteriormente avaliado usando uma metodologia inspirada no HAREM [11, 10]. Neste artigo avaliamos o *AnalisadorReferencias*, um dos módulos centrais do SUPeRB, também segundo a mesma filosofia, classificando os elementos bibliográficos de acordo com a seguinte grelha:

- Correcto (Elemento correctamente delimitado e classificado);
- Errado (Elemento correctamente delimitado mas mal classificado);
- Incompleto (Elemento correctamente classificado mas incompleto);
- Excessivo (Demasiado material. Estaria correctamente classificado se apenas se considerasse parte do elemento);
- Em falta (Elemento que falta, provavelmente erroneamente incluído num elemento excessivo).

Com base nesta classificação, calculámos depois as seguintes medidas:

$$Precisao = \frac{\#c}{\#tp} \quad (2)$$

$$Abrangencia = \frac{\#c}{\#te} \quad (3)$$

$$Precisaoalargada = \frac{\#c + \#e}{\#tp} \quad (4)$$

$$Abrangenciaalargada = \frac{\#c + \#e}{\#te} \quad (5)$$

$$Sob - geracao = \frac{\#i + \#ef}{\#tp} \quad (6)$$

$$Sobre - geracao = \frac{\#err}{\#tp} \quad (7)$$

$\#c$ denota o número de elementos correctos, $\#tp$ o número total de elementos propostos, $\#te$ o total de elementos esperados, $\#e$ o número de elementos em excesso, $\#i$ o número de elementos incompletos, $\#ef$ o número de elementos em falta, e $\#err$ o número de elementos errados. A tabela 1 apresenta alguns resultados desta avaliação, usando 33 referências bibliográficas extraídas de 33 páginas pessoais de investigadores activos na corpora-list ou fazendo parte da comunidade de processamento da língua portuguesa.

Note-se que ignorámos algumas distinções entre campos, tais como entre autores e editores (concentrando-nos na questão da correcta detecção de nomes de pessoas), entre nome da conferência ou nome abreviado da mesma, e entre local de edição e local da conferência. De 239 elementos esperados, 102 foram identificados correctamente, 47 erradamente, e 84 encontravam-se em falta. Este estudo, além dos resultados globais, permitiu-nos também analisar o desempenho na detecção de alguns elementos em particular:

- A identificação dos nomes de autores tem uma precisão elevada (sobretudo no caso de autores como nomes portugueses: os piores casos são alguns nomes estrangeiros);
- Exceptuando o ano, elementos numéricos não trazem dificuldades;
- O módulo REB pode dar origem a sobre-geração de alguns campos, provocando ruído;
- No que se refere a editoras, por outro lado, o mesmo módulo ainda tem uma cobertura deficiente;
- Os campos do REB não são necessariamente mutuamente exclusivos: universidades podem funcionar como editoras, uma mesma pessoa pode ser autor e editor, e assim para distinguir entre diversas funções mais atenção ao contexto na referência é necessário.

5. OBSERVAÇÕES FINAIS

Um problema concreto na actividade quotidiana do nosso projecto levou-nos a aplicar ferramentas desenvolvidas para

Table 1: Avaliação do módulo *AnalisadorReferencias*

	Precisão	Abrangência	Medida-F	Prec.-A	Abr.-A	Sob-ger.	Sobre-ger.
autor	0.72	0.40	0.26	1.00	0.56	0.44	0.00
ano	0.41	0.50	0.23	0.80	0.97	0.03	0.21
título	0.39	0.57	0.23	0.50	0.73	0.27	0.43
conferência	0.36	0.44	0.20	0.45	0.56	0.44	0.39
local	0.75	0.40	0.26	0.75	0.40	0.60	0.00
pages	0.83	0.77	0.40	0.92	0.85	0.15	0.08
volume	1.00	0.33	0.25	1.00	0.33	0.67	0.00
instituição	0.33	0.40	0.18	0.50	0.60	0.40	0.50
Total médio	0.60	0.427	0.25	0.74	0.62	0.38	0.20

o processamento computacional da nossa língua para o resolver e acabámos por desenvolver um sistema mais geral que esperamos que possa ajudar investigadores ou bibliotecários especializados no trabalho com referências numa área específica.

De facto, a maior parte das citações, assim como a ordenação dos pesquisadores, mesmo quando os autores são brasileiros ou portugueses, é feita em publicações internacionais, ou seja, inglês. Por essa razão, havia muito pouco trabalho feito no assunto em ou sobre o português (embora também haja publicações internacionais em português).

Não conseguimos encontrar nenhum sistema desenvolvido especialmente para lidar com referências em português, nem com referências de autores de língua materna portuguesa (em português, inglês ou outras línguas).

Além de apoiar a gestão de um catálogo de cerca de 2000 publicações, e incluindo 113 conferências, livros ou revistas distintas, os módulos do SUPeRB estão disponíveis como código aberto (na linguagem Perl), para poderem ser usados em projectos que tratem de referências em português. De notar pois que podem ser usados independentemente da interface na rede.

No futuro, tencionamos melhorar o SUPeRB com palavras-chave e resumos, assim como pretendemos incorporar ontologias relativas aos assuntos constantes na base, obtidas a partir do tratamento do texto dos próprios artigos quando acessível electronicamente.

Embora as referências possam ser consideradas como um subtipo de texto semi-estruturado, e como tal ser vistas como uma fonte relevante de extracção de informação a partir de texto científico [4], esta actividade tem até agora sido completamente negligenciada para o português.

Agradecimentos. Este trabalho foi feito no âmbito da Linguateca, projecto financiado através do contrato nº 339/1.3/C/NAC, pelo governo português e pela União Europeia, e executado pela FCCN.

6. REFERENCES

[1] Associação Brasileira das Normas Técnicas. *NBR 6023: Norma Brasileira*, Agosto de 2002.

[2] Marco Baroni e Silvia Bernardini. Bootcat: Bootstrapping corpora and terms from the web. In Maria Teresa Lino et al., editoras, *Proceedings of LREC2004, the 4th International Conference on Language Resources and Evaluation(LREC'2004, Lisboa, Portugal)*, 26-28 de Maio de 2004.

[3] Luís Miguel Cabral. Documentação online do SUPeRB. http://adamastor.linguateca.pt/super/docs/help.html, Julho de 2007. Última actualização 5 de Agosto 2008.

[4] Fabio Ciravegna, Sam Chapman, Alexiei Dingli e Yorick Wilks. Learning to harvest information for the semantic web. In *1st European Semantic Web Symposium (ESWS-2004, Heraklion, Grécia)*, Maio de 2004.

[5] Dror G. Feitelson. Cooperative Indexing, Classification and Evaluation in BoW. In *Proceedings of the 7th International Conference on Cooperative Information Systems*, 2000.

[6] Instituto português da Qualidade. *NP 405-2: Norma Portuguesa: Documentos electrónicos*, 2003.

[7] Mike Jewell. ParaTools Reference Parsing Toolkit-Version 1.0 Released. *D-Lib Magazine*, 9(2), 2003.

[8] Steve Lawrence, C. Lee Giles e Kurt Bollacker. Digital Libraries and Autonomous Citation Indexing. *IEEE Computer Society Press*, 32(6), 1999.

[9] Paulo Alexandre Rocha. Gestão das Páginas do projecto: Processamento Computacional do Português. Technical report, Departamento de Informática, Universidade de Braga, 10 de Novembro de 2001.

[10] Diana Santos e Nuno Cardoso, editores. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 12 de Novembro de 2007.

[11] Diana Santos, Nuno Seco, Nuno Cardoso e Rui Vilela. HAREM: An Advanced NER Evaluation Contest for Portuguese. In Nicoletta Calzolari et al., editores, *Proceedings of the 5th International Conference on Language Resources and Evaluation(LREC'2006, Génova, Itália)*, pp 1986–1991. ELDA, 22-28 de Maio de 2006.