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.