FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

SUPeRB
Sistema Uniformizado de Pesquisa de Referências Bibliográficas

Luís Miguel Cabral

Mestrado em Engenharia Informática
Porto, Março de 2007

Faculdade de Engenharia da Universidade do Porto

SUPeRB
Sistema Uniformizado de Pesquisa de Referências Bibliográficas

Luís Miguel Cabral
Licenciado em Ciência de Computadores pela Faculdade de Ciências da Universidade do Porto

Dissertação submetida para satisfação parcial dos requisitos do grau de mestre em Engenharia Informática

Dissertação realizada sob a supervisão de Professor Doutor Eugénio de Oliveira, Departamento de Engenharia da Faculdade de Engenharia da Universidade do Porto
e
Doutora Diana Santos, SINTEF ICT, Oslo

Porto, Março de 2007

Resumo

As publicações científicas são um elemento importante na investigação científica de qualquer domínio. Por um lado, são representativos do estado da arte desse domínio; por outro, constituem a base para outros estudos e publicações. São, em suma, uma base do conhecimento científico. Não é portanto de admirar que existam actualmente tantos esforços para manter a informação bibliográfica actualizada em repositórios e bases de dados que representam domínios, instituições, organizações ou apenas pessoas individuais. Assiste-se ainda a uma proliferação de motores de pesquisa bibliográficos que visam facilitar o acesso a uma colecção de referências bibliográficas.

O objectivo deste trabalho consiste em desenvolver um sistema de pesquisa de referências bibliográficas, o SUPeRB, que, de forma semi-automática, assista na manutenção de um repositório dedicado ao processamento computacional da língua portuguesa, o catálogo de publicações da Linguateca. O catálogo de publicações da Linguateca oferece um serviço em que qualquer pessoa pode inserir e pesquisar referências bibliográficas na área do processamento computacional da língua portuguesa. No entanto, existe um processo de validação nos bastidores, necessário para manter a qualidade do recurso, mas que é também bastante penoso para o gestor deste recurso. Com o SUPeRB, pretende-se aliviar todo o processo de inserção e validação, usando o sistema desenvolvido para pesquisar informação adicional relacionada.

O sistema proposto recorre a consultas na Web para obter documentos que possam conter informação bibliográfica relevante e usa métodos de extracção de informação da Web para obter essa informação. São também utilizadas tecnologias como os serviços Web para obter informação estruturada de repositórios bibliográficos, dado que as referências bibliográficas são por natureza um conjunto de elementos bibliográficos semi-estruturados.

A integração das várias tecnologias da Web 2.0 é uma das contribuições deste trabalho, tal como a própria arquitectura do sistema e o conjunto de módulos desenvolvidos, publicamente disponíveis e utilizáveis noutros contextos.

Abstract

Scientific publication is an important part of the research in any domain. It represents both the state of the art and represents scientific knowledge for future studies and publications. Therefore there are many efforts to maintain bibliographic references up to date, grouped both in public and private repositories and databases representing collections on certain domains, organizations or just of private persons. Furthermore, there is an upsurge of dedicated search engines that index bibliographic references with the sole aim of facilitating their future retrieval.

The objective of this thesis is to develop a semi-automatic system, SUPeRB, that assists in the discovery of bibliographic references. SUPeRB's main function is to help managing Linguateca's publication catalogue, a bibliographic repository dedicated to natural language processing of the Portuguese language. This publication catalogue allows any person to insert a publication and browse and search this repository. But the validation procedure associated to each inserted publications, required to maintain the quality of the catalogue, is very costly. Before SUPeRB it implied an entirely human effort. SUPeRB was design to relieve the human from pat of this process, by collecting possible candidates that either support, update or supply related information.

A new system is proposed that *(a)* obtains relevant information from documents on the Web; *(b)* uses Web service technologies that return structured information from bibliographic repositories; *(c)* and parses text and references into fine-grained elements. Finally, the integration of several Web 2.0 technologies is another contribution of this thesis. A novel architecture is proposed and the modules developed are freely available on the Web and can be used in other domains.

Agradecimentos

Desejo agradecer a todas as pessoas que contribuíram directa e indirectamente para a realização desta tese de mestrado e sem o qual este trabalho teria teria sido possível. Agradeço aos meus orientadores, o Professor Doutor Eugénio de Oliveira da Departamento de Engenharia da Faculdade de Engenharia da Universidade do Porto, mas principalmente à Doutora Diana Santos, do SINTEF ICT, Oslo, pela orientação e pelo encorajamento que sempre me deram e acima de tudo pela paciência. Agradeço-lhes profundamente pela confiança que depositaram em mim.

Agradeço ainda ao Luís Sarmento, pela ajuda e pelos conselhos, ao Luís Costa pelas críticas construtivas, pela troca de ideias e pela revisão do texto. Deve ser ainda mencionado que a versão original do Capítulo 5, de avaliação, foi originalmente concebida e redigida pelo, Luís Sarmento, pela Diana Santos e por mim próprio. Agradeço-lhes pela contribuição dada neste capítulo.

Aproveito para agradecer a todos os outros elementos da equipa da Linguateca que directa ou indirectamente colaboraram ou que tiveram paciência para esperar um pouco mais.

Este trabalho foi financiado pela Fundação para a Ciência e Tecnologia (FCT), através dos projectos POSI/-PLP/43931/2001, co-financiado pelo POSI, e pelo projecto POSC/339/1.3/C/NAC.

À minha mãe.

Àqueles que precisaram de mim quando eu não estava lá...

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

A partilha de informação é uma das principais bases da investigação científica. Novos avanços tecnológicos e trabalhos académicos que visam o avanço tecnológico são anualmente apresentados em conferências internacionais. Como resultado dessas conferências, e com vista à disseminação da informação, são criados volumes impressos que compilam os trabalhos que foram apresentados em cada conferência. Outros meios de divulgação de informação científica são as revistas , que têm o mesmo fim. Estas conferências, livros e revistas científicas, referem-se habitualmente a domínios bastante específicos, como a linguística, a inteligência artificial, a genética ou a bioinformática, ou até sub-disciplinas destes domínios. De facto, existe um universo de conferências, livros e revistas que abordam e apresentam domínios específicos.

Actualmente, este tipo de disseminação está a evoluir. Com o aparecimento da World Wide Web (WWW ou Web), surgiu uma nova forma de divulgação: o formato electrónico e subsequente distribuição através da Web. A Web foi criada por Tim Berners-Lee no início da década de 90, com um propósito simples mas ambicioso:

> *The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.(Berners-Lee, 1992)*

Em pouco mais de dez anos pode dizer-se que este objectivo, o de providenciar acesso universal a um universo de documentos, foi para além das expectativas. Apesar de não se saber o tamanho exacto da Web, é possível fazer estimativas recorrendo ao número de páginas indexadas pelos motores de pesquisa. Em 2004 o Google[1] anunciou 8 biliões de páginas indexadas, o MSN[2] declarou 5 biliões e o Yahoo[3] 4,2 biliões. Apenas um ano antes, os valores eram menos de metade. O maior número de páginas indexadas fora igualmente anunciado pelo Google, aproximadamente 3,5 biliões. Num estudo recente, Gulli e Signorini (2005) calculam que em 2005 existissem mais de 11,5 biliões de páginas indexáveis. O número de utilizadores que acedem à Web tem tido também um crescimento acelerado. Entre 2000 e 2005, o número de utilizadores teve um crescimento de 182%, estimando-se que existam cerca de 6,5 biliões de utilizadores da Web em 2006, ou seja 15% da população mundial (Internet users Statistics). Apenas os utilizadores do Estados Unidos da América (68% da população) e da Europa (40% da população) perfazem um bilião de utilizadores.

A comunidade científica e as entidades divulgadores de informação científica são, portanto, um dos muitos intervenientes neste crescimento da Web. Apesar da divulgação científica electrónica não apresentar um crescimento tão rápido como a sua plataforma de difusão, a Web, apresenta um crescimento entre os 50% e os 100%, em acessos a publicações, como mostrou Odlyzko (2002), que apresenta como exemplos de bibliotecas electrónicas a Biblioteca do Congresso americano[4], a biblioteca do AT&T Labs - Research [5]e as páginas pessoais.

As publicações electrónicas têm tido em geral uma boa aceitação. Ainda assim, nem todos os domínios mostram essa mesma aceitação pelo novo formato electrónico. O estudo apresentado em Anderson et al. (2001) descreve uma situação em que os autores de um artigo consideraram um erro ter publicado o artigo na versão *online-only* da revista *Pediatrics*[6]. Esse artigo foi o mais citado das publicações apenas electrónicas (*online-only*) da revista *Pediatrics*

[1] `http://www.google.com`
[2] `http://searc.msn.com`
[3] `http://www.yahoo.com`
[4] *Library of Congress*, acessível em `http://www.loc.gov/index.html`
[5] `http://public.research.att.com/`
[6] `http://www.pediatrics.org`

no período de três anos, tendo tido 38 citações, apenas menos 20 citações do que o mais citado dos artigos impressos.

Permitir disponibilizar um documento em formato electrónico, a nível mundial, e que pode ser transferido para o nosso computador em qualquer altura, é o que a Web oferece. Esta nova forma de distribuição originou um novo conceito, *Open Access* (OA), a disponibilização livre de conteúdos científicos. Normalmente é o autor (ou a instituição a que o autor pertence) que paga os custos de publicação, em alternativa ao modelo baseado na assinatura para obter o reembolso dos custos. Esta é uma alternativa sem fins lucrativos. O OA permite manter o conceito de *revisão pelos pares* (*peer-review*), ou seja, o trabalho é avaliada e revisto por outros investigadores com conhecimentos na área, de forma a comprovar a qualidade das publicações.

Rapidamente a publicação electrónica de documentos académicos na Web tomou um lugar na comunidade científica sob inúmeras formas:

- Os investigadores disponibilizam a sua bibliografia pessoal online, reunindo o conhecimento e trabalho dessa pessoa numa determinada área (por vezes mais do que uma área).

- As revistas científicas disponibilizam versões na Web, facilitando o acesso a artigos através do formato electrónico, gratuitamente ou restringindo o acesso a assinantes.

- Têm sido criados repositórios que indexam as publicações e respectivas referências bibliográficas de áreas específicas.

- Têm sido criados motores de pesquisa específicos para publicações científicas, que permitem uma pesquisa de publicações mais eficiente.

- Existem sítios online que permitem a gestão de publicações e referências bibliográficas.

Não é portanto por acaso que os documentos científicos disponibilizados na Web são cada vez mais citados, tal como é apresentado na Figura 1.1, extraída de Lawrence et al. (1999).

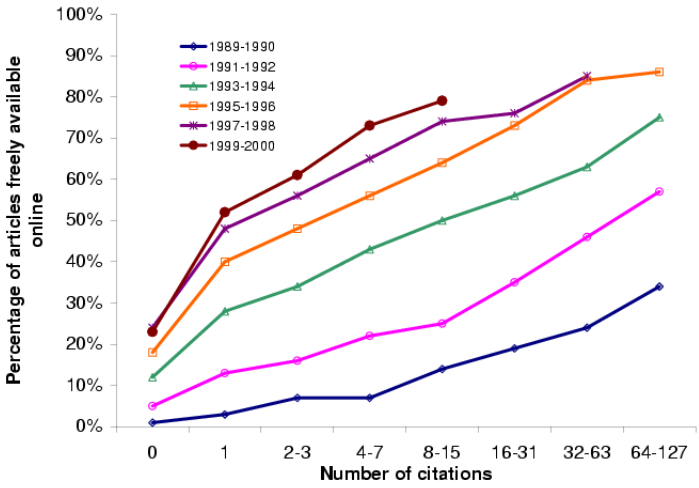


Figura 1.1: Análise a 119.924 documentos online na área de ciência de computadores e áreas relacionadas (Lawrence et al., 1999)

Apesar de se poder apenas especular sobre as razões que levam ao aumento dos acessos e das citações a publicações electrónicas, somos levados a acreditar que isto não derive inteiramente dos custos ou da qualidade das publicações. Uma das principais razões apresentadas e que justificaria este aumento é porque simplesmente as pessoas preferem aquilo a que podem aceder imediatamente (Odlyzko (2002) e Stevens-Rayburn e Bouton (1998)). Uma publicação disponível no formato electrónico pode ser encontrada e impressa em poucos minutos.

Mas esta dissertação não pretende avaliar nem comparar citações de publicações impressas e publicações online em formato electrónico. Nesta dissertação pretende-se abordar um caso real, um recurso bibliográfico, o catálogo de publicações da Linguateca[7]. O catálogo de publicações da Linguateca é um repositório bibliográfico, que contém referências bibliográficas relacionadas com o processamento computacional da língua portuguesa. Mas a tarefa de manutenção deste catálogo, como de qualquer outro, é difícil. Assim, pretende-se colmatar algumas das dificuldades sentidas na manutenção deste recurso através do desenvolvimento de um sistema capaz de complementar o repositório, sendo capaz de pesquisar documentos na Web que contenham informação relevante e de processar essa informação de forma a obter mais e

[7] `http://www.linguateca.pt`, ver catálogo de publicações

melhor informação bibliográfica.

1.2 Terminologia

Antes de prosseguir, é necessário clarificar um pouco a terminologia empregue ao longo desta dissertação, explicando de seguida alguns dos conceitos usados.

1.2.1 Referências e elementos bibliográficos

A referência bibliográfica é um conjunto de elementos bibliográficos que permite identificar um documento ou parte desse documento ((NP 405-1; NBR 6023)), quer em formato impresso quer electrónico. As referências bibliográficas podem-se referir a documentos como livros, actas, revistas, relatórios, manuais ou partes destes, como artigos em revistas ou livros ou capítulos de livros. Partes da referência como autor, título, ano, nome da conferência ou nome da revista são exemplos de elementos bibliográficos. Na figura 1.2 é possível ver os elementos bibliográficos destacados numa referência bibliográfica,

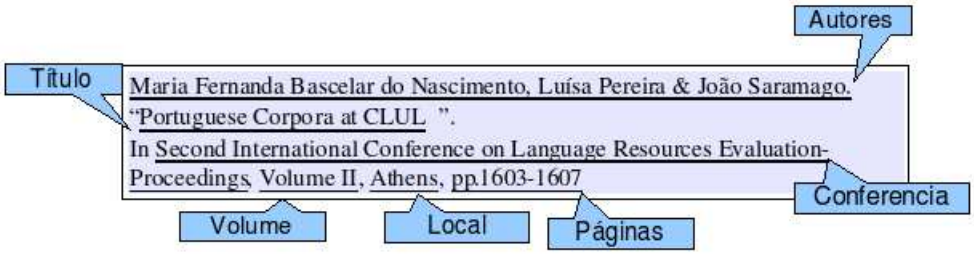


Figura 1.2: Exemplo de uma referência bibliográfica com os elementos bibliográficos autores, título, título da conferência, volume, local da conferência e páginas marcados.

Esta é a informação bibliográfica que se pretende obter e construir a partir de informação incompleta. Mas é necessário analisar como esta informação bibliográfica pode ser e é representada em documentos em geral, incluindo na Web. Podemos então considerar duas formas distintas para representar referências bibliográficas:

- Em texto simples, onde todos os elementos são apresentados sem qualquer separador específico, quase em linguagem "natural", como apresentado na figura 1.2. Uma referência bibliográfica pode ser representada em vários estilos, alterando a disposição e apresentação dos elementos bibliográficos no texto. Diferentes formas de representação gráfica constituem diferentes estilos bibliográficos.

- Outro modo de representar referências é num formato estruturado onde cada elemento bibliográfico está devidamente identificado e delimitado. Esta forma de representação será designado de formato bibliográfico. Existem vários formatos bibliográficos, mas são distintos, facilmente reconhecíveis e o seu objectivo é poderem ser processados por programas com uma certa facilidade.

São precisamente as diferenças entre cada uma destas representações que justificam o seu uso. Os estilos bibliográficos têm como finalidade ser lidos por seres humanos, necessitam ser "legíveis", ajustando-se às necessidades da publicação que representam ou do domínio a que pertencem, exibindo ou ocultando diferentes elementos bibliográficos.

Os formatos bibliográficos, por outro lado, foram desenhadas para ser legíveis por programas, de forma a serem arquivados ou para produzir representações num determinado estilo bibliográfico. É vital que se possa distinguir sem ambiguidade todas as partes da referência. É possível fazer a transformação de qualquer formato para um qualquer estilo bibliográfico. No entanto, o processo inverso não tem necessariamente de ocorrer. De seguida apresenta-se cada uma destas representações mais em pormenor.

1.2.2 Estilos bibliográficos

As referências bibliográficas são quase sempre representadas na forma de linguagem quase natural, com estilos bibliográficos diferentes. O uso de estilos bibliográficos distintos troca a ordem ou representação de diversos elementos bibliográficos. Alguns dos elementos bibliográficos podem ser abreviados ou reformatados, como ocorre frequentemente com os nomes próprios, onde as duas situações se verificam.

```
‘‘Caroline Gasperin’’
‘‘Gasperin, Caroline’’
‘‘Caroline V. Gasperin’’
‘‘Gasperin, Caroline V.’’
‘‘Caroline Varaschin Gasperin’’
...
```

A ordem dos elementos bibliográficos nos estilos bibliográficos pode variar dependendo do domínio em que se inserem. Exemplo disto são certos estilos que apresentam primeiro os nomes dos autores seguidos do título, outros apresentam o título seguido dos autores. Certos elementos bibliográficos podem mesmo ser omitidos. Por exemplo, o local de edição é usado na tradução anglo-saxónica mas geralmente é omitido na portuguesa.

É ainda possível destacar os elementos bibliográficos usando aspas (*"*), parênteses curvos (*(* e *)*), ou alterado a própria formatação do texto (*itálico*, **negrito** ou sublinhado). O tipo de destaque mais frequente é o uso do itálico. Em situações em que não é possível usar itálico (nas máquinas de escrever, por exemplo), alguns estilos recomendam o uso de outra marcação, tal como o uso de sublinhado em alternativa.

1.2.2.1 Normas internacionais e nacionais

A norma ISO 690:1987 especifica como estruturar publicações como monografias, livros, capítulos, artigos, normas, relatórios, teses, etc., nomeadamente especificando como estes documentos devem ser citados, podendo ser interpretada como um estilo bibliográfico. Foi complementada pela ISO 690-2 em 1997 para fornecer informação sobre documentação electrónica. Da mesma forma, as normas portuguesa (NP 405-1 e NP 405-2) e a brasileira (NBR 6023) são normas nacionais, harmonizadas com a ISO 690:1987 e ISO 690-2 para providenciar informação sobre como referir documentos na língua portuguesa.

De seguida é apresentado um exemplo de uma referência bibliográfica, extraída da NP 405-1.

> PAIVA, José Pedro - Medo e necessidade. Coimbra: [s.n], 1990. Trabalho de síntese apresentado à Faculdade de Letras como prova de capacidade científica.

Estas normas têm como objectivo clarificar e uniformizar a especificação de referências bibliográficas. No entanto, estas normas nem sempre são aplicadas e em alguns casos encontram-se omissas. Existem inúmeros estilos de representação bibliográfica que tentam representar a informação bibliográfica de uma forma especializada para cada domínio, dependentes da língua em que são escritos ou das conferências em que são apresentados.

1.2.2.2 Normas por domínio

Para além das normas referidas na secção anterior, existem outras formas de representar referências bibliográficas. Existem estilos, definidos geralmente para o inglês, utilizados em domínios distintos. Estes estilos bibliográficos destinam-se a representar vários tipos de publicação, inclusive artigos em formato electrónico e mesmo para repositórios específicos, como o ERIC[8], um repositório para publicações dedicadas ao ensino.

De seguida são apresentados alguns dos estilos mais conhecidos, como exemplo:

APA - O APA (*American Psychological Association*) destina-se às áreas da psicologia, educação e outras ciências sociais. É também usado em dissertações (nestas áreas).

> Anderson, K. et al. (2001). Publishing online-only peer-reviewed biomedical literature: Three years of citation, author perception, and usage experience. *Journal of Electronic Publishing*, 6(3).

Chicago - Também conhecido como CMA (*The Chicago Manual of Style*) este estilo é aplicado em livros, revistas, jornais e outros tipos de publicações não académicas.

> Anderson, K. et al. 2001. Publishing online-only peer-reviewed biomedical literature: Three years of citation, author perception, and usage experience. *Journal of Electronic Publishing*, 6 (3).

Turabian - Para aplicar em trabalhos académicos por estudantes (relatórios, monografias, dissertações). O formato Turabian é uma forma condensada do Chicago, omitindo alguns elementos, como o volume por exemplo.

[8] `http://searcheric.org/` e `http://eric.ed.gov`

> Anderson, K. et al. 2001. Publishing online-only peer-reviewed biomedical literature: Three years of citation, author perception, and usage experience. *Journal of Electronic Publishing*, 6(March).

MLA - O estilo MLA (*Modern Language Association*) é muito utilizado em documentos académicos nas áreas de Letras, artes e humanidades.

> Anderson, K. et al. "Publishing online-only peer-reviewed biomedical literature: Three years of citation, author perception, and usage experience". Journal of Electronic Publishing, 6.3 (2001).

AMA - Aplicado nas áreas da saúde, medicina e outras ciências biológicas.

> Anderson, K. et al. Publishing online-only peer-reviewed biomedical literature: Three years of citation, author perception, and usage experience. *Journal of Electronic Publishing.* 2001; 6.

1.2.3 Formatos bibliográficos

As referências bibliográficas podem ainda ser representadas em formatos estruturados, o que usualmente facilita a importação, exportação e até mesmo o armazenamento da informação bibliográfica.

A maioria destes formatos estão associados a programas (ou empresas) que acabaram por criar especificações próprias do seu próprio formato. Dada a facilidade de manipulação destas estruturas, estes formatos são frequentemente o meio de exportação por muitos dos repositórios na Web e também o meio de importação de muitos gestores online de referências bibliográficas para serem usados por utilizadores.

Alguns dos mais conhecidos e utilizados, como o BibTeX ou o EndNote/Refer, são de seguida apresentados.

1.2.3.1 BibTeX

O BibTeX é um programa e um formato que foi criado em 1986 (Lamport (1986)) para complementar o sistema de preparação de documentos em LaTeX. O formato BibTeX é provavelmente um dos formatos mais comuns para referências bibliográficas na Internet. Vários repositórios online permitem a

apresentação de referências bibliográficas neste formato, tais como o CiteSeer, ou o catálogo da Linguateca.

O formato BibTeX é um formato organizado por campos e, dado que o programa BibTeX ignora os campos desconhecidos, é facilmente expansível, podendo manter campos utilizados por outras aplicações. O seguinte exemplo inclui um campo *abstract*.

```
@article{Gettys90,
  author = {Jim Gettys and Phil Karlton and Scott McGregor},
  title = {The {X} Window System, Version 11},
  journal = {Software Practice and Experience},
  volume = {20},
  number = {S2},
  year = {1990},
  abstract = {A technical overview of the X11 functionality.
This is an update of the X10 TOG paper by Scheifler \& Gettys.}
}
```

O programa BibTeX recorre a ficheiros de estilos para assim poder gerar uma lista de citações na forma de qualquer tipo de citação desejada, podendo o próprio utilizador produzir os seus próprios estilos, com um mínimo de conhecimento de LaTeX.

O BibTeX aceita vários tipos de publicações (*article*, *book*, *booklet*, *conference*, *inbook*, *incollection*, *inproceedings*, *manual*, *misc*, *phdthesis*, *mscthesis*, *proceedings*, *techreport* e *unpublished*), sendo este identificado pela precedência de um símbolo @ e ao qual se seguem as chavetas ('{' e '}') dentro das quais ficam os vários elementos bibliográficos.

Para cada um destes tipos, o BibTeX associa um conjunto obrigatório de elementos bibliográficos, permitindo ainda um conjunto opcional, que é usado se presente mas que não causa problemas caso ausente. No entanto, é aconselhável a inclusão destes campos, não só para ter a informação mais completa, mas também para ajudar o leitor. Por exemplo, para o tipo *artigo* existem as seguintes campos:

Obrigatórios : *author*, *title*, *journal* e *year*.

Opcionais : *volume*, *number*, *pages*, *month* e *note*.

Os restantes campos são, normalmente, ignorados para este tipo de publicação. No entanto, a sua utilização pode também depender do estilo usado pelo programa BibTex para gerar uma representação da referência bibliográfica, podem ser especificados estilos que usem outros elementos. O BibTex é facilmente expansível.

Outra particularidade do BibTeX é o uso de referências cruzadas, no sentido de que utiliza mais do que uma entrada BibTex para gerar uma referência. O exemplo seguinte demonstra o uso de referências cruzadas.

```
@inproceedings{no-gnats,
    crossref = "gg-proceedings",
    author = "Rocky Gneisser",
    title = "No Gnats Are Taken for Granite",
    booktitle = "The Gnats and Gnus 1988 Proceedings"
    pages = "133-139"
}
@proceedings{gg-proceedings,
    editor = "Gerald Ford and Jimmy Carter",
    booktitle = "The Gnats and Gnus 1988 Proceedings"
}
```

A referência utiliza um campo *crossref*, que faz com que herde os elementos em falta da segunda referência, quando ausentes.

Recentemente, com o surgimento do XML, uma linguagem de marcação caracterizada por possuir uma estrutura, foram criadas diversas representação de conteúdo que baseadas em XML. Uma destas linguagens é o BibTeXML (Previtali et al., 2001), um esquema que possui também algumas ferramentas para processar a informação, uma representação de BibTeX em XML.

```
<book id="lamport:86">
 <authors>
  <name>
   <prename>Leslie</prename>
```

```
    <surname>Lamport</surname>
   </name>
  </authors>
  <title><tex code="{\LaTeX}">LaTeX</tex>:
  A Document Preparation System</title>
  <publisher>Addison-Wesley</publisher>
  <year>1986</year>
  <language>en-US</language>
  <index>LaTeX typesetting</index>
 </book>
```

A vantagem deste formato é que mantém as características do BibTeX, a estrutura e simplicidade, e possui também o poder de transformação e de representação proporcionado pelo XML.

No entanto, estas linguagens têm tido fraca aceitação, embora o uso de XML através de serviços Web tenha tido bastante sucesso, mas, predominantemente, com ontologias próprias para cada serviço ou protocolo. Protocolos como o OAI ou o SRW, apresentados mais à frente, possuem ontologias específicas de forma a fornecer mais informação para além da informação bibliográfica, tal como o conteúdo do documento ou outra informação relativa ao repositório, fontes de onde foi obtido, etc.

Esta sinergia entre serviços Web e recursos bibliográficos tem permitido aproveitar o potencial destas ferramentas e tem-se tornado cada vez mais numa constante. Neste capítulo abordaremos essas iniciativas e veremos como permitem facilitar a pesquisa entre parcerias académicas.

1.2.3.2 Refer/EndNote

Tanto o Refer como o EndNote são dois formatos semelhantes, mas programas distintos. O Refer é usado pelo *troff*, um sistema de formatação de texto comum na maioria dos sistemas Unix, enquanto que o EndNote é um programa comercial.

Os campos são identificados por um único caracter, antecedido pelo caracter %. Após o identificador, segue-se o elemento bibliográfico respectivo.

Exemplo de uma referência em formato Refer:

```
%A Jim Gettys
%A Phil Karlton
%A Scott McGregor
%T The X Window System, Version 11
%J Software Practice and Experience
%V 20
%N 20
%D 1990
%X A technical overview of the X11 functionality.
This is an update of the X10 TOG paper by Scheifler
\& Gettys
```

A identificação do tipo de publicação representado é feita com base nos elementos bibliográficos presentes. O exemplo anterior, é um artigo apresentado num jornal académico, pela presença do elemento %J.

Já o formato EndNote tem alguns identificadores adicionais, como por exemplo o %0 (digito zero) que permite especificar o tipo de publicação (*Artwork, Audiovisual Material, Book, Book Section, Computer Program, Conference Proceedings, Edited Book, Generic, Jounal Magazine, Magazine Article, Map, Newspaper Article, Patent, Personal Communication, Report* ou *Thesis*).

O exemplo anterior ficaria:

```
%0 Journal Article
%A Gettys, Jim
%A Karlton, Phil
...
```

Outra diferença entre os dois formatos é a representação dos autores. Como é visível nos exemplos, os autores são representados de formas distintas. Apesar de muito semelhantes, numa situação os nomes dos autores são armazenados na ordem natural, nome próprio no início, terminado com o apelido. No outra, a ordem altera-se, colocando-se primeiro o apelido e, separado por uma vírgula, o resto do nome, pela ordem normal.

1.2.3.3 RIS

O formato RIS (RIS, reference manual) é o formato usado pelo programa Reference Manager[9]. É um formato flexível, pensado para suportar a importação de outros formatos para o Reference Manager de forma a que sejam posteriormente mantidos no formato RIS.

Os campos são identificados por seis caracteres no inicio da linha: Duas letras maiúsculas, seguidas de dois espaços, um hífen e um espaço. "TY - ". A ordem dos campos não é relevante excepto o primeiro ,"TY - " que indica o tipo de publicação, do último, "ER -" que só indica o fim da referência e dos campos comuns, como por exemplo a ordem relativa de todos os identificadores "A1 - ", os autores. Assim, estes dois campos delimitam também as referências.

Exemplo de uma referência em formato RIS:

```
  TY  - JOUR
  A1  - Jim Gettys
  A1  - Phil Karlton
  A1  - Scott McGregor
  T1  - The X Window System, Version 11
  JO  - Software Practice and Experience
  VL  - 20
  IS  - 20
  Y1  - 1990
  N2  - A technical overview of the X11 functionality.
This is an update of the X10 TOG paper by Scheifler \&
Gettys
  ER  -
```

Este formato pode ser encontrado na Internet por exemplo nos catálogos da Springer, uma livraria online[10], permitindo a exportação das referências bibliográficas.

[9] http://www.refman.com/
[10] http://www.springerlink.com/

1.2.3.4 O formato da Linguateca

Apesar de este formato não ser usado para importar ou exportar publicações, ele está relacionado com os formatos anteriormente descritos, é um formato em texto, facilmente legível por programas, concebido por Paulo Rocha. Cada elemento ocupa uma linha, sendo identificado por uma sequência de três caracteres maiúsculos, seguidos de "=". A ordem dos elementos não é relevante e campos elementos que possam ser uma lista são duplicados, por exemplo, para cada autor existe uma entrada, onde, neste caso, interessa a ordem dos elementos da lista. Alguns dos campos são obrigatórios, tais como o tipo de publicação, o título, o autor ou editor, ano ou a língua.

```
-----------
TIP=revistas
ART=Sistema de Síntese de Fala a Partir de Texto -  DIXI
ANO=1996
AUT=M.C. Viana
AUT=L.C. Oliveira
AUT=I.M. Trancoso
AUT=P.M. Carvalho
LNG=pt
VOL=9
EDT=Conferência Nacional O Som e a Informação
REV=Revista Áudio: Dinamização Cultural
-----------
TIP=revistas
....
```

As referências são separadas por uma linha que contém apenas hífenes "-". Existe informação adicional que não é por enquanto usada na criação da referências bibliográficas mas ajuda à organização do catálogo, permitindo indexar os campos ou criar relações com colecções. Exemplos disso são campos como a língua (LNG) em que o documento foi redigido, informação sobre o documento ser uma reedição, informação sobre se o documento já foi ou não publicado, ou se pertence a algum projecto específico. O campo chave

(CHV) permite relacionar uma referência com uma colecção que fornece elementos adicionais, tal como acontece nas referências cruzadas no BibTeX. Esta funcionalidade é útil quando ocorram várias publicações no mesmo âmbito, ou seja na mesma revista ou conferência. O formato da Linguateca está descrito em pormenor em Linguateca (2005), onde se descreve também o processo de actualização de informação no catálogo.

1.3 Os vários problemas relacionados com referências bibliográficas

Com o aparecimento da Web e a divulgação de publicações em formato electrónico, a captação de artigos que sejam relevantes para o utilizador pode ser extremamente facilitada. Recorrendo a motores de pesquisa genéricos e repositórios específicos de um domínio e usando palavras-chave (*título*, *autor* ou outros termos específicos do domínio em estudo) poderemos encontrar documentos com relativa facilidade. Mas encontrar os documentos com a informação não é suficiente. É necessário extrair e identificar a informação bibliográfica, as referências e os respectivos elementos bibliográficos. E após esta fase, é necessário validar essa informação. Para poder citar o documento é necessário obter a referência o mais correcta e completa possível.

1.3.1 Soluções usadas

Este problema não é recente. Os primeiros repositórios bibliográficos online datam do início da década de 90, tendo sido adapatados de bases de dados de instituições académicas. Repositórios como o DBLP[11] ou o CiteSeer[12] são bastante utilizados no domínio da ciência de computadores. Os métodos de recolha de informação destes dois repositórios têm contudo algumas diferenças.

O DBLP (DataBase for Language Programming) surgiu na década de 80 na Universität Trier na Alemanha. A informação é inserida através da introdução completa de actas de conferências e revistas seleccionadas. A informação é

[11] `http://www.informatik.uni-trier.de/~ley/db/`
[12] `http://citeseer.ist.psu.edu/`

organizada por autores, co-autores, revistas e conferências.

O CiteSeer, desenvolvido pelo NEC Research Institute, é também um repositório bibliográfico, mas o método de recolha de informação bibliográfica é feito através da pesquisa e recolha automática de informação em documentos académicos na Web. O CiteSeer usa ainda um índice de citações, que permite procurar documentos por citações ou ordenar listas de publicações pelo impacto de citações.

Mas dada a carga no processo de actualização em ambos estes repositórios, as actualizações são periódicas e nem sempre indexam todas as conferências ou revistas do domínio que cobrem. Por vezes não encontramos imediatamente a referência que procuramos usando estes sistemas, sobretudo quando se trata de publicações em português. Nestas situações somos levados a pesquisar:

- Repositórios que indexam artigos apresentados num conjunto limitado de conferências ou jornais de um domínio específico;
- Sítios Web de instituições de investigação;
- Páginas pessoais;
- Páginas de conferências;
- Revistas online.

O esforço despendido na obtenção manual dos dados bibliográficos é considerável, mesmo recorrendo aos motores de pesquisa. Frequentemente o utilizador depara-se com dados incompletos (ausência de ano, ou do número da página), incongruências (múltiplas versões da mesma referência com anos diferentes), informação desactualizada ("*to be published*"), incorrectas (como verificar que um artigo foi de facto publicado?), ou com o facto de que não é possível descobrir a que objecto uma referência bibliográfica se refere. As referências bibliográficas podem ser representadas através de vários estilos distintos, alterando a ordem e/ou omitindo alguns dos elementos bibliográficos, conforme descrito na secção 1.2.2. Procurando na Web, uma publicação familiar pode ser encontrada representada de formas distintas:

"Corpógrafo V3: From Terminological Aid to Semi-automatic Knowledge Engine"
Luís Sarmento, Belinda Maia, Diana Santos, Ana Pinto & Luís Cabral
In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)
Génova, Itália, 22-28 de Maio de 2006

L. Sarmento et al. "Corpógrafo v3: from terminological aid to semi-automatic knowledge engine". LREC 2006 - Genoa, Italy, 2006

Neste tipo de situações em que são encontradas duas referências semelhantes surgem as seguintes questões:

- As duas referências referem-se à mesma publicação?
- Até que ponto estão completas? Em ambas falta o número das páginas e na segunda falta a data.
- A informação está actualizada?
- Onde decorreu a conferência a que se refere esta referência? (Génova, Genoa)
- Quem são os autores? O autor *Luís Sarmento* e *L. Sarmento* são a mesma pessoa? Qual a melhor forma para desambiguar nomes?

As questões podem ser simples de responder no exemplo em questão, mas num processo automatizado, com várias referências que possuem vários elementos comuns, o processo de validação está longe de ser um processo trivial.

1.3.2 Um caso prático

Nesta dissertação será considerado um caso prático: O catálogo de publicações da Linguateca[13], um repositório bibliográfico no domínio do processamento computacional da língua portuguesa. O catálogo de publicações da Linguateca tem contado com a contribuição de vários autores e dos vários membros da equipa da Linguateca para o manter. No capítulo 2 o catálogo de publicações

[13] `http://www.linguateca.pt`, ver catálogo de publicações

da Linguateca será descrito em pormenor, mas as principais dificuldades com que nos deparámos na manutenção deste catálogo são:

- A inserção de referências bibliográficas, feita manualmente, leva a que por vezes, estas se encontrem incompletas. A inserção manual de vários artigos de uma conferência ou revista é um processo penoso, e que pode levara que se insira informação repetida tantas vezes quantos os artigos. Apesar de este problema ser aliviado pelo uso de referências cruzadas, não existem meios para procurar e relacionar essa informação.
- Não existe um processo de validação (semi-)automático.
- Não há métodos para a actualização periódica das referências bibliográficas já armazenadas.

Esta situação leva a um crescimento lento deste recurso, assim como a um excesso de trabalho humano para a sua manutenção.

1.4 Objectivos

Foram apresentados alguns dos problemas para encontrar referência bibliográficas. Foi também referido o catálogo de publicações da Linguateca, um repositório que não possuía um sistema automático para obter referências bibliográficas.

Propõe-se como objectivo deste trabalho o desenho de uma plataforma modular que permita a obtenção, o tratamento, a validação e a actualização de informação bibliográfica, ou seja, referências e elementos bibliográficos, de forma a que a informação resultante seja facilmente aplicável às necessidades de um repositório bibliográfico. O desejo desta plataforma é o de minimizar o esforço de manutenção e, simultaneamente, maximizar o processo de descoberta de documentos de um domínio, facilitar a sua inserção num repositório e assim melhorar significativamente a qualidade do recurso. O sistema proposto, baptizado de **SUPeRB**, **S**istema **U**niformizado de **Pe**squisa de **R**eferências **B**ibliográficas, é uma arquitectura leve, implementado de forma modular, capaz de levar a cabo várias tarefas distintas:

- Pesquisa a motores de busca genéricos através de serviços Web, procurando encontrar documentos ou referências que completem e confirmem uma referência bibliográfica ou que correspondam a uma expressão, combinação de elementos bibliográficos;
- Análise de referências bibliográficas e extracção dos elementos bibliográficos respectivos (título, autor, ano da publicação, etc.);
- Análise periódica de páginas de colaboradores ou investigadores no domínio;
- Recolha de elementos bibliográficos (necessários para completar uma referência bibliográfica) a partir dos textos obtidos na Web;
- Validar os candidatos a referências obtidos, para garantir que os dados obtidos se refere à referência em causa/construída;
- Manutenção dos dados arquivados, nomeadamente a actualização periódica dos dados e a alteração do estado.

Os pontos propostos são processos automáticos onde a validação humana mantém-se como uma parte importante no processo de inserção de publicações. No entanto é esperado que a automatização destes pontos resulte numa redução considerável da carga do gestor humano.

Pretende-se que o sistema a desenvolver seja dedicado ao processamento computacional da língua portuguesa. No entanto a arquitectura apresentada, assim como o sistema construído, deverá ser possível de aplicar a outros domínios.

1.5 Resumo

Para o leitor poder ter uma visão da estrutura desta dissertação e dos tópicos abordados, é apresentada uma breve descrição de cada um dos capítulos que compõem a dissertação:

Capítulo 1 Este capítulo. Contém uma introdução ao tema e conceitos sobre a informação bibliográfica

Capítulo 2 Neste capítulo é apresentado em pormenor o catálogo de publicações da Linguateca, apresentando estatísticas da informação bibliográfica armazenada e os métodos de inserção, validação e actualização dos dados no catálogo. São ainda apresentadas algumas sugestões de reestruturação do catálogo que usem as funcionalidades oferecidas pelo SUPeRB.

Capítulo 3 O capítulo 3 apresenta algumas das tecnologias e métodos usados na obtenção, gestão e disseminação de referências. Este capítulo apresenta tecnologias usadas no campo da pesquisa e disseminação de referências bibliográficas, bem como outras que possam ser aplicadas ao mesmo propósito. Apresentam-se ainda algumas técnicas de extracção de informação. Outra área discutida neste capítulo é a dos sistemas já existentes que têm como função a organização de referências bibliográficas para o utilizador individual.

Capítulo 4 O capítulo 4 descreve em pormenor a arquitectura proposta e implementada no SUPeRB, decompondo em vários módulos independentes as partes mais relevantes do problema.

- pesquisa na Web, recorrendo a motores de pesquisa e repositórios bibliográficos;
- extracção de informação a partir de conteúdos Web;
- extracção de elementos bibliográficos de referências bibliográficas;
- avaliação da relevância dos elementos bibliográficos obtidos;
- arquivo e reutilização da informação bibliográfica obtida.

Cada uma destas secções descreve em pormenor as tecnologias, os algoritmos e os recursos usados para construir este sistema, assim como a interface do utilizador.

Capítulo 5 Após a apresentação da arquitectura geral do SUPeRB, no capítulo 5 é apresentada uma forma de avaliar o desempenho do sistema.

Capítulo 6 No capítulo 6 são apresentadas as conclusões extraídas desta dissertação e avaliam-se possíveis caminhos para dar continuidade ao trabalho.

Capítulo 2

O catálogo de publicações da Linguateca

Neste capítulo é apresentado o catálogo de publicações da Linguateca. Pretende-se descrever o trabalho levado a cabo para produzir o recurso que é hoje o catálogo, um recurso importante mas que necessita de ser melhorado de forma a poder servir melhor a comunidade.

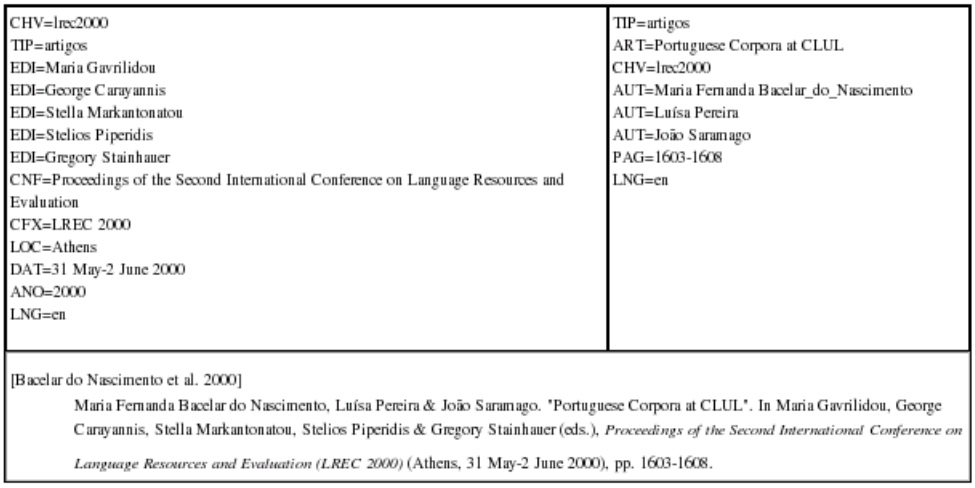
2.1 A história e função do catálogo

A Linguateca (Santos, 2000, 2002; Santos et al., 2004), um centro de recursos para o processamento computacional da língua portuguesa, disponibiliza um serviço onde é possível pesquisar e adicionar referências bibliográficas relacionadas com o domínio em questão, o processamento computacional da língua portuguesa: O catálogo de publicações da Linguateca. Este catálogo tem vindo a ser construído desde o início da Linguateca, mais precisamente desde 1999, altura em que a Linguateca ainda tinha o nome de projecto Processamento Computacional do Português.

À medida que o catálogo tem vindo a agrupar cada vez mais referências bibliográficas, todo o processo de manutenção, no que diz respeito a inserir, validar e actualizar a informação bibliográfica, tem-se tornado cada vez mais complicado. Isto deve-se ao facto deste processo ser manual e não disponibilizar

nenhuma ajuda ao utilizador ou à pessoa responsável pela gestão do catálogo, o gestor.

A informação bibliográfica está armazenada em dois ficheiros de texto, estruturados num formato próprio para as necessidades da Linguateca. O primeiro ficheiro contém informação bibliográfica para cada uma das publicações. O segundo ficheiro contém informação adicional sobre colecções a que algumas das publicações no primeiro ficheiro pertencem, nomeadamente conferências, livros ou revistas em que vários artigos no catálogo tenham sido simultaneamente publicados.



CHV=lrec2000
TIP=artigos
EDI=Maria Gavrilidou
EDI=George Carayannis
EDI=Stella Markantonatou
EDI=Stelios Piperidis
EDI=Gregory Stainhauer
CNF=Proceedings of the Second International Conference on Language Resources and Evaluation
CFX=LREC 2000
LOC=Athens
DAT=31 May-2 June 2000
ANO=2000
LNG=en

TIP=artigos
ART=Portuguese Corpora at CLUL
CHV=lrec2000
AUT=Maria Fernanda Bacelar_do_Nascimento
AUT=Luísa Pereira
AUT=João Saramago
PAG=1603-1608
LNG=en

[Bacelar do Nascimento et al. 2000]
Maria Fernanda Bacelar do Nascimento, Luísa Pereira & João Saramago. "Portuguese Corpora at CLUL". In Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis & Gregory Stainhauer (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)* (Athens, 31 May-2 June 2000), pp. 1603-1608.

Figura 2.1: Extracto dos ficheiros no formato da Linguateca. A informação das colecções encontra-se à esquerda, a dos artigos à direita. Em baixo a referência produzida pela combinação dos dois campos.

A estrutura deste formato não tem sofrido alterações significativas desde a sua criação, com excepção da adição de novos campos, para satisfazer necessidades que foram surgindo. Para processar esta informação, foram criados programas em Perl capazes de gerar páginas HTML que constituem a interface do catálogo.

A forma como o catálogo está disponível ao utilizador é uma série de páginas HTML, divididas em categorias de publicações e ordenadas quer por autor, quer por data. Categorias em que o catálogo de publicações está dividido são:

- Livro
- Capítulo de livro

- Artigo publicado numa conferência internacional
- Artigo publicado noutra conferência
- Artigo publicado numa revista
- Relatório
- Tese
- Apresentação
- Documento publicado apenas na Web

O conjunto de programas criador do catálogo tem ainda a seu cargo a geração de páginas de publicações relacionadas com projectos específicos, como, por exemplo, todas as publicações produzidas no âmbito da Linguateca. O catálogo dispõe também de uma interface de pesquisa (figura 2.2) que permite consultar o repositório em vez de navegar pelas páginas HTML.

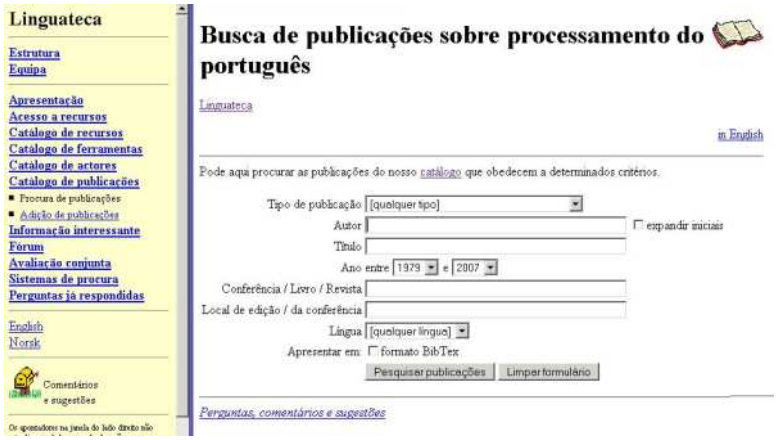


Figura 2.2: Formulário de pesquisa no catálogo de publicações da Linguateca

O catálogo de publicações foi construído com o objectivo de ser uma ferramenta cooperativa, em que os investigadores interessados podiam submeter referências bibliográficas que considerassem relevantes para o processamento computacional da língua portuguesa, quer da sua autoria, quer de outras fontes. A introdução dos elementos bibliográficos é feita através de um formulário HTML, visível na Figura 2.3, e, após a sua validação pelo gestor do catálogo, a publicação é introduzida no catálogo.

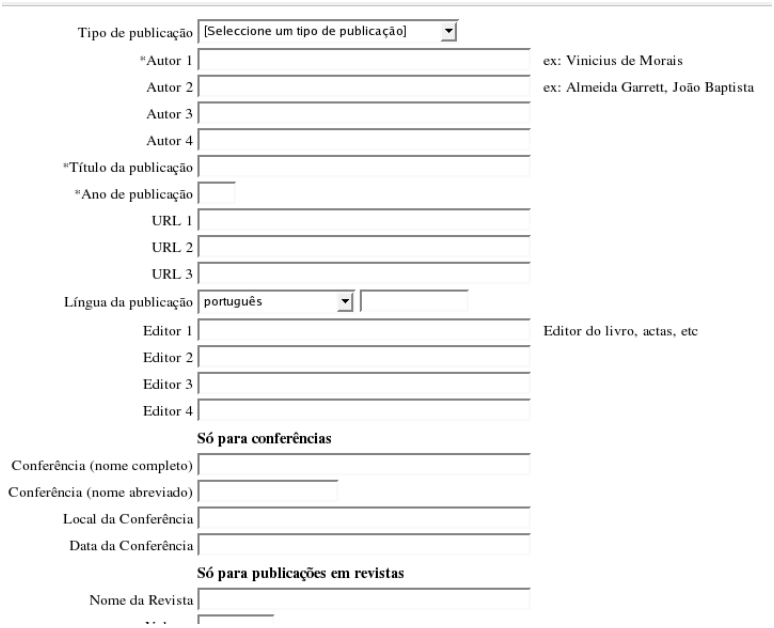


Figura 2.3: Parte do formulário de adição e publicações do catálogo de publicações da Linguateca

Todo este trabalho foi desenvolvido pelo Paulo Rocha, membro da equipa da Linguateca.

2.2 As publicações do catálogo

Em Janeiro de 2006, o catálogo de publicações continha 1050 referências bibliográficas sobre o processamento computacional da língua portuguesa. A tabela 2.1 apresenta os valores observados nas referências bibliográficas do catálogo em Janeiro de 2006 no que diz respeito às hiperligações disponíveis:

Tabela 2.1: Publicações e URL no catálogo em Janeiro de 2006

Total de publicações	1050
Publicações da Linguateca	128
Publicações com URL	550
Total de URL	710
URL activos	547
Documentos no repositório da Linguateca (em cache)	36
URL de Publicações da Linguateca	254

Cada uma destas referências pode conter uma ou mais hiperligações para versões do documento, resumos ou apresentações on-line. Ao analisar estes dados, concluiu-se que, dos 710 URL existentes, apenas 540 (77%) estavam activos. Destes, apenas 259 eram URL externos, os restantes eram URL internos de publicações armazenadas no repositório da Linguateca. Incluindo as publicações da Linguateca, existiam 438 hiperligações distintas que se mantinham activas.

Mais recentemente, em Agosto de 2006, o catálogo possuía já 1220 referências bibliográficas, com 790 hiperligações. Este aumento representa um crescimento razoável, evidenciando o ritmo de produção de trabalhos apresentados no domínio em questão, mas é necessário considerar também todo o processo seguido, sem qualquer meio automático, para a descoberta de parte destas publicações e a sua inserção.

2.3 Problemas do catálogo

O catálogo foi criado há bastante tempo e inicialmente a dimensão dos dados era bastante menor. O catálogo foi desenvolvido a pensar em agrupar e apresentar publicações no âmbito do processamento computacional da língua portuguesa. Dado que este recurso foi desenvolvido de raiz, só após o seu crescimento e a necessidade de usar o seu conteúdo noutros contextos, organizar e apresentar os conteúdos, é que foi possível conhecer as suas limitações e problemas.

Hoje, com a experiência em manter este recurso ao longo de sete anos, têm sido detectados vários problemas, principalmente no que diz respeito à usabilidade, mas também alguns problemas conceptuais.

2.3.1 Problemas de manutenção

A manutenção do catálogo é talvez o mais complicado e aquilo que mais motivou o projecto descrito nesta tese. Esta manutenção obriga a inúmeras tarefas, desde:

1) a validação de referências bibliográficas inseridas;

2) a confirmação de que as referências inseridas não existem no catálogo;

3) a actualização periódica da informação.

Todas estas tarefas são feitas manualmente e sem qualquer ajuda automática.

2.3.1.1 Validação da informação

A informação inserida requer frequentemente a verificação manual em repositórios, nas páginas das conferências, editoras ou dos próprios autores. Só assim se pode confirmar que os elementos bibliográficos dados estão correctos.

2.3.1.2 Verificar a existência no catálogo

Apesar de simples, este processo poderia ser facilitado por métodos de normalização e comparação automática. Este tipo de inserções ocorre com alguma frequência, quer por erro, quer para actualizar a referência já existente pelo autor. Apesar de este problema poder estar relacionado com a usabilidade do catálogo, isto levanta outro problema, que é a fusão de informação bibliográfica em geral.

2.3.1.3 Actualização de informação

Está relacionada, por um lado, com a possibilidade de edição (usabilidade). No entanto, não existe também nenhum método que automatize a pesquisa de informação. A possibilidade de periodicamente validar a informação do catálogo com outras fontes na Web oferece não só a possibilidade de dados actualizados mas também a possibilidade de encontrar novas publicações, por exemplo.

2.3.2 Problemas de usabilidade

Os problemas de usabilidade estão principalmente relacionados com a inserção e actualização de informação por um utilizador externo.

2.3.2.1 A inserção

A introdução de referências bibliográficas no catálogo, como se vê na figura 2.3, é feita através de um formulário complexo, composto por inúmeros campos, cada um correspondendo a um elemento bibliográfico em particular que, dependendo do tipo de publicação, pode ou não ter que ser preenchido. É necessário um conhecimento mínimo da estrutura de uma referência bibliográfica para preencher correctamente os campos do formulário. Facilmente podem ser inseridos erros das seguintes formas:

- Inserção de elementos bibliográficos no campo do formulário incorrecto.
- Omissão de elementos bibliográficos (o utilizador esquece-se ou desconhece elementos como o número de páginas)
- Introdução de elementos bibliográficos incorrectos (erros ortográficos, que por vezes são dificilmente detectáveis).

Outro problema é que o formulário é a única interface para os utilizadores introduzirem referências bibliográficas. Não é possível, por exemplo, introduzir uma referência bibliográfica nas suas formas de representação mais vulgares como é encontrada num documento (o texto todo junto ou então num formato como o BibTeX). Não existe nenhum meio automático para processar e introduzir este tipo de dados no catálogo.

2.3.2.2 A edição

Não existe nenhum meio que possibilite a edição de uma referência bibliográfica pelo utilizador. Se um utilizador detectar algum problema numa referência bibliográfica, tem que tomar a iniciativa de enviar uma mensagem de correio electrónico ao gestor do catálogo, ou de introduzir a referência novamente (tendo que introduzir a referência completa). Mesmo a edição por parte do gestor envolve a edição directa dos elementos bibliográficos.

2.3.3 Problemas conceptuais

Entende-se por problemas conceptuais lacunas na organização de dados que levam à limitação do catálogo.

2.3.3.1 Identificador da referência

Na estrutura em que as referências bibliográficas são armazenadas, não existem identificadores únicos que ajudem a processar a informação bibliográfica. Assim, a criação de hiperligações entre referências bibliográficas é dificultada. É difícil citar referências bibliográficas, criando uma hiperligação para uma outra versão (republicações) no catálogo, por exemplo. A ausência de um identificador único, da data da inserção ou da última alteração, bem como o rasto de quais as alterações sofridas, ou ainda um mínimo de informação sobre o utilizador que introduziu uma dada referência no catálogo têm sido notados como importantes numa análise posterior do catálogo. Estes problemas não estão apenas ligados à falta de especificação destes campos, implicam também a falta de métodos para gerar e processar esta informação.

2.3.3.2 Colecções bibliográficas

A geração das páginas do catálogo baseia-se no tipo de publicação que cada referência representa. São também geradas páginas para algumas colecções relacionadas com projectos. Mas este método conta com a criação de um novo campo na base de dados para identificar a referência pertencente a esse grupo. É, portanto, necessário reescrever o código para processar cada nova colecção e assim expandir o catálogo de forma a gerar uma página para esta. Idealmente deveria existir um processo mais simples de produzir facilmente colecções de documentos.

2.3.3.3 Entidades nas referências

Um outro problema tem a ver com os autores e as possíveis representações dos seus nomes. Diversos estilos bibliográficos abreviam os primeiros nomes do autor, podendo criar ambiguidade na identificação. Diversos repositórios

bibliográficos apresentam o mesmo problema, mesmo na representação num formato estruturado. Este problema deve-se ao facto de esses mesmos repositório não possuírem meios para proceder à desambiguação ou optarem por não desempenhar esta tarefa, evitando assim erros.

2.3.3.4 A exportação

A exportação de referência bibliográficas é possível de duas formas: Texto simples ou no formato BibTeX. Apesar do formato BibTeX ser um dos mais usados, é possível que exista a necessidade de exportar referências bibliográficas noutro formato. Nesta situação, mais facilmente pode surgir a necessidade de obter as referências noutros estilos.

2.3.3.5 Esquema de classificação

Outro pormenor consiste na classificação do catálogo apresentada em 2.3. Não só esta classificação é fixa, ligada a necessidades internas, como também não é equivalente a outros modelos comuns, necessitando ser mapeado se se quiser, por exemplo, usar a classificação empregue pelo CiteSeer ou pedida pela FCT em relatórios de projectos.

2.4 Extensões lógicas ao catálogo

Como foi apresentado, deparamo-nos com um repositório em pleno crescimento, para servir as necessidades dos utilizadores, que é urgente dotar de uma maior usabilidade, para os utilizadores em geral e para o tornar mais fácil de gerir. Os problemas apresentados na secção 2.3 limitam não só o crescimento do recurso mas também o seu potencial.

Queremos assim facilitar todo o processo de manutenção das referências bibliográficas. Apesar de o catálogo ser, provavelmente, o único portal para a comunidade científica dedicado ao processamento computacional da língua portuguesa, é definitivamente possível melhorá-lo. De facto, existe um universo de publicações que só será alcançado com recurso a meios automáticos que

facilitem a sua descoberta, processamento, inserção e gestão no catálogo.

No que diz respeito à inserção de dados bibliográficos, a reformulação da interface para permitir a edição de referências bibliográficas já existentes no catálogo é uma necessidade. A realidade é que os utilizadores têm mais facilidade em encontrar informação bibliográfica sob a forma de texto, seguindo formatos bibliográficos ou através da análise dos próprios documentos.

Mas estes pormenores não são os mais importantes. A tarefa que consome mais recursos humanos é a validação e a manutenção das referências bibliográficas, este processo que depende exclusivamente do gestor do catálogo e das suas capacidades.

É neste âmbito que o sistema apresentado neste capítulo surge, para aumentar a produtividade no processo de localização e gestão de referências bibliográficas candidatas no domínio em questão. O processo aqui proposto baseia-se na análise das referências bibliográficas à medida que estas são inseridas, propondo sugestões que validem ou apontem inconsistências na informação introduzida.

É também necessário monitorizar as referências bibliográficas já existentes no catálogo de forma a manter a informação o mais actualizada possível. Isto diz respeito a publicações que tenham sido introduzidas ainda incompletas mas também a informação volátil, como é o caso das hiperligações, atributos que podem facilmente sofrer alterações. É necessário prever situações em que se pretenda monitorizar páginas relacionadas com o domínio, tal como páginas de autores que produzam publicações nestes domínios ou conferências periódicas. Potencialmente podem ser encontradas muitas novas publicações relevantes.

O sistema proposto tem o principal propósito de assistir o gestor do catálogo e não de o substituir. A ideologia subjacente é a de apontar potenciais soluções e deixar que seja o gestor a decidir. Assim, é da responsabilidade do sistema encontrar e filtrar essas soluções e apresentá-las ao gestor, facilitando a sua introdução no catálogo. Outra funcionalidade é a de gerar novos recursos internos que permitam aumentar o potencial do catálogo, nomeadamente facilitar meios para permitir a desambiguação de entidades como nomes de autores, editores, editoras, conferências ou locais. É necessário organizar os recursos existentes de forma a poder utilizá-los conjuntamente com métodos que permitam resolver estas situações mas que também possam ser aplicados

a outros problemas.

É ainda necessário analisar quais as tarefas em que novas soluções possam aumentar a produtividade.

No capítulo 4 é apresentada a arquitectura para o sistema proposto, bem como especificações para cada tarefa envolvida, as ferramentas e tecnologias abordadas. É preciso salientar que o sistema aqui apresentado não tem apenas o objectivo de melhorar as funcionalidade do catálogo de publicações da Linguateca, mas que a fácil aplicação destes métodos a outros repositórios e domínios também foi tida em consideração. Os diversos módulos podem ser usados individualmente em tarefas particulares que não impliquem necessariamente a ligação de todos os passos do sistema. Isto é, pode ser possível analisar uma referência bibliográfica sem ter que a pesquisar e extrair de um texto ou sem ter que introduzi-la necessariamente num repositório.

Capítulo 3

Tecnologias e estudos relevantes

Neste capítulo são apresentados trabalhos, tecnologias e estudos que abordam a descoberta e tratamento de informação bibliográfica, bem como outros estudos relevantes para o objectivo em questão. Assim, este capítulo divide-se em duas áreas distintas:

- Numa primeira parte serão abordadas tecnologias relacionadas com a pesquisa de informação relevante na Web, descrevendo protocolos para pesquisa e obtenção de informação da Web não só no domínio bibliográfico mas também através de acesso a motores de pesquisa genéricos.

- Numa segunda parte, é abordado o processamento e extracção de informação de documentos e formas de validar essa informação.

3.1 Pesquisa na Web

Nesta secção serão abordados vários meios para aceder a informação estruturada na Web. Nomeadamente discutem-se protocolos de acesso a repositórios bibliográficos, que retornam a informação de forma estruturada, mas são também apresentados os serviços Web dos motores de pesquisa genéricos, possibilitando a pesquisa "global" da Web.

3.1.1 Acesso a repositórios bibliográficos através de serviços Web

A cooperação entre entidades responsáveis por manter e partilhar referências bibliográficas tem sido uma constante ao longo dos anos. A possibilidade de pesquisar repositórios remotos tem evoluído para acompanhar os avanços tecnológicos e as necessidades dos utilizadores. Este tipo de acesso evoluiu com o surgimento dos serviços Web. Os serviços Web são usados para proceder a pesquisas estruturadas e troca de dados entre alguns dos repositórios e motores de busca de publicações.

3.1.1.1 Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)

A Open Archives Initiative (OAI)[1] é uma organização que visa a distribuição de conteúdos. A OAI desenvolveu o OAI-PMH[2] com o intuito de proporcionar um enquadramento para a distribuição e recolha de meta-informação de repositórios. Este protocolo fornece um enquadramento com duas classes de participantes:

- Os **fornecedores de dados** com a função de administrar sistemas que suportam este protocolo como meio de divulgar meta-informação acerca do conteúdo dos seus sistemas.
- Os **fornecedores de serviços** responsáveis por emitir pedidos através do protocolo OAI e usar a meta-informação devolvida na construção de serviços refinados.

Este protocolo visa essencialmente a partilha de meta-informação entre repositórios através de XML mas proporcionando uma API para o acesso à informação obtida.

No entanto, este protocolo não se destina à pesquisa de candidatos a referências bibliográficas mas antes à disseminação e troca, em grandes quantidades, de

[1] http://www.openarchives.org/
[2] http://www.openarchives.org/OAI/openarchivesprotocol.html

referências bibliográficas. Os métodos disponíveis são: *GetRecord*, *Identify*, *ListIdentifiers*, *ListMetadataFormats*, *ListRecords* e *ListSets*.

Por exemplo, este protocolo foi usado pelo Microsoft Live Academic[3], um motor de busca na Internet, para recolher informação de vários repositórios bibliográficos.

3.1.1.2 O Z39.50

O Z39.50 é um protocolo cliente/servidor, através de TCP/IP, dedicado à pesquisa de informação bibliográfica em repositórios ou computadores remotos. Este protocolo, bastante antigo, antecede a Web, tendo surgido em 1970. A última versão deste protocolo data de 2003 (Z39.50-2003). É um dos protocolos mais utilizados entre repositórios de bibliotecas académicas e foi desenvolvido para resolver problemas relacionados com a pesquisa em vários repositórios e as especificidades de cada um, nomeadamente campos ou menus exclusivos de certos repositórios.

Têm sido feitas várias tentativas de adaptar este protocolo às novas tecnologias baseadas em XML e serviços Web. Algumas mal sucedidas, como a tentativa com o nome de *ZING*(*Z39:50: international; Next Generation*). De destacar duas das mais importantes: os protocolos gémeos SRU/SRW, que são versões HTTP do protocolo Z39.50 usando XML, que descrevemos a seguir.

3.1.1.3 O SRU e o SRW

SRU (*Search and Retrieve via URL*)/SRW (*Search and Retrieve via Web Service*) são dois protocolos gémeos que funcionam através de parâmetros no URI ou através de SOAP (Soap Version 1.2), um protocolo de troca de mensagens em XML, respectivamente. O resultado de ambos os protocolos é codificado em XML. Estes protocolos são um exemplo do que a Web 2.0 pode oferecer, permitindo disseminar informação através de HTTP e XML e facilitando a sua utilização por programas ou mesmo apresentar directamente os resultados em HTML recorrendo a folhas de estilo (*stylesheets*).

[3] `http://academic.live.com`

A diferença entre estes dois protocolos é apenas na forma como o pedido é feito. No entanto, o SRU tem vindo a superar o seu parente SRW. Este facto pode dever-se à simplicidade de formalizar o pedido através de um URL em comparação com o método do SRW, em que é necessário gerar um objecto XML.

```
http://z3950.loc.gov:7090/voyager?version=1.1
            &operation=searchRetrieve&query=dinosaur
```

Como já foi dito, estes protocolos são uma variante do protocolo Z39.50, mantendo as suas principais propriedades, mas funcionando através de pedidos HTTP. O principal propósito deste protocolo é o de permitir a pesquisa a repositórios remotos através de serviços Web. Os pedidos são feitos através de um pedido **searchRetrieveRequest**, feito por URL ou num objecto XML, dependendo do protocolo, e que o servidor processa e retorna um objecto XML **searchRetrieveResponse**, que contém uma lista de registos coincidentes com os parâmetros da pesquisa dada.

A expressão de pesquisa é feita em CQL (*Common Query Language*), uma linguagem formal, para representação de pesquisas a sistemas de informação, como repositórios bibliográficos.

Tabela 3.1: Exemplos de consultas em CQL

title all "Síntese fala"	Título contém todas as palavras entre aspas
title any "Syntactical annotation"	Título contém qualquer uma das palavras entre aspas
title exact "The Multilingual Question Answering Track at CLEF"	Título exacto
date within "2002 2006"	Datas entre 2002 e 2006
any/relevant "Syntactical annotation"	Aplica um algoritmo de relevância para determinar resultados e a ordem respectiva

Esta linguagem, apesar de intuitiva e simples de utilizar, é bastante poderosa, disponibilizando ainda funções para tratar texto e expressões regulares desde

métodos para encontrar a raiz da palavra (*stemming*), usar termos relevantes ou mesmo procurar palavras foneticamente semelhantes.

3.1.1.4 A API do CiteSeer

O repositório CiteSeer (Bollacker et al., 1998) possui uma API SOAP/WSDL, descrita em Petinot et al. (2004), dedicada a repositórios CiteSeer. Esta API disponibiliza todas as funcionalidades fornecidas pelo CiteSeer, inclusive pesquisa ao texto completo. Estas funcionalidades envolvem três tipos de recurso: documentos, citações e grupos. Cada item destes recursos, ou seja, cada documento, cada citação e cada grupo possui um URI.

```
http://<server>/document/<encoding>/<document-id>
http://<server>/citation/<encoding>/<citation-id>
http://<server>/group/<encoding>/<group-id>
```

Cada um destes URI é um identificador de uma instância de um documento em formato XML.

A API do CiteSeer permite os métodos de pesquisa **findDocumentByText** e **findCitationByText**. Estes métodos retornam uma lista de URI em vez das instâncias dos documentos. Para cada recurso, existe um método, **getDocument**, **getCitation** e **getGroup**. Além destes métodos, existem ainda outros métodos capazes de retornar o texto de uma publicação, ou descobrir quais as publicações recentemente inseridas no CiteSeer.

3.1.2 Acesso a motores de pesquisa genéricos através de serviços Web

A Web é usada na área de extracção de informação, através da recolha e análise de documentos obtidos a partir de consultas na Web. Cada vez mais a Web é usada como um recurso, uma base de dados para inúmeros fins, tais como aprendizagem para extracção de informação sobre venda de produtos (Cordeiro, 2003), resposta automática a perguntas (Costa, 2005) ou por forma a construir corpora para determinados fins (Baroni e Bernardini, 2006). A pesquisa de consultas de referências bibliográficas é também uma forma de aplicar extracção

de informação na Web como o Google Scholar[4] e o CiteSeer). Estes meios de pesquisa recorrem por norma ao uso de motores de pesquisa comuns ou a programas próprios que pesquisam páginas Web.

Motores de pesquisa, como o **Google** e o **Yahoo!** são os índices da Web. Pode-se dizer que, se vários motores de busca não encontram o objecto que se procura, então o mais provável é que essa informação não esteja disponível na Web. O acesso a estes motores de pesquisa pode ser através de programas próprios que usam as interface Web dos motores de pesquisa e extraem os resultados directamente do código HTML. Isto tem, no entanto, vários contras:

- Este procedimento pode ser entendido como um comportamento abusivo, levado a cabo por serviços automáticos;
- As interfaces estão sempre sujeita a alterações;
- O surgimento de novas tecnologias para permitir páginas dinâmicas, como o Ajax, dificulta o processamento dos resultados.

Por isso, com a disponibilização de novas tecnologias como os serviços Web, torna-se mais prático fazer os pedidos a serviços Web, usando API próprias.

Infelizmente nem todos os motores de pesquisa possuem serviços Web. Apenas alguns disponibilizam esta tecnologia e disponibilizam API[5]. Ao usarmos serviços Web, estamos a comunicar através de uma API, através de uma linguagem de alto nível, sendo mais fácil de processar a processar a informação resultante do pedido, evitando erros de análise sintáctica.

Uma análise a três das API disponíveis, nomeadamente a do Google, Yahoo e MSN, em Janeiro de 2006, permitiu uma avaliação das capacidades disponibilizadas por cada um dos motores conforme é possível visualizar na tabela 3.2.

Das três API, a que aparentemente disponibiliza mais opções é a do Yahoo, permitindo pesquisas em espaços distintos, o que permite a pesquisa exclusiva

[4] `http://scholar.google.com`

[5] É possível ir buscar as API em diversas linguagens de programação a partir dos sítios `http://www.google.com/apis/` e `http://developer.yahoo.com/search/`, respectivamente, onde é possível obter mais informação e exemplos.

[6] O Yahoo é o único cuja pesquisa em feito num espaço específico. A pesquisa é feita num dos seguintes espaços, permitindo a pesquisa, de forma distinta, a Documentos(Web), imagem, vídeo, Notícias, Yahoo, termos, spellcheck ou Relacionado

Tabela 3.2: Diferenças entre as API dos três principais motores de busca

	Google	Yahoo	MSN
Limite diário	1000	5000	10000
Limite de respostas	20	50	50
Protocolo	SOAP	REST	SOAP
Pesquisa em espaços específicos[6]	Não	Sim	Não
Devolve ultima actualização	Não	Sim	Não
Exemplo Java	Sim	Sim	Não
Exemplo Javascript	Sim	Sim	Não
Exemplo Perl	Sim	Sim	Não
Exemplo .NET	Sim	Não	Sim
Exemplo PHP	Sim	Sim	Não
Exemplo Flash	Não	Sim	Não

a documentos da Web genéricos (pesquisa toda a Web, excluindo documentos multimédia), imagens, vídeos, notícias, documentos do Yahoo, listas de termos, sugestões de escrita ou sugestões de pesquisas relacionadas. O MSN é aquele que oferece um limite diário maior, mas que apresenta menos exemplos de uso do serviço Web, disponibilizando exemplos apenas em .NET.

Estas API permitem acesso a três dos maiores motores de pesquisa, ou seja, acesso aos maiores índices da Web a partir de uma aplicação, oferecendo ainda opções para refinar a pesquisa.

3.2 Extracção de informação

A extracção de informação a partir de textos da Web não é nova. Existem já inúmeros sistemas que têm como objectivo obter informação de documentos acessíveis na Web. Nesta secção são descritas diversas metodologias bem como alguns sistemas.

3.2.1 Extracção de informação de texto

Sistemas como o Armadillo (Ciravegna et al., 2004) ou o KnowItAll (Etzioni et al., 2005), são sistemas de extracção de informação da Web, de forma

automática, sem supervisão e com capacidades de aprendizagem. Estes sistemas recorrem a vários serviços para extrair e identificar informação específica para inserir num repositório. A informação recolhida é usada na descoberta de novas instâncias, criando regras baseadas na redundância da informação. Esta metodologia permite uma expansão contínua e automática da base de conhecimento. Por exemplo, o Armadillo é capaz de extrair nomes de filmes de texto, sendo capaz de reconhecer e de relacionar títulos de filmes como "The big chill" e "big chill, The". O Armadillo extrai informação de vários serviços Web que, por sua vez, têm funções específicas e recorrem a outros sistemas. Por exemplo, um serviço de reconhecimento de entidades mencionadas de um sítio de uma universidade recorre a um sistema de reconhecimento de entidades mencionadas para identificar potenciais nomes. Outros serviços procurariam obter artigos, da autoria de um investigador identificado no serviço anterior, do CiteSeer ou do DBLP. Cada serviço produz resultados pouco fiáveis só por si, de pouca precisão, mas a combinação dos diversos serviços produz resultados com uma precisão alta.

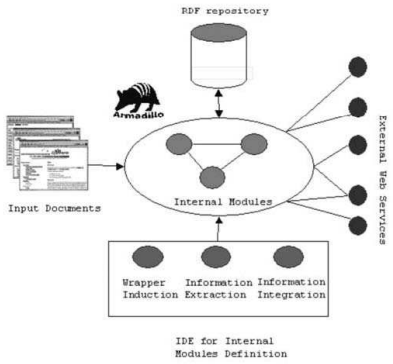


Figura 3.1: Arquitectura do Armadillo, extraída de Ciravegna et al. (2004)

A informação obtida pelos vários serviços é de seguida integrada, através de ontologias, num repositório RDF, onde é armazenada. Explorando a redundância da informação na Internet e posteriormente no repositório gerado, o Armadillo extrai informação com diferentes graus de confiança e expande a sua base de conhecimento inicial. Esta metodologia evita a aquisição de informação espúria baseada em informações erradas. O Armadillo funciona com o mínimo de intervenção humana: o utilizador fornece um URL e alguma

informação adicional, não requerendo anotações manuais. Após a intervenção do utilizador, os dados que este alterou, apagou ou adicionou, podem ser usados novamente para reiniciar a aprendizagem de forma a obter mais informação e maior precisão.

O knowItAll, por outro lado, é um sistema que permite a pesquisa e extracção de informação independente do domínio. Usa um conjunto de oito padrões para determinar candidatos a facto através da instanciação de uma classe. Por exemplo, é possível extrair os facto CIDADE("Porto") e CIDADEDE("Porto", "Portugal") a partir de texto comum, como

> ...a cidade do Porto, em Portugal ...

O KnowItAll assenta essencialmente em três métodos distintos:

- Aprendizagem de padrões capazes de serem usados tanto com regras de extracção como de validação das instâncias extraídas
- Extracção de subclasses (por exemplo, é capaz de extrair subclasses de cientista (tais como físicos, geólogos, etc.)
- Capaz de extrair listas de classes, através da aprendizagem de padrões.

Ao contrário do Armadillo, este método dispensa a marcação de textos para aprendizagem, dado que a informação extraída pelos padrões é carregada no módulo de BootStrapping de forma a gerar procuras (para usar nos motores de busca) e regras de extracção.

Estes sistemas dependem do processamento de texto e da identificação correcta da informação recolhida, quer através de heurísticas quer através da criação de recursos de informação que permitam criar regras para identificar instâncias.

3.2.1.1 Wrappers

A maioria dos sistemas de extracção de informação na Web usa *wrappers* para extrair informação de documentos no formato HTML de um sítio e converter essa informação para um formato estruturado. Os *wrappers* podem ser criados manualmente ou semi-automaticamente.

A criação e treino de *wrappers* requer o treino individual para cada sítio através de aprendizagem manual ou semi-automática (veja-se Ashish e Knoblock (1997) e Geng (2002)). No caso de sistemas que exigem a extracção de texto de sítios não especificados, tal torna-se impraticável. Um método alternativo, usando por outros sistemas, é o uso de heurísticas simples para obter a informação desejada. Estas heurísticas normalmente aplicam-se através da análise da estrutura de documentos HTML (Geng (2002), Soricut e Brill (2006) e Agichtein et al. (2004)), dando relevância e tentando construir informação que esteja interligada com base nas marcas (*tags*) de HTML. Ou seja, é possível mapear informação com base nas marcas que fornecem informação visual com listas ($< li >$), parágrafos ($< p >$), quebras de linha ($< br >$), elementos de tabelas($< tr >$ e $< td >$), etc.

3.2.2 Extracção de informação bibliográfica

A extracção de referências bibliográficas a partir de referências em texto envolve não só a separação dos elementos mas requer também que esses sejam correctamente identificados. Existem diversas técnicas para este processo.

3.2.2.1 O ParaTools

O ParaTools (Jewell, 2003) é uma colecção de módulos Perl cujo objectivo é o de processar referências bibliográficas. O ParaTools é composto por duas ferramentas específicas:

- Analisador sintáctico baseado em modelos (templates), comparando com uma lista de 400 padrões, para obter os elementos bibliográficos. Apesar de fixa, a lista de padrões pode ser facilmente aumentada.

- Analisador sintáctico compatível com o CiteBase, um serviço do OAI, que permite processar referências de revistas académicas mas é pouco útil para os restantes tipos de publicações (actas de conferências, livros, teses, etc.)

3.2.2.2 Métodos estatísticos

Existem outras implementações com um objectivo semelhante. Por exemplo, Huang et al. (2004) usa uma técnica semelhante ao Paratools mas baseada em algoritmos genéticos para fazer o alinhamento. Geng (2002) usa cadeias de Markov escondidas para identificar os elementos bibliográficos. Estas implementações recorrem a aprendizagem automática. Os sistemas são treinados com exemplos, de forma a gerarem caminhos (Geng, 2002) ou expressões genéticas (Huang et al., 2004) probabilísticas.

3.2.2.3 Reconhecimento

Outra abordagem recorre a técnicas de identificação de entidades mencionadas tal como no SIÊMES (Sarmento, 2006), comparando os elementos por identificar com um repositório de exemplos de entidades, semelhantes ao REPENTINO (?).

3.3 Organização de recursos: Pesquisa e gestão

Os programas para gestão de referências bibliográficas são já bastante conhecidos do grande público. Podem ser separados em dois grupos distintos; programas para uso individual e programas cooperativos.

3.3.1 Programas para uso individual

Os programas para uso individual são provavelmente os mais comuns. Destinam-se a funcionar localmente no computador do utilizador, permitindo gerir uma lista de referências bibliográficas e proporcionar meios para gerar listas para associar com outros programas, nomeadamente editores de texto, como o Microsoft Word, Latex, OpenOffice, etc. Com o surgimento de serviços Web, tem-se tornado também possível a estes programas disponibilizarem consultas a repositórios bibliográficos, permitindo a inserção de informação estruturada nas bases de dados dos utilizadores. Exemplos destes programas são:

- Jabref[7]
- EndNote
- RefTeX[8]
- Reference Manager

3.3.2 Programas cooperativos

Em paralelo com catálogos de bibliotecas ou de editoras, existem repositórios dedicados a domínios específicos, sistemas de gestão na Web para serem usados de forma cooperativa. Os programas cooperativos são mais usados em ambientes Web, e estão associadas ao surgimento da Web 2.0. Estes recursos recorrem a métodos de inserção manuais, métodos automáticos através de extracção de informação de documentos Web, e à partilha de informação através do *download* das referências bibliográficas em diversos formatos, como o BibTeX por exemplo.

Estes gestores têm uma particularidade: Fornecem um meio de classificação manual que é bastante poderoso, uma vez que é cooperativo. Ou seja, cada referência que o utilizador insira pode ser classificada por diversas marcas (*tags*), não só pelo utilizador que a inseriu mas também por todos os outros utilizadores. Isto permite criar uma rede de interesses de utilizadores conhecida como *folksonomy* (Mika, 2005; Feitelson, 2000; Golder e Huberman, 2006), o que será descrito em pormenor ainda neste capítulo. Para terminar, é apresentada uma lista de alguns dos programas cooperativos disponíveis:

- CiteUlike[9]
- Connotea[10]
- eprints[11]

[7] http://jabref.sourceforge.net/
[8] http://staff.science.uva.nl/ dominik/Tools/reftex/
[9] http://www.citeulike.org/
[10] http://www.citeulike.org/
[11] http://www.eprints.org/

- Bibsonomy[12]

3.4 A Web 2.0 e as tecnologias associadas

Dado que esta dissertação pretende abordar essencialmente a descoberta de informação na Web, pretendendo facilitar a interacção com o utilizador, não se pode concluir este capítulo sem referir algumas das tecnologias e métodos usados hoje em dia para satisfazer este propósito. É necessário fazer menção à Web 2.0, um conjunto de tecnologias e conceitos que revolucionou a interoperabilidade e usabilidade de aplicações Web (O'Reilly, 2005).

3.4.1 O Ajax

O Ajax, que significa *Asynchronous JavaScript and XML*, é uma combinação de tecnologias que permite uma maior interacção das aplicações Web, alterando a maneira de pensar na arquitectura de aplicações Web. As tecnologias que compõem o Ajax são:

- o XHTML (ou HTML) e CSS para apresentação
- o DOM para manipulação da página
- troca de dados assíncrona entre o browser e o servidor através do objecto **XMLHttpRequest**
- o XML e XSLT para transmissão de dados entre o browser e o servidor
- JavaScript

A combinação destas tecnologias permitiu criar um enquadramento para aumentar a interacção em aplicações Web. O Ajax usa comunicação assíncrona através do objecto **XMLHttpRequest** para trocar pequenas quantidades de dados com o servidor. A informação recebida é depois usada para refrescar a página actual, ou apenas parte, para ser mais preciso. Recorrendo ao JavaScript, é possível substituir o conteúdo de qualquer objecto da página. Não é necessário recarregar a página completa. O uso de Ajax em aplicações Web

[12]http://www.bibsonomy.com

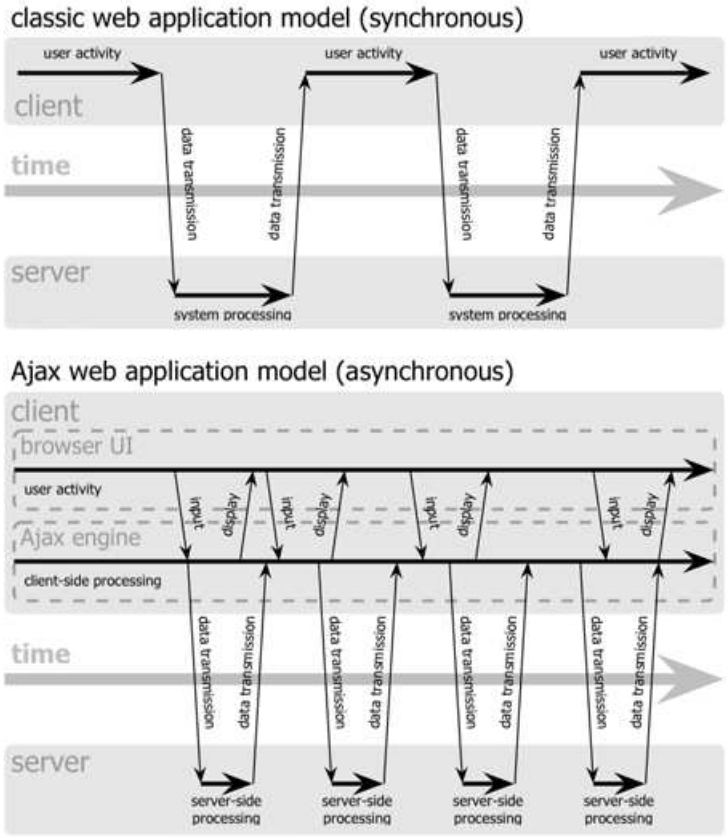


Figura 3.2: Comparação entre a comunicação clássica e através de Ajax(retirado de `http://www.adaptivepath.com/publications/essays/archives/000385.php`)

cresceu imenso, desde o seu uso inicial, experimental, em aplicações simples como o *Google Suggest*[13] ou *Google Maps*[14]. Actualmente existem aplicações mais complexas, como:

- Ambientes de correio electrónico;
- Editores de texto, como o Writely[15];
- Editores de imagens;
- Ou mesmo ambientes de trabalho completos, possuindo engenhocas (*Widgets*) como editores de texto, reprodutores de áudio, ou outras

[13] `http://www.google.com/webhp?complete=1&hl=en`

[14] `http://maps.google.com`

[15] Em 2006 foi comprado pela Google passando a ter o nome de Google Docs. `http://docs.google.com`

funcionalidades fornecidas por terceiros. Um exemplo é a capacidade de visualizar um álbum de fotos do Flickr[16].

3.4.2 Folksonomias e ontologias

3.4.2.1 Ontologias

Uma ontologia é uma especificação do conhecimento de um domínio (Gruber, 1993). Usa um vocábulo controlado e uma estrutura hierárquica para descrever objectos e as relações entre eles. As ontologias representam-se através de meta-informação, como o RDF, uma linguagem para representar informação. As ontologias são usadas em diversas áreas como a Web semântica (através, por exemplo, da linguagem OWL[17]), a inteligência artificial ou mesmo para descrever documentos académicos, com o *Dublin Core*[18], para representar conhecimento. Têm ainda como objectivo integrar a informação de diversas fontes e aumentar a interoperabilidade entre os sistemas.

3.4.2.2 Folksonomias

Folksonomias, em inglês *Folksonomies*, é um termo recente. Curiosamente, a definição para folksonomia pode ser encontrada em Wikipedia. Esta definição é citada em diversos artigos (Vazquez et al., 2006), mas sem referir versão ou data da página da Wikipédia, onde os conteúdos estão. Talvez por isso o criador do termo, Thomas Vander Wal, deu a seguinte definição (Wal, 2005) após encontrar 15 citações à definição da Wikipédia:

```
Folksonomia é o resultado da marcação pessoal e livre de informação
 e objectos (qualquer conteúdo com URL) para uso próprio. É executado
 num ambiente social, partilhado e aberto a outros.
```

As folksonomias são um meio de evitar a criação de taxonomias próprias, que levam tempo a construir podendo não ser apropriadas para um grupo de utilizadores. São facilmente adaptáveis a novos conteúdos ou alterações, no

[16] http://www.flickr.com/
[17] http://www.w3c.org/2001/sw/WebOnt/
[18] http://dublincore.org

sentido de que os utilizadores podem criar novas marcas para se adaptarem a novos conteúdos. Por último, as folksonomias proporcionam como maior benefício a capacidade de as marcas representarem a relevância dos conteúdos para que apontam. As folksonomias oferecem uma alternativa aos métodos tradicionais de pesquisa através de motores de pesquisa, podendo facilitar as pesquisas com base em marcas em vez de conteúdo das páginas. O uso de folksonomias é, portanto, uma forma inovadora de categorizar conteúdos possibilitando o uso de taxonomias pessoais e partilhando essa informação com toda a comunidade (Mika, 2004). Sítios como o Del.icio.us[19] permitem categorizar URL, o Flickr permite categorizar fotos, CiteULike ou o Bibsonomy permitem categorizar referências bibliográficas. A marcação livre do Gmail[20], por exemplo, não deve ser considerada uma folksonomia *privada*, uma vez que consiste simplesmente num sistema de *tagging*, não sendo feito num ambiente social nem partilhado.

No entanto, as folksonomias não são perfeitas:

- Limitação a uma palavra apenas. Muitos sítios limitam a marcação a uma palavra, não permitindo o uso de expressões.

- O uso de sinónimos não tem qualquer controlo, o que leva a múltiplas marcações com o mesmo significado ("carro", "automóvel", "car"), inclusive entre singular e plural ("livro", "livros").

- A marcação através de vocábulos não controlados pode levar a ambiguidades em situações que as marcas usadas sejam muito subjectivas. Por exemplo, dois documentos, um sobre linguística computacional e outro sobre inteligência artificial, podem ambos ser marcados como "programação", no entanto podem ser considerados assuntos distintos, uma ramificação que pode ser necessário distinguir.

Estas propriedades são apontadas pelos detractores das folksonomias como causas para a geração de demasiado "ruído", reduzindo assim a utilidade da informação. Adicionalmente, os defensores do uso de taxonomias/ontologias

[19] http://del.icio.us/
[20] http://www.gmail.com

defendem que o uso de tags livres reduz consideravelmente a eficiência na indexação de dados.

Capítulo 4

SUPeRB - Um sistema de tratamento de informação bibliográfica

Neste capítulo é proposta uma arquitectura para o SUPeRB, um sistema que tem como objectivo a automatização da pesquisa bibliográfica na Web, para ser facilmente integrada num repositório existente.

O SUPeRB é um sistema interactivo, ou seja, pressupõe a interacção com um utilizador, mas sem exigir deste demasiados conhecimentos técnicos. Assume-se que o utilizador pretende pesquisar e processar informação bibliográfica partindo de três tipos de parâmetros:

- uma expressão simples, que contenha informação suficiente para identificar uma publicação ou um conjunto de publicações. Um exemplo pode ser uma combinação de palavras-chave ou o nome de um autor;
- uma referência bibliográfica incompleta de onde se pretendem obter os outros elementos bibliográficos de forma a que esta fique completa;
- um URL, em que o utilizador sabe que existem referências bibliográficas relevantes.

Dado um destes parâmetros ao SUPeRB, o sistema recorre à Web através de serviços Web para obter documentos ou informação bibliográfica estruturada,

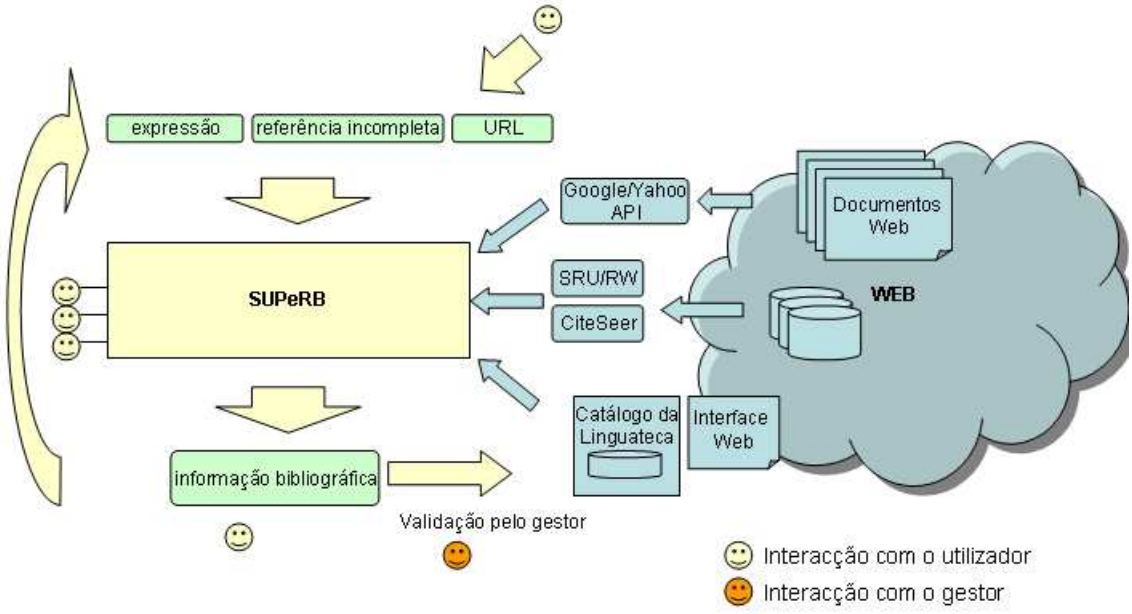


Figura 4.1: O sistema SUPeRB

como é apresentado na figura 4.1. O resultado da consulta à Web é processado e combinado de forma a obter informação bibliográfica relevante em relação aos parâmetros dados. Por fim, essa informação deve ainda ser organizada de forma a ser fácil de inserir no repositório bibliográfico, neste caso o catálogo de publicações da Linguateca. Outra alternativa é a de utilizar a informação obtida para obter mais informação, quer por interacção do utilizador, quer automaticamente, em acções periódicas de actualização.

Todo o processo de obter documentos relevantes e de extrair e fundir a informação obtida é monitorizada em diversas fases. Esta monitorização permite a validação dos resultados obtidos nas várias fases, além da validação final dos resultados.

4.1 A arquitectura geral do SUPeRB

O SUPeRB é constituído por um conjunto de módulos em que cada módulo é responsável por uma tarefa específica.

Alguns destes módulos têm a capacidade de serem invocados remotamente,

dado que não requerem acesso a informação adicional. Por exemplo, o processamento de referências bibliográficas é uma tarefa que pode ser executada remotamente ou em paralelo. Assim, é possível que determinadas tarefas possam obter um desempenho melhor. É também possível considerar cada tarefa como uma componente distinta do SUPeRB. Por outro lado, a supervisão e validação humana é outra das tarefas facilitadas pelo sistema, sendo possível validar ou avaliar os resultados produzidos por cada componente através de interfaces próprias.

Esta divisão em componentes é relativa ao processamento de informação. Os módulos nestas camadas necessitam de interagir com vários tipos de informação. De um lado, existem as interfaces que permitem a interacção com utilizadores, do outro lado existe a informação que é acedida e armazenada fisicamente. Existem portanto outras camadas com responsabilidades diferentes. O SUPeRB apresenta assim 3 camadas, conforme é visível na figura 4.2:

- A camada de interface, que permite a interacção com o utilizador através de páginas dinâmicas e serviços Web.
- A camada lógica, responsável por diversas tarefas de processamento de informação. Incluem-se aqui também os métodos de acesso a serviços externos.
- A camada de base de dados, responsável pelo armazenamento dos dados. Esta camada é composta pelos recursos físicos e pelas interfaces que permitem o acesso a esta informação a partir das camadas acima.

Neste capítulo será focada a camada lógica, revelando como alguns do problemas foram resolvidos. Apresenta-se ainda a camada de interface, apresentando o funcionamento e as suas vantagens.

4.1.1 Interligação entre componentes

As componentes desenvolvidas facilitam a interoperabilidade entre si e entre as camadas que comunicam, nomeadamente as interfaces dos utilizadores. O funcionamento de cada componente pode ser obtido sob a forma de um

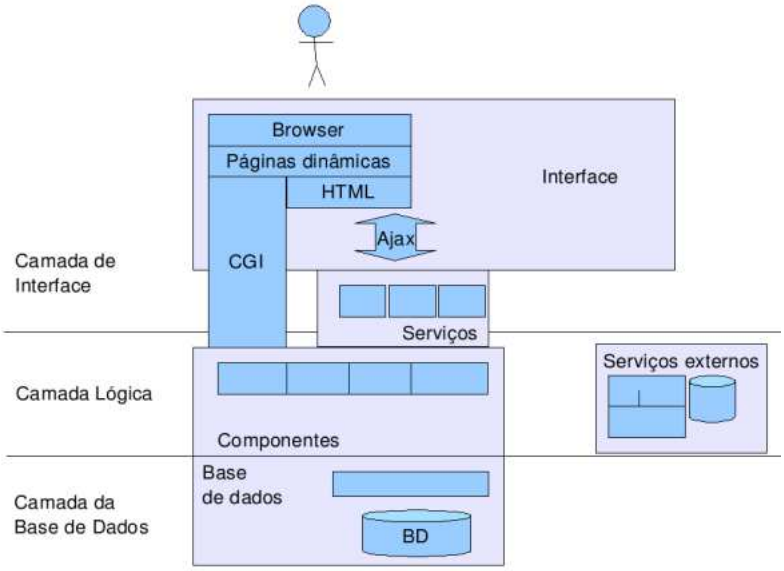


Figura 4.2: Camadas do SUPeRB

serviço, possibilitando a execução em máquinas remotas ou com programas concorrentes. O uso de XML com método para comunicar entre as diversas componentes apresentou-se como uma solução fácil e eficaz. O XML permite a troca de informação entre componentes de forma estruturada e sem restringir a execução a uma única máquina. É possível igualmente transmitir os dados a serem processados ou para serem exibidos numa interface do utilizador, permitindo a este prever e escolher sobre quais efectuar as tarefas seguintes. A figura 4.3 apresenta um exemplo de informação em XML.

Além disso, usando este método, é possível usar os mesmos meios para trocar informação estruturada entre as interfaces Web e o servidor. Isto será discutido na secção 4.3, onde se descreve a interface e as metodologias e tecnologias aplicadas.

4.2 As tarefas do SUPeRB

O processo de descoberta de referências bibliográficas pode ser dividido em diversas fases, sendo possível decompor o sistema em diversos módulos cujo objectivo é distinto. Assim, a estruturação em módulos visa facilitar a

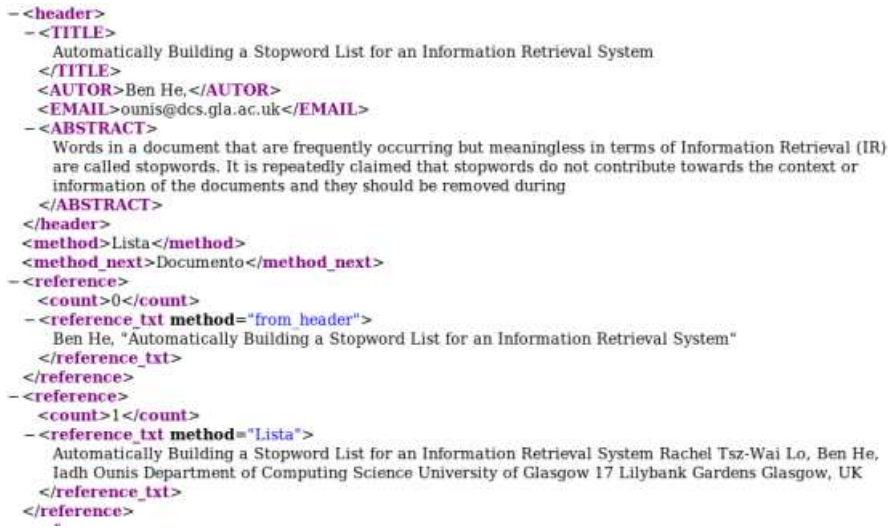


Figura 4.3: Exemplo de informação em XML contendo informação bibliográfica extraída de um documento

construção de programas que possam usufruir de algumas das funcionalidades fornecidas. Na figura 4.4 é apresentado o conjunto das tarefas que compõem o SUPeRB.

A primeira tarefa, a pesquisa na Web, recebe um conjunto de parâmetros e devolve como resultado um conjunto de URL para documentos relevantes na Web. Os resultados desta tarefa são dados como argumento à tarefa seguinte. O fluxo do sistema decorre desta forma, até chegar finalmente à última tarefa, a classificação, após a qual os dados são submetidos para posterior validação pelo gestor do catálogo. Na figura são também apresentadas as diversas fases de validação dos parâmetros obtidos por algumas das tarefas.

4.2.1 Pesquisa na Web

A pesquisa na Web recebe um parâmetro de entrada, que é:

- Uma expressão, texto simples dado pelo utilizador.
- Ou uma referência bibliográfica incompleta.

Na primeira situação, quando é dada uma expressão como parâmetro, o sistema pode utilizar essa expressão, sem qualquer refinamento, para invocar motores de

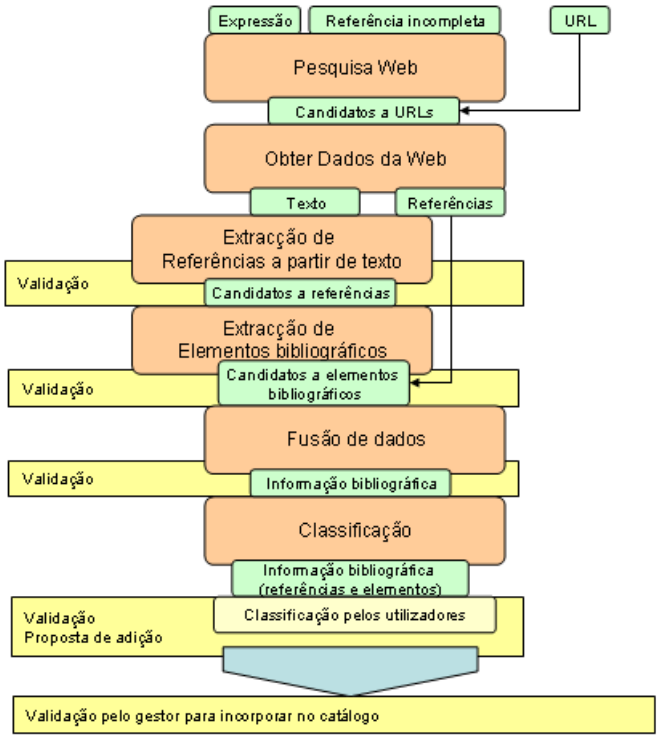


Figura 4.4: Tarefas do SUPeRB

pesquisa Web e obter um conjunto de URL. No entanto os resultados produzidos dependem da expressão usada, podendo ocorrer duas situações nesta etapa:

- A expressão é demasiado específica e não produz suficientes resultados;
- A expressão é muito simples e não produz resultados relevantes.

Assim, tendo em vista melhorar a possibilidade de se obter um conjunto de resultados relevantes, são efectuadas várias operações de refinamento. Para além de se usar a expressão dada como parâmetro, são produzidos vários tuplos, combinações de um número limitado de palavras extraídas da expressão dada como parâmetro. Para evitar a geração de expressões pouco significativas, a lista de palavras extraída exclui palavras muito pequenas, que não são indexadas pelos motores de pesquisa. Esta abordagem oferece-nos duas vantagens:

- Permite simplificar a expressão dada de forma a obter resultado melhores.

- Baroni e Bernardini (2004) mostraram que a geração de múltiplas pesquisas ao motor de pesquisa Google alterando a ordem das palavras, produz resultados diferentes.

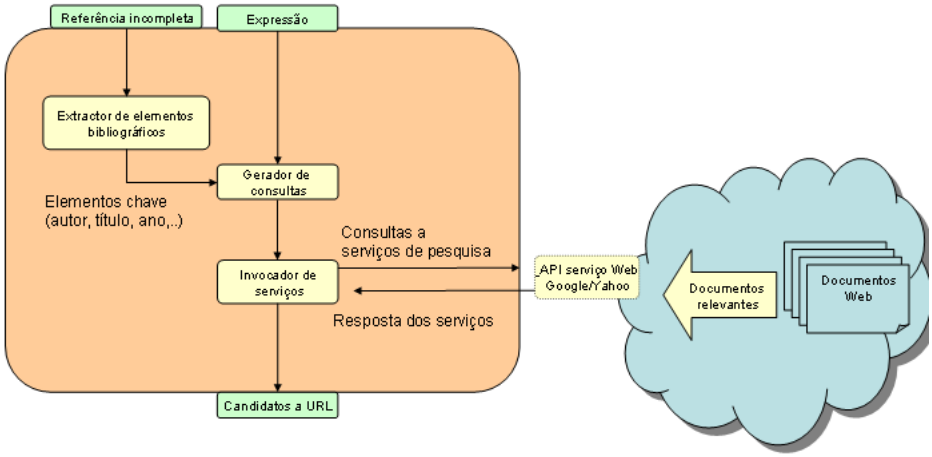


Figura 4.5: Tarefa de pesquisa na Web

Dado que ao aplicar este método é possível que se perca informação, tornando os tuplos gerados em expressões demasiado gerais que podem produzir resultados pouco relevantes, é tomada uma precaução adicional. Numa segunda fase são adicionadas palavras específicas do contexto bibliográfico aos tuplos gerados. Ou seja para além de se gerar tuplos com palavras fornecidas pelo utilizador, adiciona-se ainda uma palavra adicional, de uma lista de palavras apresentada na tabela 4.1.

Construídos os diversos tuplos, a próxima fase do sistema consiste em usar os tuplos gerados, do qual faz parte a expressão original, para invocar motores de pesquisa como o Google e o Yahoo através de interfaces próprias, API que permitem interagir com o serviço Web fornecido e devolver informação estruturada. O uso de serviços Web evita o processamento adicional das páginas HTML de respostas dos motores de pesquisa, possibilitando tratar mais facilmente a informação recolhida, URL, título e sumário.

Após a resposta dos serviços dos motores de pesquisa, o sistema possui uma lista de URL para documentos na web que possam ter informação bibliográfica relevante. Um exemplo de uma lista de tuplos gerados a partir de uma expressão dada como parâmetro é apresentado a seguir. A expressão "*The semantic Web*

Tabela 4.1: Lista de palavras usadas para adicionar aos tuplos gerados

publicações
publications
referências
references
artigo
article
academic
pdf
documentos
documents
bibliografia
bibliography

Revisited Shadbolt"pode assim produzir os seguintes tuplos para consulta na tabela 4.2:

Tabela 4.2: Lista de expressões geradas a partir de palavras usadas para adicionar às expressões geradas

The semantic Revisited Shadbolt publicações
The semantic Revisited Shadbolt references
The Shadbolt semantic Web referencias
Shadbolt The Revisited semantic referencias
Shadbolt The Revisited semantic publications
Shadbolt The Web Revisited documents
Shadbolt The Web Revisited referencias
The Revisited Web semantic documentos

Falta ainda descrever a situação em que é dada uma referência bibliográfica incompleta. Para aliviar o processamento desta informação, pode-se admitir que os diferentes elementos bibliográficos possam ser transmitidos pelo utilizador através de um formulário. Caso a informação seja dada em texto, é necessário extrair e identificar os elementos bibliográficas, usando o módulo de extracção de elementos bibliográficos que vai ser apresentado na secção 4.2.4.

Obtidos os elementos bibliográficos dados como parâmetros, pretende-se usar apenas os elementos mais relevantes, como o AUTOR, TÍTULO, CONFERÊNCIA ou ANO. A escolha dos elementos bibliográficos usados na pesquisa tem como objectivo gerar expressões que sejam produtivas. São utilizadas combinações

dos elementos, por exemplo os da tabela 4.3

Tabela 4.3: Lista de combinações possíveis

AUTOR + TITULO + CONFERENCIA + ANO
AUTOR + TITULO + ANO
AUTOR + CONFERENCIA + ANO
AUTOR + TITULO
TITULO + CONFERENCIA

São geradas expressões como no caso anterior em que é dada uma expressão, mas, neste caso, a geração de expressões pode ser mais organizada. Outras palavras, que não são utilizadas nesta tarefa, têm interesse para tarefas seguintes. Expressões como o nome de um autor, a data completa (dias, mês e ano), número das páginas, editores, "proceedings", são pouco relevantes para a fase de pesquisa na Web se se possuir os elementos mais relevantes como nome completo ou último nome do autor, o título, ano e o nome ou abreviatura da conferência.

4.2.2 Análise dos URL e obtenção de conteúdos

Na última secção foi descrito como obter uma lista de URL relevantes a partir da Web. Nesta fase procede-se à análise e tratamento desses URL dados. Assim, dependendo da situação, os URL são processados da seguinte forma:

- O URL pertence a uma lista de URL a ignorar. Neste caso o URL é descartado. Esta situação ocorre com sítios que possuam documentos que não se queiram considerar pelo SUPeRB. Um exemplo é o caso das páginas do próprio catálogo, uma vez que é possível aceder a esta informação directamente. Outros casos em que não se queira considerar os documentos são:
  - ou por se ter conhecimento de que não contêm informação bibliográfica relevante para o domínio;
  - ou se trate de documentos com listas de palavras, por exemplo;

- O URL já foi descarregado e processado recentemente? É possível que a informação já tenha sido processada ou pelo menos parte dessa informação?
- O URL pertence a um repositório a que o SUPeRB possa aceder através de serviços Web. Pode-se em alternativa usar um serviço próprio para recolher a informação bibliográfica.
- O URL refere outros documentos da Web, dos quais é possível extrair informação processando o próprio documento.

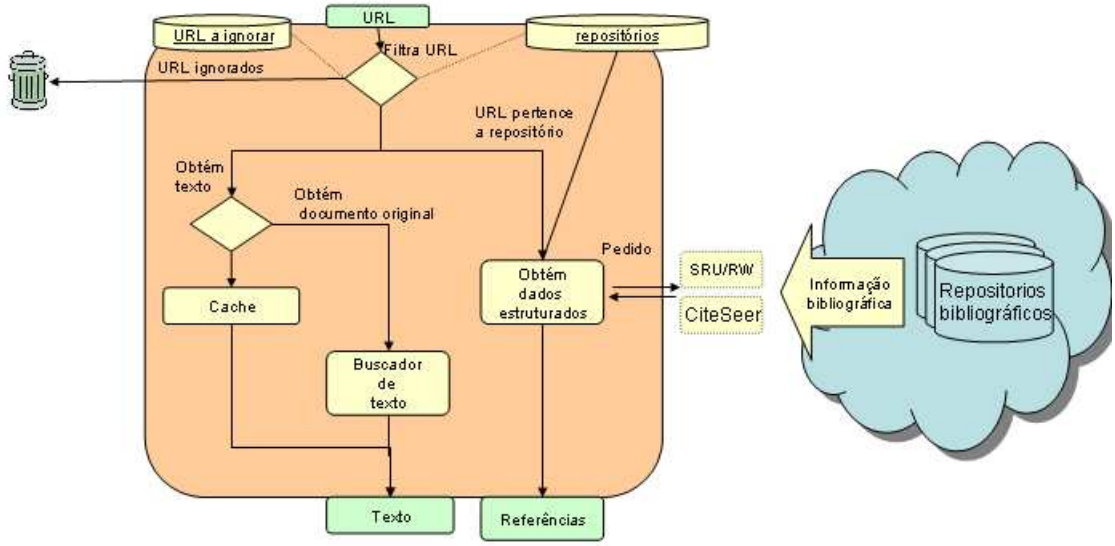


Figura 4.6: Tarefa de análise e obtenção da informação dos respectivos URL

Considerando estas opções é possível determinar o melhor método de obter a informação pretendida. Nos dois primeiros casos, a informação pode ser ignorada ou pode já ter sido acedida e processada. Nos restantes casos, é possível obter a informação através de métodos distintos, extraindo e processando o documento em questão ou recorrendo a serviços próprios.

4.2.2.1 Obtenção de informação a partir de documentos Web

A informação na Web pode ser encontrada em inúmeras formas. A informação bibliográfica não é excepção. Recorrendo a pesquisas Web, podemos encontrar

informação bibliográfica em todos os tipos de formatos. No entanto apenas são relevantes para o SUPeRB documentos de onde seja possível extrair texto. É necessário identificar o tipo de documento para poder escolher o programa correcto a utilizar e assim extrair correctamente a informação que este contém. Até ao momento foram especificados tratamentos para os seguintes formatos, quer pelo uso de aplicações já existentes, quer pela criação de aplicações para esse fim:

- Postscript (PS)
- Acrobat format (PDF)
- Rich Text Format (RTF)
- Word Document (DOC)
- PowerPoint (PPT)
- Hiper Text (HTML)

Os documentos são copiados para o servidor e o tipo do documento é determinado pela sua extensão. Quando este método falha, é ainda possível recorrer ao *Mime Type* do documento. Após determinado o tipo, é escolhido o conversor correcto e o texto é extraído. Imagens ou outro tipo de multimédia que possa estar presente no documento são descartados.

O texto passa ainda por um processo de limpeza, nomeadamente para remover caracteres ilegíveis ou com problemas na acentuação, normalmente causados pelo programa de extracção. Por exemplo, é frequente que os acentos nos caracteres sejam colocados antes ou depois do caracter. Este método tenta identificar qual o caso e corrigir da forma adequada.

Quando não for possível processar um determinado documento, pode-se recorrer ao uso de outros serviços que transformem os documentos para um formato que o SUPeRB seja capaz de processar. Por exemplo, o Google possui uma cache que armazena documentos em formato HTML. É possível assim que, em determinadas situações, se utilize a cache do Google em alternativa ao documento original. Esse processo encontra-se descrito na figura 4.7

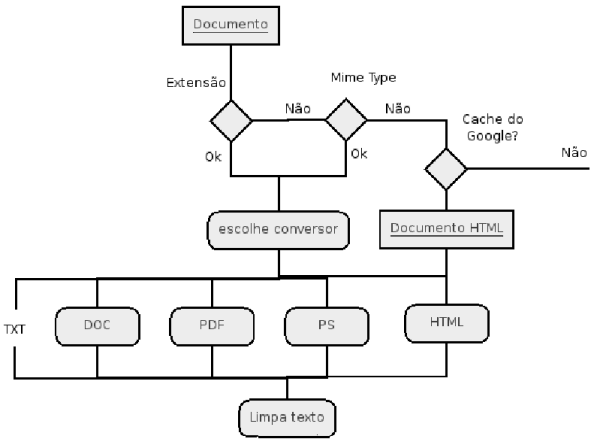


Figura 4.7: Decisão da aplicação a usar para obter o conteúdo no formato de texto

4.2.2.2 Obtenção da informação de repositórios bibliográficos

Como descrito em 3.1.1, mapeando repositórios Web que possam ser acedidos através de serviços Web permite obter informação bibliográfica estruturada. Para aceder aos repositórios bibliográficos foram usados os seguintes métodos:

- Os protocolos SRU /SRW;
- A API do CiteSeer.

A API do CiteSeer proporciona o acesso estruturado a um recurso enorme na área de ciência de computadores, e que está indexado pelos motores de busca usados, apresentando resultados relevantes.

4.2.3 Extracção de referências a partir de texto

A fase seguinte consiste na extracção de possíveis candidatos a referências bibliográficas ou informação bibliográfica em geral a partir dos textos extraídos. Os textos extraídos de documentos Web podem provir de diversas fontes, tais como:

- Listas de referências bibliográficas de páginas de autores;

- Páginas de conteúdos de actas, com uma ou mais referências bibliográficas;
- Documentos académicos, com uma estrutura padrão, com dados relevantes, no início do documento ou no fim do documento;
- Apresentação (conjunto de slides) que pode conter alguma informação relevante (por vezes possuem uma estrutura semelhante aos documentos académicos, com informação bibliográfica tanto no início como no fim).

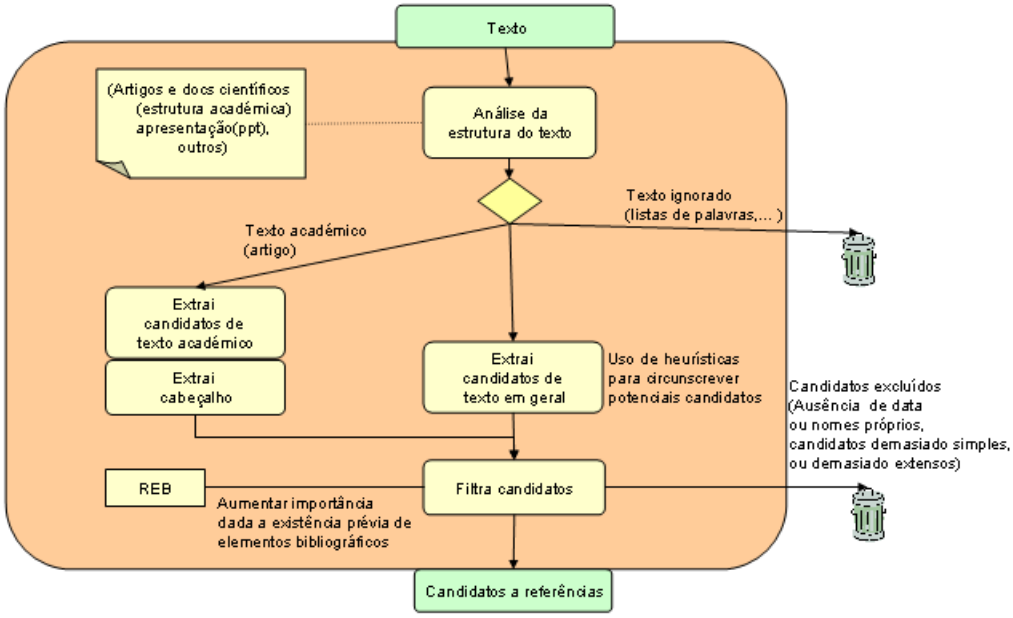


Figura 4.8: Tarefa de extracção de referências do texto

Torna-se portanto importante determinar o tipo de documento de onde o texto foi obtido para optar pela melhor forma de extrair a informação bibliográfica relevante. Assim, é possível seguir diferentes abordagens, aplicando regras específicas para cada caso.

4.2.3.1 Identificação da estrutura do documento

Este módulo tem como objectivo identificar a estrutura dos documentos face a um conjunto de estruturas pré-definidas. Esta informação será depois utilizada para tentar deduzir outras informações, tais como:

- Tipo de publicação;
- Relação com outros documentos, como por exemplo, identificar uma apresentação ou um poster relacionado com um artigo.

Para atingir este objectivo, é necessário recorrer à aplicação de várias heurísticas simples. Se necessário, pode-se ainda:

- Considerar a extensão do documento original. Documentos em Powerpoint, com extensão *ppt* ou *pps*, são potenciais apresentações;
- Aplicar as regras directamente ao documento original, como no caso do HTML, onde é possível analisar a estrutura do hipertexto.

Por outro lado, os documentos em hipertexto podem ser também analisados tendo em conta a sua estrutura interna. É possível encontrar documentos académicos como artigos, relatórios, manuais, etc., neste formato. Mas é também possível encontrar outros tipos de informação, como simples listas de referências bibliográficas, por exemplo.

De seguida, na tabela 4.4, são apresentadas heurísticas em linguagem natural, que são usadas para determinar a estrutura do documento.

Tabela 4.4: Exemplos de heurísticas para determinar a estrutura do documento

Tipo de estrutura	Heurísticas
Documento académico	Bloco inicial que começa com "resumo" (primeiros 10% do documento) Bloco final identificado por "referências" (últimos 10% do texto)
Listas de publicações	Início do texto (ou da lista) identificado por "Publicações", "Referências", etc. Em hipertexto a frequência de marcas separadores é alta como por exemplo "LI"(listas) ou "P"parágrafo)
Apresentação em slides	Formato do documento é Powerpoint Média de palavras por frase é baixa
Lista de palavras	Média de palavras por frase/linha é baixa Poucos caracteres de pontuação

Estas heurísticas podem ser usadas para identificar tipos de documentos, podendo inclusive excluir um documento. Por exemplo, o último caso na

tabela 4.4 refere-se a dicionários de palavras que podem ser excluídos. Estas heurísticas podem ser adaptadas para documentos em diversas línguas. No entanto, o português e o inglês foram considerados mais relevantes para este trabalho. Assim, actualmente apenas estão a ser considerados documentos nestas duas línguas.

Após determinar o tipo de documento, adopta-se a melhor estratégia para obter candidatos com informação bibliográfica. As estratégias escolhidas variam dependendo do tipo de estruturas encontrados no documento.

4.2.3.2 Extracção de informação bibliográfica do cabeçalho de um documento (Auto-referência)

Este processo consiste em obter do próprio documento informação que o identifique. Os documentos académicos possuem por norma um cabeçalho com informação bibliográfica onde se pode encontrar o nome do autor ou autores e o título. Outras informações podem também ser encontradas no início do documento, nomeadamente, o resumo, afiliações dos autores ou moradas e contactos. No exemplo da figura 4.9, é apresentado um exemplo de um texto, extraído do início de um documento académico.

Extracting Citation Metadata from Online Publication Lists Using BLAST

I-Ane Huang1, Jan-Ming Ho1, Hung-Yu Kao1, Wen-Chang Lin2
1
Institute of Information Science, Academia Sinica, Taiwan
2
Institute of Biomedical Sciences, Academia Sinica, Taiwan

hia@iis.sinica.edu.tw, hoho@iis.sinica.edu.tw,
bobby@iis.sinica.edu.tw, wenlin@ibms.sinica.edu.tw

Abstract. Scientific research reports require a great deal of citation, therefore an automatic citation tool would be of great use. Due to numerous models, it is
....

Figura 4.9: Exemplo de um bloco de texto extraído do início de um documento PDF

Quando é possível identificar um cabeçalho deste género, é possível aplicar

algumas heurísticas simples para obter a informação bibliográfica. Na figura 4.10 são apresentados alguns dos dados que é possível obter.

título	Extracting Citation Metadata from Online Publication Lists Using BLAST
autor	Ane Huang
autor	an-Ming Ho
autor	Hung-Yu Kao
autor	Wen-Chang Lin
afiliação	Institute of Information Science, Academia Sinica, Taiwan
afiliação	Institute of Biomedical Sciences, Academia Sinica, Taiwan
email	hia@iis.sinica.edu.tw
email	hoho@iis.sinica.edu.tw
email	bobby@iis.sinica.edu.tw
email	wenlin@ibms.sinica.edu.tw
resumo	Scientific research reports require a great deal of citation, therefore an automatic citation tool would be of great use. Due to numerous models, it is...

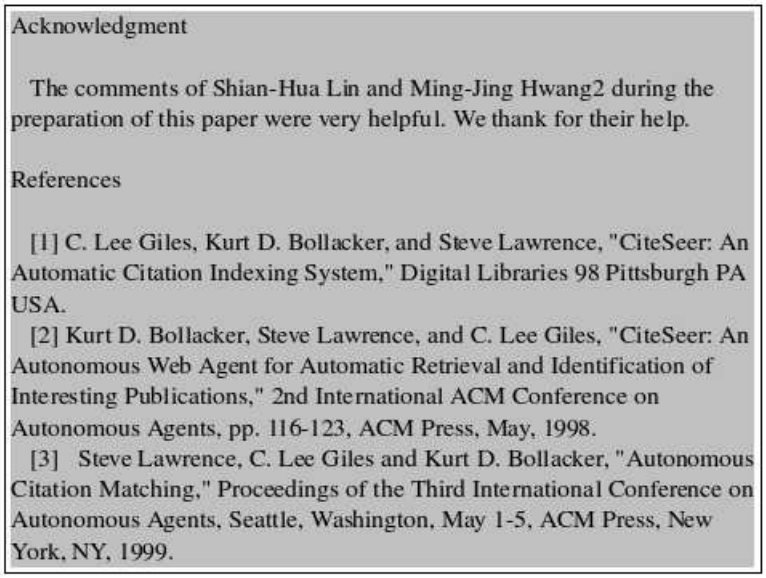
Figura 4.10: Informação extraída do exemplo da figura 4.9

4.2.3.3 Extracção de informação do fim do documento

Outra característica dos documentos académicos é possuírem um bloco de referências bibliográficas no final do documento. Este está identificado por uma expressão "Referências bibliográficas" ou equivalente. O bloco em questão possui depois uma lista de referências bibliográficas citadas ao longo do documento. Esta lista, ou pelo menos parte dela, pode ser relevante para a pesquisa, donde esta informação é também extraída. Numa fase posterior poderá ser filtrada para excluir candidatos menos prováveis.

Assim, usando como exemplo o mesmo documento citado anteriormente, desta vez o seu fim, podemos ver na figura 4.11 um exemplo do texto a analisar.

Para obter a informação bibliográfica do bloco de texto com as referências bibliográficas é necessário determinar o separador ou o identificador das



Acknowledgment

The comments of Shian-Hua Lin and Ming-Jing Hwang2 during the preparation of this paper were very helpful. We thank for their help.

References

[1] C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence, "CiteSeer: An Automatic Citation Indexing System," Digital Libraries 98 Pittsburgh PA USA.

[2] Kurt D. Bollacker, Steve Lawrence, and C. Lee Giles, "CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications," 2nd International ACM Conference on Autonomous Agents, pp. 116-123, ACM Press, May, 1998.

[3] Steve Lawrence, C. Lee Giles and Kurt D. Bollacker, "Autonomous Citation Matching," Proceedings of the Third International Conference on Autonomous Agents, Seattle, Washington, May 1-5, ACM Press, New York, NY, 1999.

Figura 4.11: Exemplo de um bloco de texto extraído do fim de um documento PDF

referências bibliográficas. Devido às limitações das aplicações para extracção do texto, a quebra de linha não é necessariamente um separador. Pela mesma razão, é possível que uma referência se encontre partida por uma quebra de linha. Assim, torna-se necessário determinar o separador ou pelo menos o tipo de identificador da citação que, quando conhecido, pode tomar o lugar do separador se necessário. São usadas as seguintes heurísticas para obter esta informação:

1) Determinar a sequência de caracteres que segue imediatamente a expressão "References"(ou similar);

2) Procurar índices comuns como expressões dentro de parênteses rectos ou parênteses curvos;

3) Encontrar citações no texto que coincidam e usá-las para determinar o início de cada referência bibliográfica.

No exemplo dado em 4.11, o identificador é facilmente reconhecido e a informação obtida será algo semelhante ao apresentado na figura 4.12.

Este tipo de estratégia pode ser aplicado a documentos académicos, relatórios,

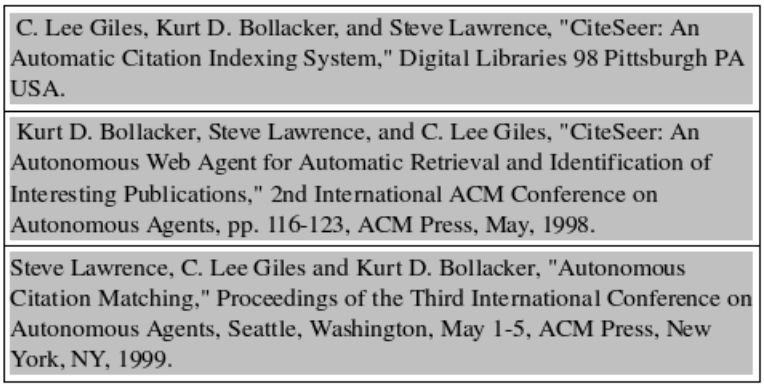


Figura 4.12: Exemplo de informação obtida do exemplo 4.11

apresentações ou outro tipo de documento onde seja possível identificar um bloco de texto com referências e que estas possuam um identificador.

4.2.3.4 Extracção de informação de texto em geral, usando heurísticas

É possível encontrar colecções de informação bibliográfica em documentos. É frequente encontrar documentos com listas de publicações de uma organização, de um domínio ou de um autor.

Procurando encontrar separadores ou identificadores comuns que identifiquem listas é possível, usando um conjunto de heurísticas simples, determinar um bloco de texto que é um potencial candidato a referência bibliográfica.

1) Marcas em documentos hipertexto como parágrafos ($< p >$) ou listas ($< li >$);

2) Marcas identificadoras de listas em texto no início da linha como o asterisco ($*$) ou hífen ($-$). A aplicação responsável pela conversão de HTML substitui listas por estes caracteres.

3) Identificadores comuns de referências no início da linha, como [1] ou *(Lawrence et al 1999)*;

4) Blocos de texto que contenham expressões que existam na expressão dada como argumento.

Esta estratégia tem demonstrado ter uma precisão baixa mas com uma abrangência alta, produzindo um número bastante alto de candidatos. Para melhorar a precisão, são aplicados filtros para reduzir a lista de candidatos inicial a uma lista de candidatos mais provável.

4.2.3.5 Outros métodos não abordados

Outras abordagens foram consideradas mas não foram aplicadas, como o uso de *wrappers* (secção 3.2.1.1) que, através do reconhecimento de um padrão, é capaz de extrair informação de documentos estruturados (Ashish e Knoblock, 1997). No entanto para reconhecerem cada padrão, necessitam de ser treinados individualmente para cada página, sítio ou então para páginas semelhantes. Considerando que se pretende que o SUPeRB recolha informação da Web em geral, os *wrappers* seriam mais eficientes se fossem aplicados a um conjunto restrito de sítios.

4.2.4 Extracção de elementos bibliográficos

A extracção de elementos bibliográficos consiste em distinguir e identificar um elemento bibliográfico de uma referência bibliográfica. Como apresentado em 3.2.2, existem vários métodos para executar esta tarefa. No caso do SUPeRB, foram adoptadas duas abordagens distintas:

- Usar a ferramenta ParaTools,
- Criar uma aplicação própria, baseada em identificar elementos bibliográficos através de repositórios de entidades com nomes.

Cada uma destas abordagens tem diferentes vantagens. O ParaTools, como já foi referido, também recorre a duas abordagens diferentes.

Por outro lado, o uso de um repositório de entidades com nome permite identificar pessoas, locais e conferências que façam parte do repositório. Para esta tarefa, são usados dois repositórios:

1) O REPENTINO, com cerca de 450.000 exemplos de entidades com nome. Tendo sido inicialmente construído para assistir na tarefa de reconheci-

mento de entidades mencionadas, possui 111 sub-categorias nas quais os exemplos são classificados. Apesar de bastante genérico, o REPENTINO possui diversas categorias que podem ser importantes no âmbito do SUPeRB e dos elementos bibliográficos, tais como SER::Pessoa, ORGANIZACAO, LOCAL que podem ser mapeadas para AUTOR, ou EDITOR, AFILIACAO, INSTITUICAO ou LOCAL para o SUPeRB, permitindo determinar ou pelo menos delimitar o tipo de elemento que se trata.

2) O REB (Repositório de Entidades Bibliográficas),que é um repositório semelhante ao REPENTINO mas apenas com elementos bibliográficos, obtidos a partir das referências bibliográficas do catálogo da Linguateca. Como foi construído apenas a partir de referências bibliográficas com os elementos bibliográficos devidamente identificados, possui apenas categorias no âmbito bibliográfico. Contém categorias como Autores, Editores, Conferências, Abreviaturas de conferências, revistas, editoras, Locais, etc. Categorias compostas apenas por elementos numéricos ou que possuem uma estrutura específica facilmente identificada recorrendo a heurísticas, não foram inseridas. Adicionalmente este recurso pode ser melhorado com a inserção de novos itens que sejam encontrados pelo SUPeRB e validados pelo utilizador.

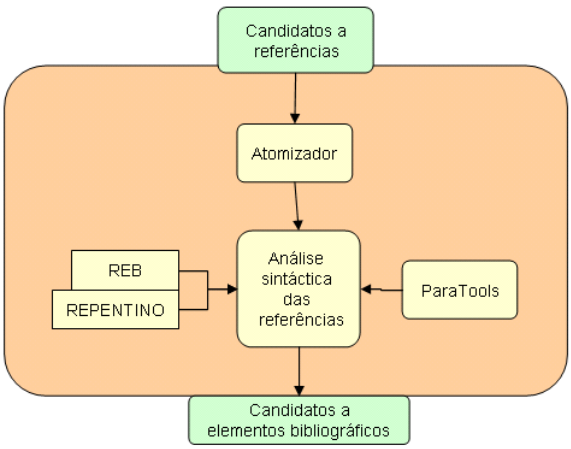


Figura 4.13: Tarefa de extracção dos elementos bibliográficos

Mas antes de se tentar classificar os respectivos elementos bibliográficos, é necessário ainda extrair primeiro os elementos bibliográficos. Para extrair os elementos bibliográficos, é necessário determinar os separadores correctos. Assim, numa primeira fase, a referência bibliográfica é quebrada, como apresentado na tabela 4.5, linha 2. Esta separação é feita, mantendo a precedência dos separadores, por ordem de maior precedência: parentêses, ponto e vírgula, vírgula, ponto, e dois pontos. De seguida, tenta-se determinar

Tabela 4.5: Fases para extracção e identificação de elementos bibliográficos

1	Steve Lawrence, C. Lee Giles and Kurt D. Bollacker, "Autonomous Citation Matching,"Proceedings of the Third International Conference on Autonomous Agents, Seattle, Washington, May 1-5, ACM Press, New York, NY, 1999.
2	Steve Lawrence, C. Lee Giles and Kurt D. Bollacker, "Autonomous Citation Matching,"Proceedings of the Third International Conference on Autonomous Agents, Seattle, Washington, May 1-5,ACM Press, New York,NY,1999.
3	NOME, ?. NOME, NOME. NOME, "TITULO", CONFERENCIA, LOCAL, LOCAL, DATA, EDITORA, LOCAL, LOCAL, DATA.
4	NOME, NOME and NOME, "TITULO", CONFERENCIA, LOCAL, DATA, EDITORA, LOCAL, DATA.

o tipo de elemento bibliográfico recorrendo quer a repositórios de exemplos, quer a expressões regulares, para verificar casos como datas, páginas, volumes, etc. Este passo pode ser visto na tabela 4.5, da linha 2 para a linha 3. O nome dos autores, ou o nome da conferência, pode ser obtido, recorrendo aos repositórios, que identificarão *Steve*, *Lawrence* e *Giles* como nomes de pessoas, e *Seattle*, *Washington* ou *New York* como locais. Expressões regulares permitem deduzir que a expressão entre parênteses é um título, o título da conferência é identificado por possuir a palavra *proceedings* e as datas são também identificadas através de expressões regulares.

Por último, aplicam-se regras que permitem unir ou separar vizinhos com o mesmo tipo de estrutura, podendo se utilizar regras específicas em determinados casos. No exemplo, os tipos NOME que são vizinhos são reorganizados; são determinados como separadores a vírgula e o *and*, concatenando algumas expressões e separando outras. Isto porque se identifica a vírgula como identificador, considerando o ponto como um caracter não separador. Noutras

situações, os vizinhos são simplesmente concatenados, como é o caso de campos adjacentes identificados como LOCAL.

4.2.5 Fusão da informação bibliográfica

A tarefa de fusão da informação bibliográfica tem como objectivo, dado um conjunto de referências bibliográficas, com os elementos bibliográficos devidamente estruturados, identificar as referências bibliográficas que se referem ao mesmo documento, e tentar concatenar os diferentes elementos bibliográficos numa única referência.

Esta não é uma tarefa simples. Pretende-se não só desambiguar os dados, mas qualificar a qualidade e a relevância tendo em conta as fontes, as semelhanças, e a redundância dos dados bibliográficos obtidos.

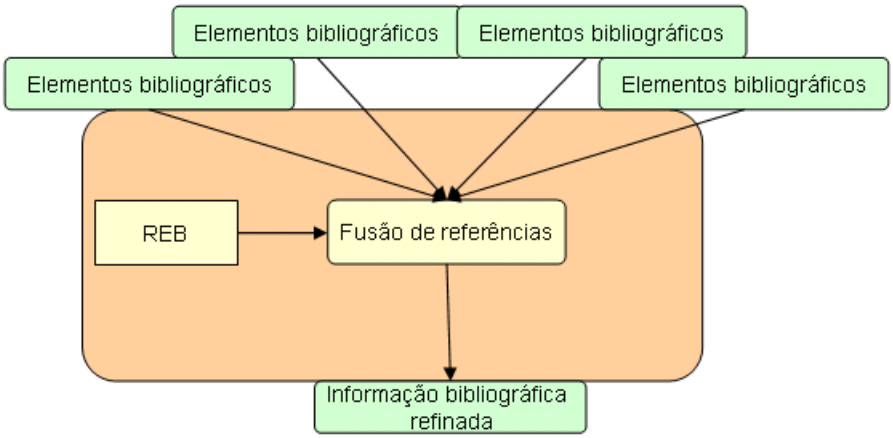


Figura 4.14: Fusão da informação bibliográfica a partir das diferentes fontes

4.2.5.1 Desambiguação dos elementos bibliográficos

O processo de desambiguação dos elementos bibliográficos consiste em simplificar através de remoção de acentos e da tentativa de expandir nomes de autores. Este tratamento dos dados permite comparar os diversos candidatos e agrupar os mais semelhantes, prevendo erros causados pela omissão voluntária

ou involuntária de quem criou o documento original ou erros causados pelos métodos de extracção de texto.

Para efectuar a desambiguação de nomes, recorre-se ao REB uma vez mais. A comparação entre os dados obtidos e os dados no catálogo é efectuada de forma idêntica à descrita por Feitelson (2000), através de tentativa de expansão de iniciais e remoção de acentos.

Note-se que este processo de desambiguação, removendo acentos ou expandindo iniciais, tem como único objectivo poder proceder a uma comparação entre vários elementos. Não se pretende alterar o conteúdo dos elementos descobertos. Ou seja, o nome "*J.J. Almeida*" não é substituível por "*José João Almeida*" a não ser que uma das referências a ser concatenada contenha essa forma. Os principais elementos passíveis de tentar fundir várias referências são:

1) Título, verificando se é possível que a mesma referência tenha sido encontrada várias vezes ou já exista no catálogo;

2) Autor, para manter informação adicional sobre os autores;

3) Conferência, podendo obter informação mais completa quer do repositório quer de outras referências (tal como o nome completo ou abreviatura da conferência, os editores, a editora, a data ou o local)

Para permitir a fusão entre diversas referências, são considerados alguns limites, como terem pelo menos um autor em comum, o título ser idêntico e terem a mesma data.

4.2.5.2 Qualidade da informação

Após a desambiguação de elementos bibliográficos que permite agrupar as diversas referências numa única referência, é necessário escolher quais os elementos bibliográficos que devem fazer parte da referência gerada. De cada grupo de referências bibliográficas, apenas uma referência é produzida. Mas não basta concatenar as referências. Entre os elementos ambíguos, é necessário escolher o correcto, ou mais indicado. A figura 4.15 apresenta um exemplo onde é possível observar duas referências com o **Autor**, o **Título** e a **Conferência** em comum.

Figura 4.15: Exemplo de fusão de duas referências que se referem à mesma publicação

4.2.6 Classificação da informação bibliográfica

A fase de classificação da informação possibilita a classificação da informação bibliográfica encontrada, quer automaticamente, quer através da marcação livre do utilizador.

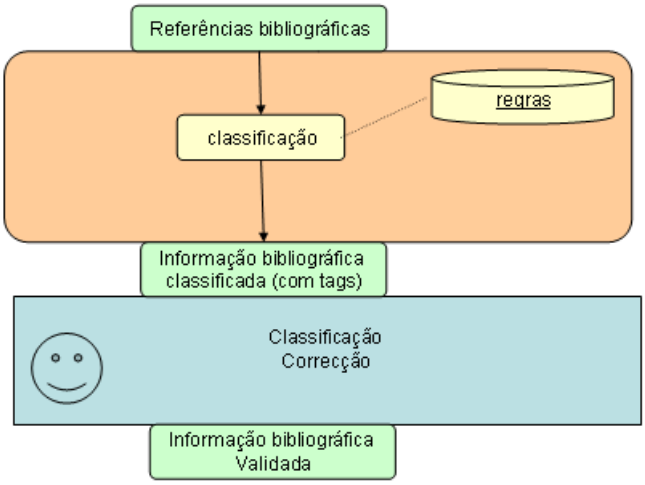


Figura 4.16: Classificação da informação

Este processo vem facilitar e oferecer novos meios de organização do catálogo proporcionando agrupamentos ou pesquisas com base na classificação atribuída.

4.2.6.1 A classificação manual

A classificação manual pelo utilizador é um processo simples, que pode ser executado facilmente através de uma interface Web apropriada. Consiste no acto de atribuir marcas a cada referência bibliográfica, ou a um grupo de referências bibliográficas. O utilizador tem a liberdade para escolher as marcas que pretende atribuir a cada referência, sem qualquer restrição de vocabulário, com a excepção do tamanho, devendo conter pelo menos 4 caracteres.

É ainda possível facilitar um método de sugerir ao utilizador marcas já existentes, através de um menu *popup*. Este método permitiria reduzir a variedade das marcas, dando a conhecer ao utilizador marcas já existentes, e que podem ser semelhantes às que o utilizador pretende atribuir.

O processo de classificação, ou *tagging* como é geralmente conhecido actualmente, é um processo rápido que, como se pode observar em outros sistemas Web como o del.ici.ous[1] ou o bibsomony[2], o utilizador tem facilidade em colaborar.

No SUPeRB, foi testada a classificação em conjunto com a interface de pesquisa de publicações no catálogo da Linguateca, permitindo a pesquisa e inserção de novas marcas nas referências apresentadas.

4.2.6.2 A classificação automática

A classificação do conteúdo de forma automática poderá ter como objectivo proporcionar uma ferramenta que possa facilmente identificar o tipo de informação bibliográfica em questão de acordo com as preferências dos utilizadores. Isto é feito recorrendo a um conjunto de regras simples, pré-definidas e que possam ser facilmente utilizadas, com os dados disponíveis e a informação relevante recolhida, tal como a estrutura do documento, a fonte do documento, o texto, os autores, o título, o resumo, nome da conferência ou da revista. Os resultados

[1] `http://del.icio.us`
[2] `http://www.bibsinomy.org`

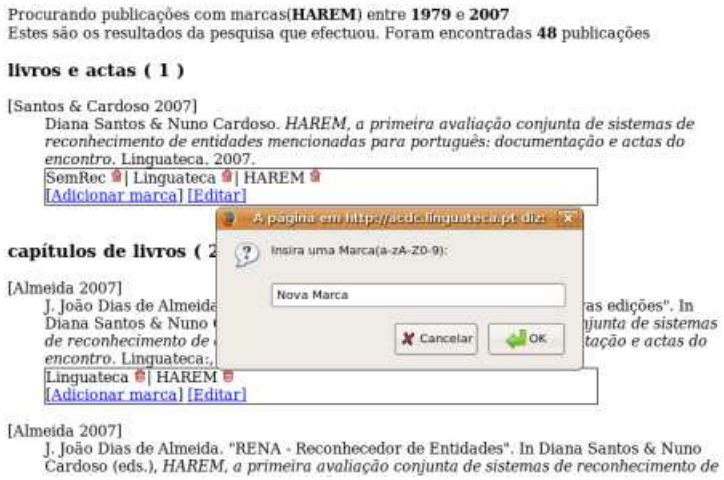


Figura 4.17: Classificação da informação, em pesquisa de publicações

desta classificação são normalmente intrínsecos a uma classificação interna do repositório, pelo que não têm que ser necessariamente apresentados ao utilizador que está a inserir a informação bibliográfica.

4.3 Interface Web do SUPeRB

O objectivo do SUPeRB é permitir que qualquer pessoa seja capaz de efectuar tarefas de pesquisa de referências bibliográficas através da Web, fornecendo ao utilizador as ferramentas para processar referências bibliográficas e de interagir com o catálogo da Linguateca, nomeadamente permitindo a inserção ou actualização de informação bibliográfica.

É portanto importante que o SUPeRB possua uma interface fácil. Assim, com vista a aumentar a usabilidade da interface do SUPeRB, esta tem vindo a ser desenvolvida com a tecnologia Ajax. O uso de Ajax possibilita uma interactividade maior, capaz de efectuar acções sem ter que recarregar as páginas. Para o SUPeRB isto representa uma vantagem, dado que as tarefas levadas a cabo podem ter um longo tempo de espera, sendo no entanto possível obterem-se resultados antes do fim da execução da tarefa. Da mesma forma é possível iniciar outras tarefas antes do final da execução de uma tarefa.

A tarefa de validação, por exemplo, em que o utilizador interage com o

SUPERB em várias fases é um momento onde este tipo de interacção representa uma forma de melhorar a eficiência e a usabilidade das tarefas. Recorrendo ao Ajax, o utilizador é capaz de aceder à informação calculada antes do final da execução de uma tarefa. É possível ao utilizador interagir antecipadamente com o SUPeRB, editando ou validando a informação recolhida. Em tarefas que apresentem muitos resultados, o utilizador pode efectuar uma parte da tarefa de validação antes do final da execução.

Figura 4.18: Apresentação dos resultados dos URL processados no módulo de extracção de texto

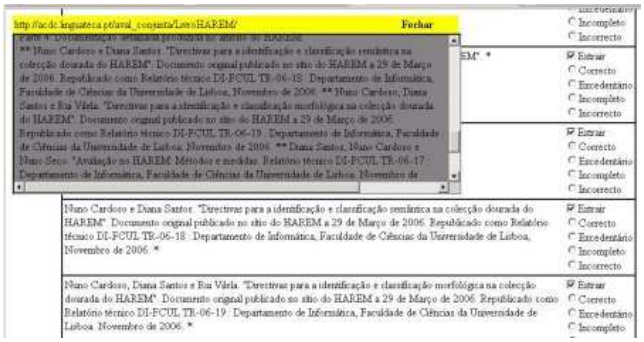


Figura 4.19: Apresentação dos resultados obtidos no módulo de extracção de referencias a partir de texto

O Ajax é uma tecnologia que oferece interfaces realmente dinâmicas, sendo possível alterar o conteúdo de parte da interface em tempo real e submeter informação ao servidor sem que isso afecte o resto da interface. No SUPeRB é possível editar uma referência dada, e submeter as alterações sem que isso afecte o resto das referências.

Assim, cada funcionalidade do SUPeRB possui uma interface capaz de

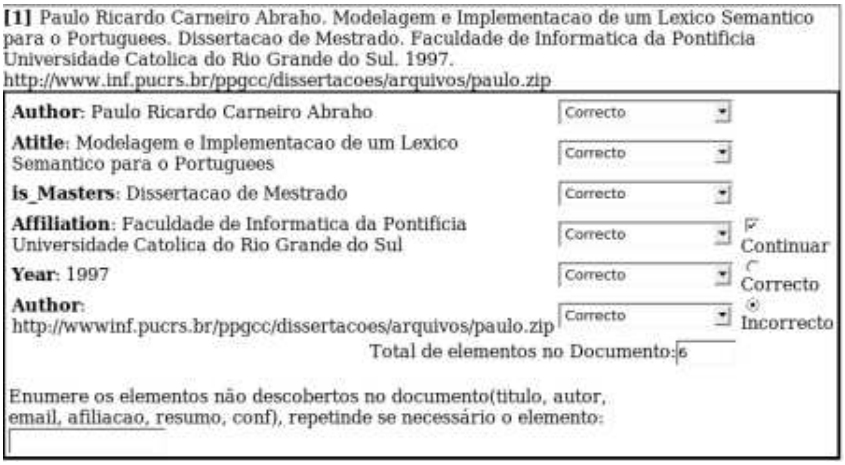


Figura 4.20: Apresentação dos resultados obtidos a partir do módulo de extracção de elementos bibliográficos

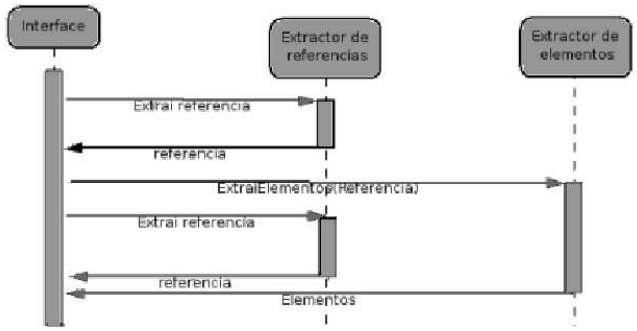


Figura 4.21: Exemplo de pedidos entre a interface usando Ajax

comunicar com o servidor e fazer pedidos, por exemplo pedir para extrair os elementos de uma referência. É possível ainda executar várias acções em simultâneo. Por exemplo ao processar vários documentos, em que estejam a ser extraídas referências, é possível iniciar a extracção de elementos bibliográficos de uma ou mais referências já extraídas e apresentadas ao utilizador, ainda que não se tenha terminado a extracção de referências bibliográficas de todos os documentos. Da mesma forma, pode ser possível corrigir a informação de uma referência em particular sem afectar o estado da interface e da restante informação bibliográfica.

4.4 Interacção com o SUPeRB

O utilizador pode interagir com o SUPERB através da Web mas esta interacção pode processar-se de diversas formas. Por um lado, temos todas as funcionalidades do SUPeRB que podem ser levadas a cabo separadamente. Por outro, pretende-se que seja possível executar todas as tarefas sequencialmente de forma a que, partindo de uma informação limitada, se obtenha os dados bibliográficos relevantes numa forma estruturada que seja possível de inserir no catálogo ou que seja a representação da informação que o utilizador esperava.

Mas a interacção com o SUPeRB não é limitada à introdução de expressões ou URL pelos utilizadores. É possível utilizar o SUPeRB para verificar informação recolhida em pesquisas anteriores ou para verificar e actualizar informação bibliográfica do catálogo.

4.4.1 Por omissão

O fluxo normal do SUPeRB consiste numa interacção simples onde o utilizador fornece informação na forma de expressão ou de uma referência. A informação inserida é utilizada para recolher informação da Web e usada posteriormente para processar e filtrar a informação obtida. Este processo complexo utiliza várias componentes, desde a pesquisa, extracção de texto da Web, extracção de referências, extracção de elementos bibliográficos, a fusão e por fim a classificação. Todos estas tarefas são executadas sequencialmente, uma vez que os parâmetros de cada componente dependem do processos anteriores.

Opcionalmente pode ser possível validar os resultados de cada componente antes de submetidos à tarefa seguinte.

4.4.2 Em ciclo

Obtidos os elementos bibliográficos, estes podem ser sugeridas ao catálogo de publicações. Outra forma de utilizar os dados consiste em utilizar a informação recolhida para obter novos dados. As referências extraídas devem servir para obter nova informação bibliográfica. A aplicabilidade deste método pressupõe

que o utilizador pretende:

1) obter novos resultados que completem a informação;

2) alargar a pesquisa, procurando novos resultados;

3) ou obter novos resultados que sejam comuns a diversas das referências obtidas.

Por exemplo, pode-se prever um caso em que são seleccionadas três referências em que os resultados obtidos sejam comuns a todas as situações apresentadas, tal como um artigo que cite todas as referências.

4.4.3 Interacção com algumas componente específicas

Como mencionado, é possível interagir individualmente com alguns dos componentes do SUPeRB através de interfaces Web. Aqui descreve-se a interacção com alguns dessas componentes, nomeadamente a extracção de referências e a extracção de elementos bibliográficos.

4.4.3.1 Interacção com a componente de extracção de referências

O utilizador pode indicar ao sistema textos que contenham referências bibliográficas que se pretendam obter. O utilizador pode introduzir um ou mais URL para os documentos que queira processar através de uma caixa de texto (ver figura 4.22).

O sistema recorre automaticamente a outro módulo para extrair texto dos documentos e de seguida procede à sua análise. Alternativamente, é ainda possível inserir um texto directamente numa caixa de texto.

Durante a fase de análise, os resultados vão sendo apresentados ao utilizador à medida que cada documento é processado, permitindo ao utilizador validar ou editar os dados bibliográficos (figura 4.23). É ainda possível ver alguma informação a respeito do documento de onde a informação foi extraída, sendo possível ver o documento original ou o texto extraído.

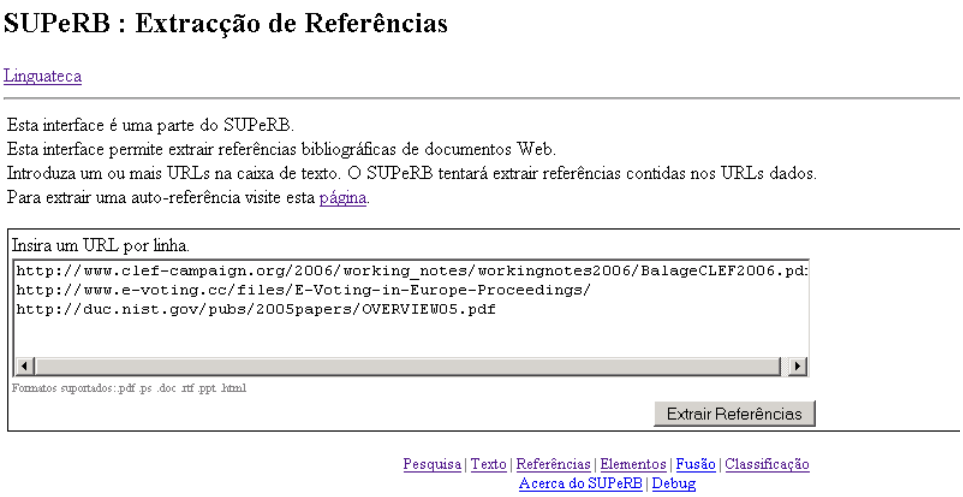


Figura 4.22: Introdução de URL para extrair referências

Figura 4.23: Resultados apresentados da extracção de referências

Os dados obtidos pelo utilizador podem depois ser utilizados directamente pelo utilizador, ou por outras componentes para processar a nova informação. Podem, por exemplo, ser usados directamente pelo extractor de elementos bibliográficos de forma a serem mais tarde incorporados no catálogo.

4.4.3.2 Interacção com a componente de extracção de elementos bibliográficos

A interface desta componente é semelhante à anterior, mas os parâmetros são diferentes. Este módulo extrai elementos bibliográficos a partir de referências bibliográficas. A interface desta componente (Figura 4.24) consiste numa área de texto que pode receber uma ou mais referências. Estas têm que estar devidamente delimitadas, devendo ser colocada apenas uma referência por

linha. Caso exista mais do que uma referência por linha, será considerada como uma só referência.

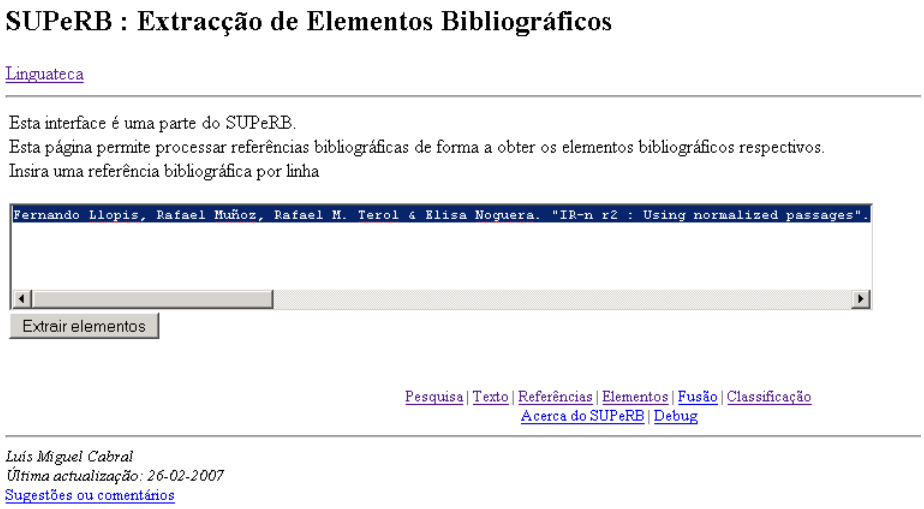


Figura 4.24: Interface de entrada de referências

As referências são depois processadas uma a uma, exibindo os resultados ao utilizador, que pode comparar com o texto original, tal como ilustra a figura 4.25. É ainda possível editar e corrigir os elemento bibliográficos obtidos, sendo possível:

- editar o texto;
- alterar o tipo de elemento, corrigindo por exemplo autor para editor;
- remover elementos;
- adicionar elementos.

É possível ainda utilizar os resultados obtidos para serem utilizados por outras componentes, sendo possível, por exemplo, guardar os dados ou converter os dados para um formato comum, como o BibTeX.

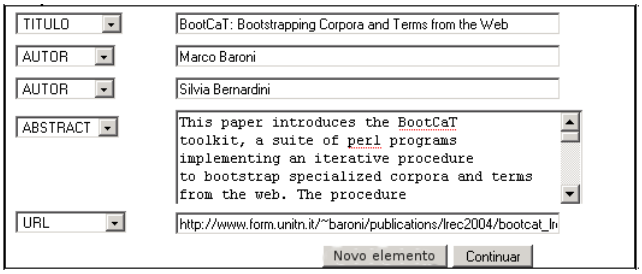


Figura 4.25: Exemplo de uma interface de validação, que permite a edição de elementos

Capítulo 5

Avaliação do SUPeRB

Nota: A primeira versão deste capítulo foi elaborada em colaboração com a Diana Santos e com o Luís Sarmento.

A modularização do sistema com vista a poder invocar individualmente diversas funcionalidades permite avaliar independentemente cada componente, além de poder avaliar a eficiência do sistema completo. Neste capítulo apresenta-se uma metodologia para avaliar em pormenor alguns módulos que compõem o SUPERB, nomeadamente os módulos de:

- extracção de referências a partir de texto;
- extracção de auto-referências;
- extracção de elementos bibliográficos a partir das referências.

Obviamente, existem outros módulos que deverão ser testados, no entanto, este capítulo pretende mostrar o problema em avaliar um sistema desta complexidade, em vez de descrever exactamente toda a avaliação necessária. Os módulos considerados nesta avaliação são pontos fulcrais ao longo de todo o processo e deles depende crucialmente o desempenho do sistema global.

Como será interessante medir a evolução do SUPeRB em vários momentos, além dos resultados de avaliação apresentados, detalha-se a metodologia de criação de materiais de teste para ser possível replicá-la mais tarde.

5.1 Diferença entre validação e avaliação

É importante esclarecer que a avaliação, embora seja semelhante em espírito à validação, foi concebida para avaliar o sistema, enquanto que a validação é para ser parte integrante do SUPeRB para um utilizador qualquer. Assim, embora mais tarde os resultado da validação sejam passíveis de incorporar noutro tipo de avaliação a acompanhar o sistema, as interfaces são distintas e os seus objectivos (e os seus utilizadores) diferentes.

Assim, uma interface de validação tem como objectivo permitir o menor esforço ao utilizador para usar os resultados automáticos do SUPeRB, assim como permitir facilmente a continuação do trabalho (e o deitar fora de sugestões).

Por outro lado, a interface de avaliação pretende medir rigorosamente a qualidade dos resultados do SUPeRB, mesmo que implique bastante trabalho de classificação de coisas que não seriam utilizáveis num fluxo normal (e que fariam com que um utilizador abandonasse naturalmente aquela proposta).

Algumas das interfaces de validação foram apresentadas no capítulo anterior. Aqui apresentam-se as de avaliação.

5.2 Avaliação do módulo de extracção de referências bibliográficas a partir de listas

Como entrada deste módulo, é dado um texto qualquer, sob a forma de URL, como foi exemplificado na secção 4.4.3.2. Este módulo terá de extrair todas as referências bibliográficas nele presentes.

Uma referência bibliográfica considera-se correctamente extraída se o excerto de texto que a contém contiver todos os elementos bibliográficos presentes no texto e apenas esses, independentemente do estilo bibliográfico que é usado no documento. Exemplos de referências correctamente extraídas são apresentados na figura 5.1.

5.2.1 Como avaliar?

Como referido na secção 3.2.2, o excerto de texto pode encontrar-se partido por quebras de linha ou por hifenização. Deverá, contudo, apresentar toda a informação que permita a sua decomposição posterior nos elementos bibliográficos que o constituem, e apenas essa informação. Ou seja, caso o excerto de texto correspondente apresente informação em excesso ou em falta por incorrecta delimitação da referência, considera-se que a referência se encontra incorrectamente extraída. Três situações com incorrecções podem ocorrer:

1) **Erro**: o excerto de texto apresentado não apresenta qualquer informação que permita a extracção dos elementos bibliográficos, pelo que é completamente inútil para propósitos de extracção de referências, tal como é exemplificado na figura 5.2

2) **Informação excedentária**: o excerto de texto extraído apresenta mais informação para além da correspondente à referência bibliográfica, quer por inclusão de informação bibliográfica de referências adjacentes, quer por inclusão de texto avulso. É, no entanto, possível encontrar uma referência completa no excerto de texto em causa. Exemplos de referências com informação excedentária são apresentados na figura 5.3. Não se considera como informação excedentária caracteres isolados que não causem ambiguidade. Por exemplo, o "*[13]*" em "*[13] Rohini Srihari*

David McDonald. "Internal and External Evidence in the Identification and Semantic Categorization of Proper Names". Corpus processing for lexical acquisition, págs. 21-39, 1996. 90

D. A. Smith and G. S. Mann. Bootstrapping toponym classifiers. In Proceedings of the HTL/NAACL Workshop on The Analysis of Geographic References, 2003.

Hofmann, T. Probabilistic latent semantic indexing. In Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval (1999).

Figura 5.1: Exemplo de referências correctamente extraídas

... " não causa ambiguidade, pelo que não é considerado informação excedentária.

3) **Informação incompleta**: o excerto de texto não contém toda a informação bibliográfica disponível no texto original. A figura 5.4 exemplifica algumas situações do género. Na linha superior está a referência como se encontra disponível no texto original. Na linha seguinte é apresentado o texto capturado.

Este trabalho foi parcialmente financiado pela Fundação para a Ciência e Tecnologia, através do projecto com referência POSI/SRI/40193/2001 e da bolsa de doutoramento com referência SFRH/BD/10757/2002.
AVALON: Encontro de avaliação conjunta em sistemas de processamento computacional do ortuguês, organizado pela Linguateca.
Examples of such systems include LCC([7]), QuASM, IONAUT([1]), START([11]) and Webclopedia([10]).

Figura 5.2: Exemplo de erros na extracção de referências

Referências 1. E. Amitay, N. Ha'El, R. Sivan, and A. Soffer. Web-a-Where: Geotagging Web content. In Proceedings of SIGIR-04, the 27th conference on research and development in information retrieval, 2004.underline{ 2. N. Cardoso}
[13] Rohini Srihari and Wei Li. Information Extraction Supported Question Answering. Eighth Text REtrieval Conference (TREC-8). Gaithersburg, MD. November 17-19, 1999. Carlton, John
XU, J . AND CROF T, W. B. 2000. Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems (TOIS) 18, 1, 79-112. ACM Transactions on Internet Technology, Vol. V, No. N, Month 2002.

Figura 5.3: Exemplo de referências com informação excedentária (a vermelho a informação excedentária)

4) **Omisso**: a referência bibliográfica não foi extraída do texto original.

5.2.2 Medidas de desempenho

As medidas de desempenho propostas são as usuais em RI (veja-se Santos e Cardoso (2007); Santos et al. (2006b,a)) - precisão, abrangência, medida F, sub-geração e sobre-geração.

1) A precisão é dada pela fórmula

$$Precisao_{referencias} = \frac{\#ReferenciasCorrectas}{\#ReferenciasIdentificadas} \tag{5.1}$$

onde $ReferenciasCorrectas$ é o número de referências correctamente identificadas pelo SUPeRB e $ReferenciasIdentificadas$ o número total de referências identificadas pelo sistema.

2) A abrangência define-se como

$$Abrangencia_{referencias} = \frac{\#ReferenciasCorrectas}{\#ReferenciasDocumento} \tag{5.2}$$

Onde $ReferenciasDocumento$ é o número total de referências que

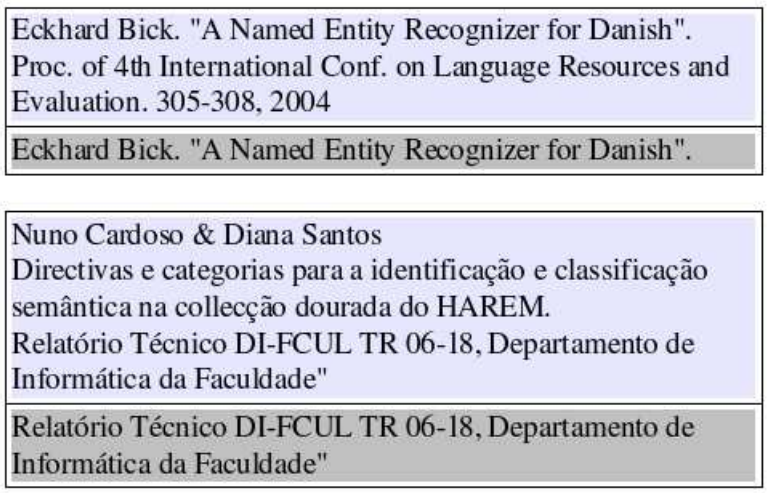


Figura 5.4: Exemplo de referências com informação incompleta (a claro o texto original, num tom mais escuro o texto da referência delimitada)

existem no documento ou na colecção de documentos considerados

3) A Medida F é a média harmónica da precisão e da abrangência

$$MedidaF_{referencias} = 2.\frac{Precisao.Abrangencia}{Precisao + Abrangencia} \tag{5.3}$$

4) A precisão alargada considera também as referências com informação excedentária

$$PrecisaoAlargada_{referencias} = \frac{\#ReferenciasCorrectas + \#ReferenciasExcedentarias}{\#ReferenciasIdentificadas} \tag{5.4}$$

Onde $ReferenciasExcedentes$ é o número de referências com informação excedentária.

5) A abrangência alargada considera também as referências com informação excedentária

$$AbrangenciaAlargada_{referencias} = \frac{\#ReferenciasCorrectas + \#ReferenciasExcedentarias}{\#ReferenciasDocumento} \tag{5.5}$$

6) A sub-geração mede a informação incompleta e omissa

$$Sub{-}geracao_{referencias} = \frac{\#ReferenciasIncompletas + \#ReferenciasOmissas}{\#ReferenciasDocumento} \tag{5.6}$$

onde $ReferenciasIncompletas$ é o número de referências incompletas e $ReferenciasOmissas$ o número de referências não encontradas..

7) A sobre-geração quantifica as referências erradas

$$Sobre - geracao_{referencias} = \frac{\#ReferenciasErradas}{\#ReferenciasIdentificadas} \tag{5.7}$$

5.2.3 Materiais de teste

A primeira decisão a tomar refere-se ao ponto de partida a usar para a operação de extracção. A este nível existem duas opções:

1) Utilizar como informação de entrada documentos em vários formatos (por exemplo, `PDF`, `RTF` e `HTML`) contendo várias referências. Cabe ao módulo de análise do URL e obtenção de conteúdos (secção 4.2.2) obter o texto a ser processado a partir de URL de documentos.

2) Utilizar o conteúdo de texto já devidamente extraído dos documentos.

A segunda decisão prende-se com a distribuição dos géneros de documentos a serem testados. A divisão dos documentos em partes iguais pelos dois géneros mais significativos para esta tarefa – documentos académicos, por um lado, e listas de referências Web, por outro – parece adequada. Pode-se, no entanto, realizar uma divisão de segunda ordem entre vários tipos de documentos académicos, que poderão ser separados em artigos científicos, relatórios técnicos, dissertações, etc..

Assim, para testar este módulo sugeriu-se uma lista de URL, em que para cada documento seja extraído manualmente o número de referências nele contidas.

- Uma parte dos URL apontando para documentos Word ou PDF publicados em conferências da area do processamento computacional do português, como o PROPOR, o encontro da APL, o TIL, etc, com artigos em vários formatos e nas duas línguas consideradas no SUPeRB, o inglês e o português;
- Outra parte dos URL apontando para páginas web com listas de referências bibliográficas, como páginas pessoais e de instituições relevantes na área do processamento computacional da língua portuguesa e outras.

5.2.4 Exemplo de avaliação

A tabela 5.1 apresenta uma lista de URL, seguindo a metodologia descrita na secção anterior.

A tabela 5.2 apresenta a classificação detalhada para a avaliação da extracção dos resultados sobre os dados da tabela 5.1. Não foram considerados, no cálculo das medidas, os casos em que o sistema não foi capaz de extrair o texto dos URL apresentados (a). A tabela 5.3 apresenta os resultados das medidas referentes à classificação dada.

Tabela 5.1: URL e número de referências de cada um, avaliados para a extracção de referências; o primeiro grupo (1-10) contém páginas com listas de referências; o segundo grupo (11-21) refere-se a documentos.

ID	URL	#Referências
1	`http://www.emeraldinsight.com/Insight/viewContentItem.do?contentType=Article&contentId=1520174`	*a*
2	`http://www.di.fc.ul.pt/sobre/\?reports\&entry\_type=M`	21
3	http://istpress.ist.utl.pt/	9
4	http://en.scientificcommons.org/k_bontcheva	*a*
5	http://en.scientificcommons.org/8865457	*a*
6	http://www.clef-campaign.org/clef-bibliography.pdf	193
7	http://www.pget.ufsc.br/curso/dissertacoes_defendidas.php	16
8	http://gate.ac.uk/projects/sekt/	13
9	http://www.linguateca.pt/Diana/interesses.html#cont	13
10	http://acdc.linguateca.pt/aval_conjunta/LivroHAREM/	23
11	http://www.inesc-id.pt/pt/indicadores/Ficheiros/3277.pdf	*a*
12	http://www.linguateca.pt/documentos/SantosESP2004.pdf	28
13	http://www.linguateca.pt/Diana/download/Geyetal2006.pdf	11
14	http://www.inesc-id.pt/pt/indicadores/Ficheiros/2900.pdf	*a*
15	https://repositorium.sdum.uminho.pt/bitstream/1822/4457/1/XATA06-0.11.pdf	*a*
16	http://www.di.uminho.pt/ jcr/XML/publicacoes/artigos/2005/RLH05-EML.pdf	12
17	ftp://ftp.ime.usp.br/pub/mfinger/2004/FingerWassermann-jlc2003final.pdf	*a*
18	http://acdc.linguateca.pt/LuisCabral/publicacoes/Proposta_SUPERB.pdf	*a*
19	http://centria.di.fct.unl.pt/ lmp/publications/online-papers/proc_APSD06.pdf	19
20	http://centria.di.fct.unl.pt/ lmp/publications/online-papers/Reformar_ES.pdf	*a*
21	http://rod.do.sapo.pt/Rod_Web/Publications_files/limalopes.pdf	73
Total		431

5.3 Avaliação do módulo de extracção de referências bibliográficas a partir do próprio documento

Considerámos que seria também interessante avaliar separadamente o caso da obtenção da auto-referência, ou seja, da referência que é possível obter a partir

Tabela 5.2: Classificação detalhada dos URL da tabela 5.2

ID	Encontradas	Errados	Incompletos	Excedentes	Omissos	Total
2	16	8	0	5	0	29
3	0	5	0	0	9	5
5	2	0	0	0	191	2
6	6	23	29	0	0	58
8	12	7	0	0	1	19
9	0	1	1	2	10	4
10	6	5	0	0	17	11
12	22	2	3	2	1	29
13	8	1	3	0	0	12
16	0	1	0	0	12	1
19	15	3	5	0	0	23
21	8	6	0	1	64	15
Total	95	62	41	10	305	208

Tabela 5.3: Cálculo das medidas de avaliação referentes à extracção de referências das tabelas anteriores

	Listas	Documentos académicos	Cálculo global
Precisão	0.328	0.663	0.457
Abrangência	0.146	0.371	0.220
Medida F	0.202	0.475	0.297
Precisão Alargada	0.383	0.7	0.505
Abrangência Alargada	0.170	0.391	0.244
Sub-Geração	0.909	0.633	0.802
Sobre-Geração	0.376	0.197	0.298

do próprio artigo que se quer catalogar.

Para este caso, o que faz sentido é verificar quais os campos que era possível reconhecer (se encontravam no objecto electrónico), tornando este tipo de tarefa semelhante em termos de resultados à tarefa de extracção de elementos bibliográficos (embora o processo de os encontrar seja totalmente diferente).

A figura 5.5 apresenta a interface de avaliação.

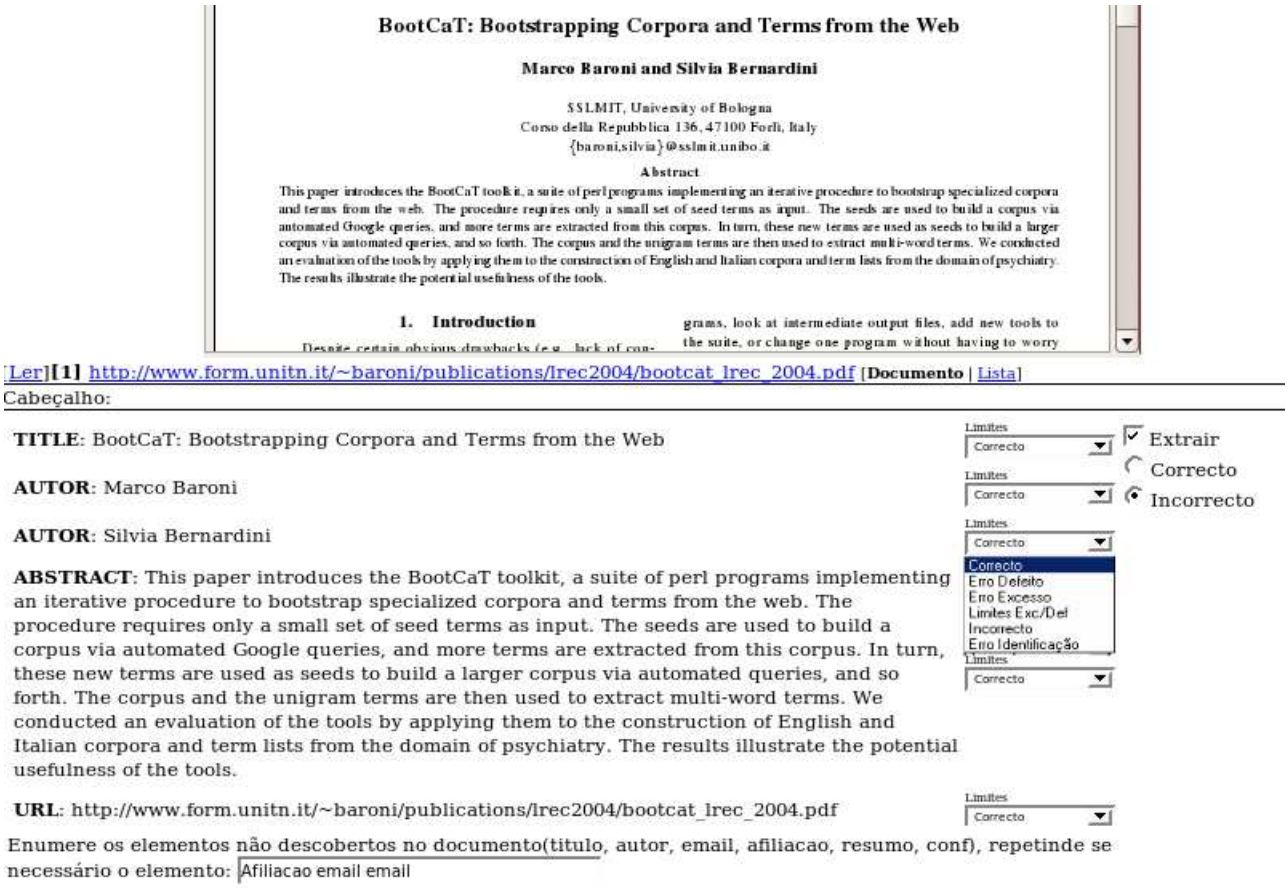


Figura 5.5: Exemplo de avaliação de uma auto-referência

5.3.1 Exemplo de avaliação

A tabela 5.4 apresenta a lista de URL com o número de elementos presentes.

A tabela 5.5 apresenta os resultados detalhados para cada URL, a tabela 5.6 apresenta as medidas obtidas para todos os elementos enquanto a tabela 5.7 representa a avaliação individual por elemento. Mais uma vez, não foram considerados, no cálculo das medidas, os casos em que o sistema não foi capaz de extrair o texto dos URL apresentados.

Tabela 5.4: URL avaliados para a extracção de auto-referências

ID	URL	# Referências
1	http://www.emeraldinsight.com/Insight/viewContentItem.do?contentType=Article&contentId=1520174	*a*
2	http://www.cs.cmu.edu/ acarlson/semisupervised/million-fact-aaai06.pdf	18
3	http://infolab.stanford.edu/pub/jannink/janthesis.pdf	8
4	http://www.alta.asn.au/events/altss_w2003_proc/altss/courses/molla/qa_roadmap.pdf	58
5	http://eprints.sics.se/55/01/registerReply.pdf	*a*
6	http://www.e-voting.cc/files/E-Voting-in-Europe-Proceedings/	*a*
7	0ttp://duc.nist.gov/pubs/2005papers/OVERVIEW05.pdf	6
8	http://www.cs.washington.edu/homes/mjc/papers/cafarella_databul06.pdf	10
9	http://arxiv.org/PS_cache/cmp-lg/pdf/9712/9712004.pdf	8
10	http://www.hpl.hp.com/personal/jjc/tmp/matching.pdf	*a*
11	http://www.cs.unt.edu/ rada/papers/mihalcea.cicling06a.pdf	11
12	http://acdc.linguateca.pt/LuisCabral/publicacoes/Proposta_SUPERB.pdf	5
13	http://www-db.stanford.edu/pub/gio/2001/westpoint-ieee3.htm	5
14	http://www.di.fc.ul.pt/tech-reports/06-07.pdf	12
Total		141

5.4 Avaliação do módulo de extracção de elementos bibliográficos

O objectivo do módulo de extracção de elementos bibliográficos consiste em identificar e separar correctamente todos os elementos existentes no interior de uma referência bibliográfica (que se assume correctamente extraída).

5.4.1 Como avaliar?

Vamos considerar que todos os tipos de elementos bibliográficos presentes numa referência bibliográfica devem ser extraídos, embora estes possam ser divididos em duas categorias:

Tabela 5.5: Resultados dos URL avaliados para a extracção de auto-referências

ID	Correcto	Errado	Incompleto	Excedente	Classificação	Omisso	Total
2	7	0	1	0	0	3	8
3	6	0	0	0	0	2	6
4	1	2	0	0	0	57	3
7	5	0	0	0	1	0	6
8	6	0	0	0	0	0	6
9	4	1	0	0	0	4	5
11	7	0	0	0	1	3	8
12	2	0	0	0	0	3	2
13	1	0	1	0	0	2	2
14	7	0	0	0	2	3	9
Total	46	3	2	0	4	88	55

Tabela 5.6: Cálculo dos resultados do URL avaliados para a extracção de auto-referências

Medida	**Valor**
Precisão total	0.836
Abrangência total	0.326
Precisão Alargada total	0.836
Abrangência Alargada total	0.326
Medida F total	0.469
Sub-Geração total	0,638
Sobre-Geração total	0.055

Tabela 5.7: Resultados da avaliação por elemento

	Precisão	Abrangência	Medida F
Autor	0,666	0.186	0.291
Título	0.57	0.8	0.667
Resumo	0.889	1	0.941
Email	1	0.333	0.5
Filiação	0	0	0

- Elementos obrigatórios: lista de autores, título da publicação, título do livro onde se encontra a publicação no caso de ser um artigo pertencente a actas ou semelhante, e ano de publicação.

- Elementos opcionais: lista de editores, informação acerca de volume,

número ou série da publicação, as páginas, o local de publicação, o mês de publicação, organização responsável pela edição (empresa editora), resumo, o URL, o DOI, etc..

Parece ser complicado quantificar a importância relativa da extracção dos elementos obrigatórios e dos elementos opcionais. Por um lado, os elementos obrigatórios são essenciais para a correcta identificação da publicação. Por outro lado, a pesquisa dos elementos opcionais é muitas vezes aquela que obriga a mais trabalho de pesquisa por parte do operador de manutenção de um catálogo, por serem elementos que estão frequentemente dispersos por várias fontes.

Por esse motivo, se é certo que os elementos obrigatórios têm de ser sempre correctamente identificados, também os elementos opcionais deverão ser correctamente extraídos, pois essa informação é valiosa e poupará muito trabalho a qualquer utilizador e ao operador humano responsável por validar a informação bibliográfica que é proposto para o catálogo, que é, no fundo, o principal objectivo do SUPeRB.

Assim sendo, propõe-se que a avaliação pondere igualmente todos os elementos existentes na referência a extrair.

Embora se pudesse seguir uma abordagem semelhante à do HAREM (Santos et al., 2006a) para o reconhecimento de entidades mencionadas em texto português, em que se separa a identificação (ou delimitação) pura e simples da classificação atribuída ao que foi delimitado, tal não faz grande sentido no âmbito de uma análise sintáctica de publicações onde a própria estrutura de cada campo é que leva à hipótese de identificação daquele campo e, portanto, qualquer que seja o método utilizado se está à procura de números para o número das páginas, datas para a data, etc. Assim a delimitação entra em conta com a classificação que pressupõe, e não faz sentido atribuir uma pontuação correcta a um editor que foi analisado como autor (mesmo que o nome esteja bem delimitado).

Ou seja, apenas vamos classificar como correcto se o nome do autor está classificado como autor. Se algum autor faltar, é marcado **Em Falta**, se alguém for considerado como AUTOR e não o é, considera-se como **Excedentário**, e

será medida a precisão e a abrangência do campo AUTOR seguindo o processo normal (idem para todos os elementos presentes na chave e no resultado do SUPeRB).

Cada elemento (exemplificando com AUTOR) pode ser classificado como:

1) **Correcto** Quando o elemento AUTOR foi correctamente extraído.

2) **Excedentário** Quando o elemento AUTOR apresentado contém informação excedente. (Nota: Não se considera informação redundante que permita a identificação do elemento como excedentário. Por exemplo, "*pp. 65-72*" e "*65-72*" são igualmente correctas uma vez que "*pp.*" é um identificador que permite a classificação correcta do elemento.)

3) **Incompleto** Quando ao elemento AUTOR apresentado faltem partes, por exemplo o valor de VOLUME conter *Volume no* em vez de *Volume no. 10.*

4) **Em falta** Quando existe um ou mais elementos marcados como AUTOR na chave que não aparecem como resultado do SUPeRB.

5) **Espúrio** Quando um elemento AUTOR não o é na referência (é talvez editor ou outra coisa qualquer)

Veja-se o seguinte exemplo na figura 5.6: O elemento *Rafael M* é considerado **Incompleto** enquanto que *Terol & Elisa Nogera* é considerado **Excedentário**. O elemento *15-17 September 2004* classifica-se com **Espúrio**. É ainda considerado um valor **Em Falta** pela editora *IST-CNR.*

Considerando a avaliação dos resultados na figura 5.6, obteríamos valores semelhantes à tabela 5.8.

5.4.2 Medidas de desempenho

As medidas globais de desempenho da extracção de elementos serão então a soma (para todos os elementos incluídos no resultado do sistema e na chave) destes valores. Exemplificando para a precisão, a precisão da extracção de

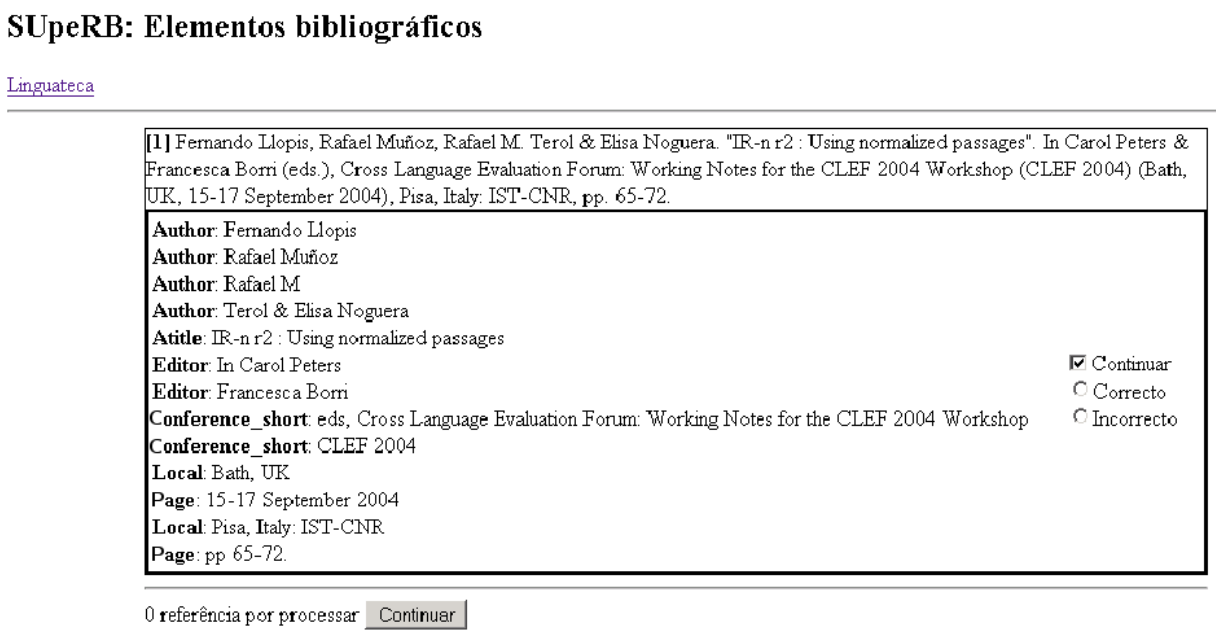


Figura 5.6: Exemplo de uma referência extraída

Tabela 5.8: Classificação pormenorizada do exemplo da figura 5.6

Elemento	Corr.	Esp.	Incomp.	Exced.	Clas.	Em Falta	Total	Esperado
Autor	2	0	1	1	0	0	4	4
Título	1	0	0	0	0	0	1	1
Editores	2	0	0	0	0	0	2	2
Conferencia	1	0	0	0	0	0	1	1
Abreviatura	1	0	0	0	0	0	1	1
Local	2	0	0	0	0	0	2	2
Página	0	1	0	1	0	0	2	1
Data	0	0	0	0	0	1	0	1
Editora	0	0	0	0	0	1	0	1
Total	9	1	1	2	0	2	13	14

elementos será a soma do número de elementos (autores, editores, páginas, etc.) correctos sobre o número de elementos que o sistema identificou.

Dada esta categorização inicial, é possível especificar medidas de desempenho idênticas às usadas anteriormente. Assim consideramos as seguintes medidas para a avaliação da extracção de elementos:

1) Precisão

$$Precisao_{AUTOR} = \frac{\#ElementosCorrectos}{\#ElementosIdentificados} \tag{5.8}$$

2) Abrangência

$$Abrangencia_{AUTOR} = \frac{\#ElementosCorrectos}{\#ElementosReferencia} \tag{5.9}$$

3) Medida F

$$MedidaF_{AUTOR} = 2.\frac{Precisao.Abrangencia}{Precisao + Abrangencia} \tag{5.10}$$

4) Precisão alargada

$$PrecisaoAlargada_{AUTOR} = \frac{\#ElementosCorrectos + \#ElementosExcedentarios}{\#ElementosIdentificados} \tag{5.11}$$

5) Abrangência alargada

$$AbrangenciaAlargada_{AUTOR} = \frac{\#ElementosCorrectos + \#ElementosExcedentarios}{\#ElementosReferencia} \tag{5.12}$$

6) Sobre-geração

$$Sobre{-}geracao_{AUTOR} = \frac{\#ElementosIncompletos + \#ElementosOmissos}{\#ElementosReferencia} \tag{5.13}$$

7) Sub-geração

$$Sub - geracao_{AUTOR} = \frac{\#Elementoserrados}{\#Elementosidentificados} \tag{5.14}$$

5.4.3 Materiais de teste

Aqui fica descrita uma metodologia possível para obter grande número de dados de avaliação semi-automaticamente, através de estudos de mutilação ("ablation studies", em inglês).

De facto, são conhecidos os valores dos vários elementos constantes do catálogo da Linguateca. Seria possível não só extrair como mutilar ou truncar muitas das entradas, de forma a testar o resultado do SUPeRB sobre as referências (mutiladas) obtidas. Além disso, podia também usar-se a capacidade de gerar formatos vários a partir da informação no catálogo, de forma a poder ter um leque mais variado de referências a analisar.

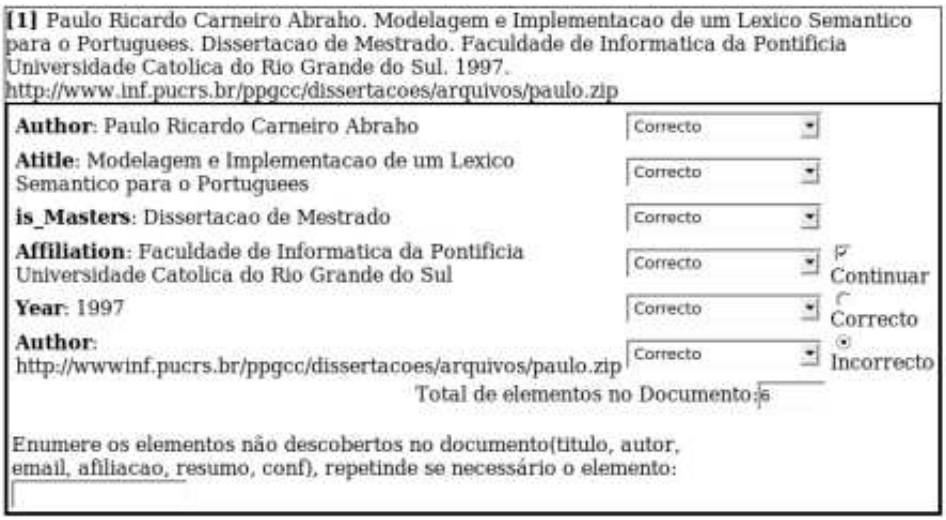


Figura 5.7: Interface de avaliação da extracção de elementos bibliográficos

A figura 5.7 mostra como, para cada elemento extraído, existe uma caixa de opções para escolher a opção correcta. É possível preencher o formulário e guardar os dados. Os dados e as respectivas medidas de avaliação poderão ser futuramente consultados.

5.4.4 Exemplo de avaliação

Por razões de tempo, não foi possível efectuar testes de desempenho reais no módulo de extracção de referências (excepto o caso da auto-avaliação descrito acima). Contudo, exemplificamos alguns dos cálculos usando como exemplo o elemento AUTOR e os valores da tabela 5.8.

$Precisao_{AUTOR} = \frac{2}{4} = 0,5$

$Abrangencia_{AUTOR} = \frac{2}{4} = 0,5$

$MedidaF_{AUTOR} = 2.\frac{0,5.0,5}{0,5+0,5} = 2.\frac{0,25}{1} = 0,5$

$PrecisaoAlargada_{AUTOR} = \frac{2+1}{4} = 0.75$

$AbrangenciaAlargada_{AUTOR} = \frac{3+1}{4} = 0,75$

$Sobre - geracao_{AUTOR} = \frac{1+0}{4} = 0,25$

$Sub - geracao_{AUTOR} = \frac{0}{4} = 0$

5.5 Avaliação global

Um avaliação parcial, em que cada módulo é considerado individualmente de forma a encontrar soluções que tornem esse módulo mais robusto não impede que o sucesso do SUPeRB não tenha de ser avaliado no seu conjunto, com utilizadores a executar tarefas reais no seu ambiente de trabalho quotidiano.

Assim, foi pedido à equipa da Linguateca que, durante 15 dias, ao procurar referências, guardasse:

- A informação de que dispunha inicialmente;
- A sua necessidade de informação;
- A referência bibliográfica final (depois de ter encontrado);
- Eventualmente o tempo que demorou a encontrá-la e a preenchê-la;

A informação recolhida poderá depois vir a ser utilizada no SUPeRB para comparar e medir o quanto a invocação totalmente automática do SUPeRB, assim como o uso de validação intermédia, ajuda ou não em cada um dos casos.

Finalmente, é possível ainda imaginar experiências com utilizadores, a quem é dado um conjunto de "problemas bibliográficos"para resolver. Estes problemas seriam resolvidos metade com a ajuda do SUPeRB, outra metade não, permitindo futuramente comparar os resultados.

Capítulo 6

Comentários finais

Inicialmente propôs-se, como objectivos, a criação de uma plataforma capaz de pesquisar informação bibliográfica na Web, extrair essa informação, e processá-la, de forma a se obter informação relevante e refinada. Propôs-se ainda a criação de meios para reutilizar e gerir essa informação recolhida em conjunto com o catálogo de publicações da Linguateca.

Esta dissertação abordou um leque alargado de áreas para tornar realizáveis as tarefas necessárias para cumprir os objectivos a que se propôs. Só assim foi possível específicar uma arquitectura e construir um sistema capaz de providenciar as funcionalidades necessárias para atingir esses objectivos. Ou seja, a abordagem tomada passou por analisar cada problema independentemente e estudar as soluções possíveis, procurando em seguida encontrar uma solução viável que servisse as necessidades impostas, e implementar essa solução, tomando em consideração a existência de recursos que pudessem ser reutilizados (por exemplo o ParaTools e o REPENTINO). Assim, foi possível criar meios para pesquisar informação bibliográfica na Web e processar essa informação bibliográfica, integrando os resultados das várias tarefas.

6.1 Cômputo geral

Nesta tese foi proposta uma arquitectura para realizar os objectivos propostos, em que a solução consistiu em desenvolver diversos módulos independentes,

cada um responsável por uma tarefa; mas que, integrados na arquitectura, podem ser executados em cadeia, e produzir informação mais refinada.

O desenvolvimento do SUPeRB ainda não está terminado. Ainda não dispomos de um sistema capaz de processar uma expressão a pesquisar na Web e apresentar como resultado as referências bibliográficas relevantes no formato desejado. No entanto, muitos dos processos intermédios já são possíveis com alguma fiabilidade, proporcionando funcionalidades úteis por si só, tais como:

1) a extracção de texto a partir de documentos em vários formatos,

2) a extracção de referências a partir de texto,

3) a extracção de elementos bibliográficos a partir de uma referência bibliográfica,

4) a conversão entre formatos bibliográficos,

5) e a possibilidade de permitir ao utilizador marcar referências bibliográficas.

Algumas destas tarefas podem inclusive já ser executadas em sequência, trocando informação entre si. Por exemplo, a partir de um dado URL, pode já obter-se informação bibliográfica estruturada, o que inclui a sequência de 3 módulos.

Foram também criadas interfaces Web que permitem a interacção mais fácil de um utilizador com as funcionalidades em questão, permitindo validar os resultados apresentados pelo sistema.

Por outro lado, foi proposto um método de avaliação para algumas das funcionalidades já disponíveis, e primeiras avaliações foram levadas a cabo segundo essa metodologia Este estudo permitiu-nos ponderar quais dessas funcionalidades podem ser melhoradas para providenciar um melhor serviço.

Finalmente o SUPeRB encontra-se disponível em `http://www.linguateca.pt/SUPeRB`, podendo ser utilizado publicamente por qualquer utilizador.

6.2 Trabalho futuro

Da secção anterior pode contudo concluir-se que ainda não atingimos todos os nossos objectivos, sendo além disso ainda necessárias algumas melhorias. Mais especificamente:

1) É necessário integrar todas as funcionalidades de forma a, a partir de uma expressão, obter-se um conjunto refinado de informação bibliográfica precisa e que possa ser utilizada para diversos fins, desde armazenar no catálogo de publicações a poder reutilizar a informação obtida para obter mais informação bibliográfica.

2) É preciso ainda integrar o SUPERB no catálogo de publicações da Linguateca, melhorando a troca de informação entre ambos.

3) Falta ainda implementar a automatização das tarefas, permitindo que periodicamente o sistema procure obter informação adicional ou corrigir informação no catálogo de publicações através da calendarização individual ou de um conjunto de referências.

4) É ainda necessário considerar a personalização de utilizadores de forma a possibilitar pesquisas personalizadas, e identificar o utilizador que inseriu determinadas referências submetidas no catálogo, bem como manter um historial das acções levadas a cabo por cada utilizador e permitir o armazenamento de referências privadas.

Existe ainda algum trabalho de documentação e de disponibilização do código, em forma de pacotes Perl. Como referido, todos os módulos do SUPeRB têm sido desenvolvidos de forma independente. Ainda é necessário algum esforço para a clarificação de todas as dependências necessárias entre os vários módulos de Perl para poder tornar público estes módulos, de forma a serem facilmente instalados e utilizados por outros programas.

6.3 Áreas de investigação em aberto

Muitas áreas e problemas científicos podem ainda ser investigados tendo em vista a expansão das capacidades do SUPeRB.

Um destes problemas pode ser por exemplo a classificação automática de textos, mencionada na secção 4.2.6.2. A classificação automática é um processo paralelo à classificação manual, cujo meio de funcionamento seria de tentar atribuir a mesma classificação atribuída manualmente com base em grupos (*clusters*) ou através de regras que pudessem ser facilmente introduzidas. (Geffet e Feitelson, 2001; Montejo-Ráez et al., 2005; Sarmento, 2005) apresentam várias aplicações destes métodos.

Outro caminho possível é o de expandir as funcionalidades na área da Web Semântica, permitindo a integração do SUPeRB com outros programas ou repositórios bibliográficos (Shadbolt et al., 2006), permitindo a integração de diferentes recursos e integrar diferentes ontologias para um mesmo fim.

Também pode ser interessante não só completar a avaliação global prevista na secção 5.5 mas também proceder a avaliações com utilizadores noutras áreas, que possam revelar outras necessidades que não tenham sido abordadas até ao momento.

É também importante considerar a usabilidade das interfaces. O SUPeRB é um caso apropriado para fazedr um estudo de usabilidade, quer para aplicar técnicas já existentes, quer para sugerir novas metodologias de interacção que possam usufruir da Web 2.0, de forma a aumentar a satisfação dos utilizadores.

Finalmente, é possível pensar na criação de um sistema de resposta a perguntas, específico a questões bibliográficas, que permitisse que os utilizadores comunicassem através de linguagem natural com o SUPeRB, por exemplo fazendo perguntas como *Quem é* ***X****?*, *Qual o domínio do trabalho de* ***Y****?* ou *Com quem publica* ***Z****?*

No seguimento desta última questão, outra mais valia pode ser a implementação de co-citações, permitindo agrupar referências bibliográficas que estejam relacionadas.

O SUPeRB deu apenas os seus primeiros passos, como uma ferramenta de

descoberta e processamento de informação bibliográfica. Espera-se que o trabalho futuro possa vir a justificar o soberbo nome com que foi baptizado.

Apêndice A

Características da implementação

Aqui é descrita a implementação do SUPeRB, nomeadamente a linguagem de programação e os recursos utilizados.

A.1 Características genéricas

O sistema base em que o SUPeRB está disponível ao público é um Linux Red Hat, kernel 2.4.20, tendo grande parte do trabalho de desenvolvimento sido levado a cabo num sistema com o linux Fedora Core 4, kernell 2.6.11. Também foram feitos testes num Linux Ubuntu, kernell 2.6.17.

O Linux foi criado por Linus Torvalds, mais especificamente o kernel do Linux. É dos sistemas operativos onde mais predomina a existência de software *Open Source* e software livre. O Linux é dos mais utilizados como servidores Web (NetCraft), como é exemplo o servidor onde se pretende manter o SUPeRB é Linux, um servidor que alberga já algumas aplicações e recursos da Linguateca, inclusive o catálogo de publicações da Linguateca.

O SUPeRB foi implementado em Perl[1], uma linguagem de *scripting*, criada em 1987 por Larry Wall. Uma das principais razões para esta escolha é por razões históricas de forma a permitir a integração com o catálogo da Linguateca. Mas

[1] `http://www.perl.com`

a implementação em Perl deve-se também à portabilidade desta linguagem para vários sistemas operativos, a facilidade em desenvolver CGI e também por ser uma linguagens mais versáteis no que diz respeito ao processamento de expressões regulares.

Como recurso para armazenamento de dados, utiliza-se o MySQL[2], um sistema de gestão de base de dados, um sistema multi-plataforma, funcionando em Linux e possuindo API para inúmeras linguagens, de entre as quais o Perl. O uso de uma base de dados tem em vista o armazenamento dos resultados de forma estruturada para fácil acesso e o recurso a algumas funcionalidades de pesquisa em texto que são fornecidas pelo MySQL.

O SUPeRB possui uma interface Web (também é possível executar o SUPeRB através da linha de comando) executado pelo *Apache HTTP Server*[3]. O Apache é um servidor de conteúdo estático e dinâmico (como é o caso de CGI Perl) multi-plataforma que desenvolveu um papel importante no enriquecimento da WWW e é um dos servidores mais utilizados (NetCraft).

A combinação do software utilizado é caracterizada como LAMP, (Linux + Apache + MySQL + Perl), uma plataforma *Open Source* para aplicações Web. Ou seja, o SUPeRB corre sobre esta plataforma sem recurso a software proprietário. Apesar do SUPeRB ter sido desenvolvido nesta plataforma, nada impede a implementação da arquitectura numa combinação diferente. As opções feitas foram tomadas considerando a necessidade de integrar a aplicação com uma outra aplicações já existente, e de possibilitar o funcionamento no sistema Linux.

A.2 Optimização do processamento de pedidos

Todo o processo é relativamente pesado, o que atrasa a sua finalização, mas existem partes que consomem mais tempo. Os pedidos a serviços Web e a obtenção de documentos na Web são caso disso. Estes não dependem da carga do processador mas sim do tempo de resposta individual de cada serviço Web ou dos servidores/sítios que alojam os documentos que se pretende obter.

[2] `http://www.mysql.com`
[3] `http://httpd.apache.org/`

Nesta fase podem ocorrer mais de uma dúzia de pedidos a vários serviços Web e serem descarregados mais de uma dezena de documentos da Web. Esta fase é um ponto crítico, onde o sistema facilmente pode ficar paralisado quer por um serviço Web demorar em enviar uma resposta ou um servidor Web demorar a enviar um documento. Para reduzir o perigo de isto suceder, esta secção foi optimizada para correr várias *threads* para fazer pedidos a serviços Web e simultaneamente ir buscar os resultados que vão sendo obtidos. As *threads* partilham a memória, pelo que a comunicação é feita através de filas partilhadas. Na figura A.1 pode ver-se um exemplo demonstrando a sequência de mensagens e a inicialização de *threads*.

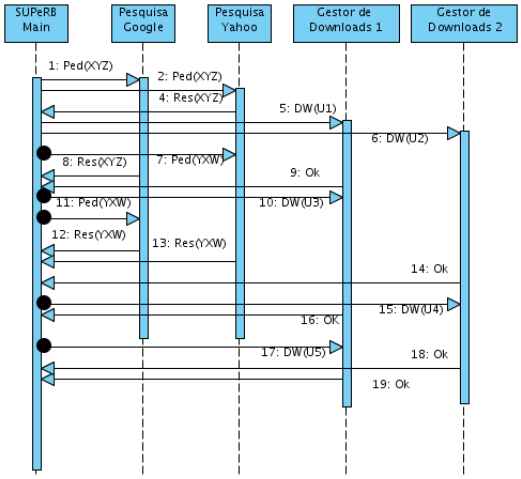


Figura A.1: Diagrama de sequência das *threads* na pesquisa.

É feito um pedido usando expressão XYZ a diversos serviços Web simultaneamente. Quando chegam os primeiros resultados, as threads iniciais acedem a uma fila que contém um número limitado de expressões. Ao mesmo tempo, é preenchida uma lista com os resultados e são iniciadas várias *threads* para ir buscar cada um dos resultados que, quando terminam, verificam se existem outros documentos para ir buscar, obtendo essa informação a partir da lista de resultados.

A.3 Módulos desenvolvidos de raiz

Esta secção descreve alguns módulos Perl desenvolvidos até ao momento ou em desenvolvimento e que poderão ser futuramente disponibilizados.

SUPeRBTools::Search Módulo que implementa várias metodologias para recolher informação de motores de pesquisa da Web.

SUPeRBTools::FileUtils Módulo que implementa os algoritmos para processar ficheiros, nomeadamente a extracção de texto de documentos, a conversão entre códigos de caracteres (*charsets*) e a correcção de acentos. Este módulo depende de outros programas, responsáveis pelas transformações de diversos formatos, tais como o xpdf (pdf2text), o ghostscript (ps2ascii) e o Jakarta POI[4](doc e ppt).

SUPeRBTools::ReferenceExtractor Módulo responsável por extrair referências bibliográficas de texto. Pode extrair listas de texto ou auto-referências.

SUPeRBTools::ReferenceParser Módulo responsável por extrair elementos bibliográficos de referências.

SUPeRBTools::ReferenceConverter Módulo responsável por converter entre vários formatos.

A.4 Alguns módulos utilizados

Esta secção refere os módulos, e o autor ou responsável pelo módulo, mais importantes que são utilizados ou que foram avaliados ao longo do desenvolvimento do SUPeRB.

Biblio::Citation::Parser Plataforma para o parsing de referências bibliográficas.
Desenvolvido por *Mike Jewell*

[4]`http://jakarta.apache.org/poi/`

`http://search.cpan.org/~mjewell/Biblio-Citation-Parser-1.10/lib/Biblio/Citation/Parser.pm`

DBI Módulo de interface a bases de dados.
Desenvolvido por *Tim Bunce*.
`http://search.cpan.org/~timb/DBI-1.52/DBI.pm`

HTML::TokeParser Um dos diversos módulos para fazer a análise sintáctica de HTML.
Desenvolvido por *Gisle Aas*.
`http://search.cpan.org/~gaas/HTML-Parser-3.55/lib/HTML/TokeParser.pm`

HTML::TokeParser::Simple Um dos diversos módulos para fazer a análise sintáctica de HTML.
Este módulo é uma interface simplificada para o **HTML::TokeParser**.
Desenvolvido por *Curtis Poe*.
`http://search.cpan.org/dist/HTML-TokeParser-Simple/lib/HTML/TokeParser/Simple.pm`

Lingua::Identify Módulo que permite identificar a língua de um determinado texto. Suporta 33 linguas, de entre as quais o português.
Desenvolvido por *José Alves Castro*.
`http://search.cpan.org/~cog/Lingua-Identify-0.18/lib/Lingua/Identify.pm`

Lingua::PT::PLNbase Módulo de PLN para o português. Contém métodos para separar frases e atomizar texto em português.
Desenvolvido por *Alberto Manuel Brandão Simões*.
`http://search.cpan.org/~ambs/Lingua-PT-PLNbase-0.14/lib/Lingua/PT/PLNbase.pm`

LWP::UserAgent Este módulo é uma implementação de um agente Web. É usado para fazer pedidos via HTTP.
Desenvolvido por *Gisle Aas*.
`http://search.cpan.org/~gaas/libwww-perl-5.805/lib/LWP/UserAgent.pm`

Net::Google Modulo de interface para a API publica do motor de pesquisa Google.
Desenvolvido por *Aaron Straup Cope*.
`http://search.cpan.org/~bstilwell/Net-Google-1.0.1/lib/Net/Google/Search.pm`

Repentino.pm Um módulo que contém a versão local do Repentino, `http://www.linguateca.pt/repentino/`.
Desenvolvido por *Luís Sarmento*.
`http://paginas.fe.up.pt/~las/conteudo/soft/REPENTINO_0.01.tgz`

SOAP::Lite Colecção de módulos que fornecem uma interface para o SOAP tanto como cliente como servidor.
Desenvolvido por *Byrne Reese*.
`http://search.cpan.org/~byrne/SOAP-Lite-0.69/lib/OldDocs/SOAP/Lite.pm`

XML::Simple API para processar facilitar a análise sintáctica de XML.
Desenvolvido por *Grant McLean*.
`http://search.cpan.org/~grantm/XML-Simple-2.14/lib/XML/Simple.pm`

Yahoo::Search Módulo de Interface para a API pública do motor de pesquisa Yahoo!.
Desenvolvido por *Jeffrey Friedl*.
`http://search.cpan.org/~jfriedl/Yahoo-Search-1.7.10/lib/Yahoo/Search.pm`

Apêndice B

Lista de servidores SRW/SRU conhecidos

Vários servidores SRU/SRW[1]:

- Depósito de Dissertações e Teses Digitais `http://dited.bn.pt/mitra/jsp/sru.jsp`
- BIOME (Internet Resources in the Health and Life Sciences) `http://tweed.lib.ed.ac.uk:8080/elf/search/biome?operation=explain\&version=1.1`
- COPAC Database `http://tweed.lib.ed.ac.uk:8080/elf/search/copac?operation=explain\&version=1.1`
- Cheshire3 Sample Databases `http://srw.cheshire3.org/services/l5` `http://srw.cheshire3.org/services/spy` `http://srw.cheshire3.org/services/syrinnia`
- EEVL (Internet Guide to Engineering, Mathematics, and Computing) `http://tweed.lib.ed.ac.uk:8080/elf/search/eevl?operation=explain\&version=1.1`

[1] A maioria dos links desta lista estão também disponíveis em `http://www.loc.gov/standards/sru/servers.html`.

- Index Data – Gateway to LC
  `http://www.indexdata.dk:9000/voyager?operation=explain\&version=1.1`

- Koninklijke Bibliotheek / The European Library
  `http://krait.kb.nl/cgi-zoek/sru.pl?operation=explain\&version=1.1`

- Library of Congress Online Catalog
  `http://z3950.loc.gov:7090/voyager?operation=explain\&version=1.1`

- National Library of Scotland
  `http://tweed.lib.ed.ac.uk:8080/elf/search/nls?operation=explain\&version=1.1`

- OAI Registry at University of Illinois Library at Urbana-Champaign
  `http://gita.grainger.uiuc.edu/registry/sru/sru.asp?operation=explain\&version=1.1`

- OCLC GSAFD Database
  `http://alcme.oclc.org/srw/search/GSAFD?operation=explain\&version=1.1`

- OCLC PICA SRU Test Database
  `http://greta.pica.nl:1080/sru/?operation=explain\&version=1.1`

- OCLC SOAR Database
  `http://alcme.oclc.org/srw/search/SOAR?operation=explain\&version=1.1`

- Open University
  `http://tweed.lib.ed.ac.uk:8080/elf/search/open?operation=explain\&version=1.1`

- Oxford University
  `http://tweed.lib.ed.ac.uk:8080/elf/search/oxford?operation=explain\&version=1.1`

- Resource Discovery Network ResourceFinder
  `http://www.rdn.ac.uk:8080/xxdefault/?operation=explain\&version=1.1`
- Social Science Information Gateway
  `http://tweed.lib.ed.ac.uk:8080/elf/search/sosig?operation=explain\&version=1.1`
- University of Edinburgh
  `http://tweed.lib.ed.ac.uk:8080/elf/search/edinburgh?operation=explain\&version=1.1`
- University of Glasgow
  `http://tweed.lib.ed.ac.uk:8080/elf/search/glasgow?operation=explain\&version=1.1`
- University of Southampton
  `http://tweed.lib.ed.ac.uk:8080/elf/search/southampton?operation=explain\&version=1.1`
- University of Toronto
  `http://ibridge.library.utoronto.ca:2200/unicorn?operation=explain\&version=1.1`
- British Library
  `http://herbie.bl.uk:9080/Gateway/index.html`

Glossário

Ajax (*Asyncronous JavaScript and XML*) É um conjunto de tecnologias que aumenta a interacção de aplicações Web. As tecnologias que o compõem são: Javascript, DOM, CSS, XML, e comunicação assíncrona entre o cliente e o servidor.
Ver também Web 2.0, DOM, XML.

AMA (*American Medical Association*) Estilo padrão no domínio da medicina.

APA (*American Psycological Association*) Estilo padrão no domínio da psicologia e outras ciências sociais.

API (*Application Programming Interface*) é um conjunto de rotinas e padrões estabelecidos por um software para utilização de suas funcionalidades por programas aplicativos – isto é: programas que não querem envolver-se em detalhes da implementação do software, mas apenas usar seus serviços.

BibTex Ferramenta para formatar listas de referências bibliográficas usado pelo LaTeX. Também conhecido como o formato BibTex.

Chicago Estilo padrão em livros, revistas jornais e outros tipos de publicações não académicas.

CiteSeer (*Scientific Literature Digital Library*) Repositório on-line de publicações na área de ciência de computadores.

DBLP (*Computer science bibliography*) Repositório on-line de publicações na área de ciência de computadores.

DOM (*Document Object Model*) É uma especificação da W3C, independente da linguagem e plataforma, para poder manipular a estrutura do documento HTML.

Elemento bibliográfico Unidade definida de informação numa referência bibliográfica.

Google Motor de pesquisa na Internet, `http://www.google.com`

HTML (*HyperText MArkup Language*) é linguagem utilizada para produzir páginas Web.

Interface de Programação de Aplicativos *ver* API

JavaScript Uma linguagem de programação que oferece interactividade a páginas HTML.

MLA (*Modern Language Association*) Estilo padrão no domínio da literatura, artes e humanidades.

MSN Portal de pesquisa Web da Microsoft, `http://www.msn.com`

OAI *ver* Open Access Initiative

OPAC (*Online Public Access Catalog*) Índice online de conteúdos pertencentes ou licenciados a uma biblioteca.

Open Access Initiative Iniciativa para a disponibilização livre de conteúdos científicos. `http://www.openarchives.org/`

OWL (*Web Ontology Language*) É uma linguagem para definir e povoar ontologias para a Web.

Peer-review Avaliação de publicações científicas por pares com conhecimento na área.

Perl (*Pratical Extraction Report Language*) Linguagem de scripting bastante forte no uso de expressões regulares.

Referência bibliográfica Conjunto de elementos bibliográficos que identificam uma publicação ou parte dela.

Semantic Web *ver* Web Semântica

Serviço Web Sistema que permite a interoperabilidade através de uma rede. A sua interface é descrita através de WSDL.
Ver também WSDL, XML.

SGML (*Standard Generalized Markup Language*) hecido como SGML, é uma metalinguagem através da qual se podem definir linguagens de marcação. Exemplos de linguagens derivadas do SGML são o XML ou o HTML.

SOAP (*Simple Object Access Protocol*) É um protocolo de comunicação que permite a troca de mensagens XML em redes de computadores.

SUPeRB **S**istema **U**niformizado de **Pe**squisa de **R**eferências **B**ibliográficas.

URI (*Uniform Resource Identifier*) Cadeia de caracteres (*string*) num formato padrão que descreve um recurso na Web.

URL (*Uniform Resource Locator*) Sinónimo de URI.
Ver também URI.

W3 *ver* World Wide Web

W3C (*World Wide Web Consortium*) Consórcio de empresas que tem como objectivo desenvolver tecnologias e protocolos comuns e promover a interoperabilidade através da Internet `http://www.w3c.org`.
Ver também World Wide Web.

Web *ver* World Wide Web

Web 2.0 O termo Web 2.0 refere-se à segunda geração de serviços, aplicações e recursos da Web. A Web 2.0 pode ser sinónimo para semantic Web e apesar de se complementarem, a Web 2.0 e maiso provávelmetne um passo em direcção à Web Semântica.
Ver também Web Semântica

Web semântica Projecto que visa a criação de um meio universal para a troca de informação através de documentos passíveis de serem processados por programas através da Web.

Web services *ver* Serviços Web

World Wide Web É um sistema de documentos em hipertexto e outros tipos de média (imagens, videos, sons, etc.), que corre sobre a Internet. Recorrendo a um navegador (*Browser*), é possível navegar entre esses documentor usando as hiperligações que os ligam.

WSDL (*Web Services Description Language*) Formato XML para descrever serviços Web.
Ver também XML, Serviços Web.

WWW *ver* World Wide Web

XHTML (*eXtensible HyperText MArkup Language*) É uma reformulação da linguagem de marcação HTML, baseada em XML.
Ver também HTML, XML.

XML (Extensible Markup Language) Linguagem de marcação recomendada pela W3C. O seu objectivo principal é o de facilitar a troca de informação através da Internet. O XML deriva de uma outra linguagem, o SGML.
Ver também W3C, SGML.

Yahoo Motor de pesquisa na Internet, `http://www.yahoo.com`

Referências

Eugene Agichtein, Steve Lawrence e Luis Gravano. "Learning to find answers to questions on the Web". *ACM Trans. Inter. Tech.*, 4(2):129–162, 2004.

Kent Anderson, John Sack, Lisa Kraus e Lori O'Keefe. "Publishing Online-Only Peer-Reviewed Biomedical Literature: Three Years of Citation, Author Perception, and Usage Experience". *Journal of Electronic Publishing*, 6(3), 2001.

Naveen Ashish e Craig Knoblock. "Wrapper generation for semi-structured Internet sources". *ACM SIGMOD Record*, 26(4):8–15, 1997.

NBR 6023. *NBR 6023: Norma Brasileira*. Associação Brasileira das Normas Técnicas, Agosto 2002.

Marco Baroni e Silvia Bernardini. "BootCat: Bootstrapping corpora and terms from the web". Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editores, *Proceedings of Language Resources and Evaluation Conference (LREC 2004)*, págs. 1313–1316. ELDA, 26-28 May 2004.

Marco Baroni e Silvia Bernardini. *WaCky: Working papers on the Web as a Corpus*. Bologna. September 2006. ISBN 88-6027-004-9. GEDIT.

Tim Berners-Lee. World Wide Web, 3 November 1992. URL: `http://www.w3.org/History/19921103-hypertext/hypertext/WWW/TheProject.html`.

Kurt D. Bollacker, Steve Lawrence e C. Lee Giles. "CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications". Em Katia P. Sycara e Michael Wooldridge, editores, *Proceedings*

of the Second International Conference on Autonomous Agent (Agents'98), págs. 116–123, New York, May 9-13 1998. ACM Press.

Fabio Ciravegna, Sam Chapman, Alexiei Dingli e Yorick Wilks. "Learning to Harvest Information for the Semantic Web". Em *The Semantic Web: Research and Application*, volume 3053 de *Lecture Notes in Computer Science*, págs. 312–326. Springer Berlin/Heidelberg, September 09 2004.

João Paulo Cordeiro. Extracção de Elementos Relevantes em Texto/Páginas da World Wide Web. Dissertação de mestrado, Faculdade de Ciências da Universidade do Porto, Porto, Junho 2003.

Luís Costa. "Esfinge - Resposta a perguntas usando a Rede". Em José María Gutiérrez, Flavia Maria Santoro e Pedro Isaías, editores, *Proceedings da conferência IADIS Ibero-Americana WWW/Internet 2005*, págs. 616–619. IADIS Press, 2005.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu Tal Shaked, Stephen Soderland, Daniel S.Weld e Alexander Yates. "Unsupervised Named-Entity Extraction from the Web: An Experimental Study". *Artificial Intelligence Journal*, 165(1):91–134, 2005.

Dror G. Feitelson. "Cooperative Indexing, Classification, and Evaluation in BoW". *Proceedings of the 7th International Conference on Cooperative Information Systems*, págs. 66–77, 2000.

Maayan Geffet e Dror G. Feitelson. "Hierarchical indexing and document matching in BoW". Em *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, págs. 259–267, New York, NY, USA, 2001. ACM Press.

Junfei Geng. Automatic Extraction and Integration of Bibliographic information on the Web Using Hidden Markov Models. Dissertação de mestrado, Duke University, 2002.

Scott Golder e Bernardo A. Huberman. "The Structure of Collaborative Tagging Systems". *Journal of Information Science*, 32(2):198–208, 2006.

T. R. Gruber. "A translation approach to portable ontologies". *Knowledge Aquisition*, 5(2):199–220, 1993.

A. Gulli e A. Signorini. "The indexable web is more than 11.5 billion pages". Em *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, págs. 902–903, New York, NY, USA, 2005. ACM Press.

I-Ane Huang, Jan-Ming Ho, Hung-Yu Kao e Weng-Chang Lin. "Extracting Citation Metadata from Online Publication Lists Using BLAST". *Lecture Notes in Computer Science*, págs. 539–548, 2004.

NP 405-1. *NP 405-1: Norma Portuguesa: Documentos Impressos*. Instituto português da Qualidade, Janeiro de 1994.

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

Internet users Statistics. World Internet Users and Population Stats, 2006. URL: `http://www.internetworldstats.com/stats.htm`.

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

Leslie Lamport. *LaTeX: a document Preparation System*. 2ª edição. 1986. Addison-Wesley Publishing Company.

Steve Lawrence, C. Lee Giles e Kurt Bollacker. "Digital Libraries and Autonomous Citation Indexing". *IEEE Computer Society Press*, 32(6):67–71, 1999.

Linguateca. Nos bastidores do projecto, 2005. URL: `http://acdc.linguateca.pt/bastidores.html`. `http://acdc.linguateca.pt/bastidores.html`.

Peter Mika. "Social Networks and the Semantic Web". Em *IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*, págs. 285–291. IEEE Computer Society, 20-24 September 2004.

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 SemanticWeb Conference, ISWC 2005*, volume 3729 de *Lecture Notes in Computer Science*, págs. 522–536. Springer, November 6-10 2005.

Arturo Montejo-Ráez, L. Alfonso Ureña-López e Ralf Steinberger. "Text Categorization using bibliographic records: beyond document content". *Processamiento del Lenguage Natural*, (35):119–126, Septiembro 2005.

NetCraft. Netcraft: September 2006 web server survey, 2006. `http://news.netcraft.com/archives/web_server_survey.html`.

Andrew Odlyzko. "The rapid evolution of scholarly communication". *Learned Publishing*, 15(1):7–19, January 2002.

Tim O'Reilly. What is the Web 2.0: Design Patterns and Business Models for the Next Generation of Software. O'reilly Media, `http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html`, 30 September 2005.

Yves Petinot, C. Lee Giles, Vivek Bhatnagar, Pradeep B. Teregowda2, Hui Han e Isaac Councill. "CiteSeer-API: Towards Seamless Resource Location and Interlinking for Digital Libraries". Em *CIKM'04: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, págs. 553–561, New York, NY, USA, 2004. ACM Press.

Luca Previtali, Brenno Lurati e Erik Wilde. "BibTeXML: An XML Representation of BibTeX". Em *World Wide Web Conference, WWW 10*, 2001.

RIS, reference manual. *RIS Format Specifications*, 10 edição, February 2004. `http://www.adeptscience.co.uk/kb/article/A626`.

Diana Santos. "O projecto Processamento Computacional do Português: Balanço e perspectivas". Em Maria das Graças Volpe Nunes, editor, *V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*, págs. 105–113, São Paulo, 2000. ICMC/USP.

Diana Santos. "Um centro de recursos para o processamento computacional do português". *DataGramaZero - Revista de Ciência da informação*, 3(1), 2002.

Diana Santos e Nuno Cardoso. *HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para português: documentação e actas do encontro*. Linguateca. 2007.

Diana Santos, Nuno Cardoso e Nuno Seco. "Avaliação no HAREM: Métodos e medidas". Relatório Técnico TR-06-17, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa, Novembro 2006. URL: `http://www.di.fc.ul.pt/tech-reports/06-17.pdf`.

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 Miriani, Jan Odjik e Daniel Tapias, editores, *Proceedings of Language Resource and Evaluation Conference (LREC'2006)*, págs. 1986–1991, May 22-28 2006.

Diana Santos, Alberto Simões, Ana Frankenberg-Garcia, Ana Pinto, Anabela Barreiro, Belinda Maia, Cristina Mota, Débora Oliveira, Eckhard Bick, Elisabete Ranchhod, José João Dias de Almeida, Luís Cabral, Luís Costa, Luís Sarmento, Marcirio Chaves, Nuno Cardoso, Paulo Rocha, Rachel Aires, Rosário Silva, Rui Vilela e Susana Afonso. "Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa". Em Guillermo De Ita Luna, Olac Fuentes Chávez e Mauricio Osorio Galindo, editores, *Proceedings of the international workshop "Taller de Herramientas y Recursos Linguísticos para el Espanõl y el Português", IX Iberoamerican Conference on Artificial Intelligence*, págs. 147–154, 2004.

Luís Sarmento. "A Simple and Robust Algorithm for Extracting Terminology". Em *META Simposium - For a Proactive Translatology*, Québec, Canada, April 2005. Université de Montréal.

Luís Sarmento. "SIEMÊS - a named entity recognizer for portuguese relying on similarity rules". Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias,

editores, *7th Workshop on Computational Processing of Written and Spoken Language*, volume LNAI 3960, págs. 90–99. Springer, 2006.

Nigel Shadbolt, Wendy Hall e Tim Berners-Lee. "The Semantic Web Revisited". *IEEE Intelligent Systems*, 21(3):96–101, 2006.

Radu Soricut e Eric Brill. "Automatic question answering using the web: Beyond the Factoid". *Information Retrieval*, 9(2):191–206, 2006.

Sara Stevens-Rayburn e Ellen N. Bouton. "If it is not in the Web it does not exist at all". *Library and information services in astronomy III*, págs. 195–203, 1998.

ISO 690:1987. *ISO 690:1987*. Technical Commitee (TC)46, 1987.

ISO 690-2. *ISO 690 - Part 2*. Technical Commitee (TC)46, 1997.

Juan Ignacio Vazquez, Joseba Abaitua e Diego Lópes de Ipiña. "The Ubiquitous Web as a model to lead our environments to their full potential". Em *W3C Workshop on the Ubiquitous Web*, March 2006.

Soap Version 1.2. *SOAP Version 1.2*. W3C, 24 June 2003. `http://www.w3.org/TR/soap/`.

Thomas Vander Wal. Folksonomy definition and wikipedia, November 2005. URL: `http://www.vanderwal.net/random/entrysel.php?blog=1750`.

Wikipedia. Folskonomies — Wikipedia, the free encyclopedia, 2006. URL: `http://en.wikipedia.org/wiki/Folksonomy`. Versão de 29 November 2006.

Z39.50-2003. Information retrieval (z39.50): Application service definition and protocol specification, November 2003. ISSN 1041-5653. Approved November 27, 2002 by the American National Standards Institute.