

Integração de Sistemas

Relatório - Trabalho #1

Gustavo Fernandes
2006125010

Ricardo Lopes
2008114843

Rui Chicória
2008115099

Introdução

Este trabalho tem como objectivo a implementação de uma aplicação que extraia apenas a informação relevante de um site, recorrendo à técnica de *screen scrapping*, de forma a dar-lhe um significado como *objecto* de forma ser compreendido pelas *máquinas*. De forma a demonstrar a importância desta prática apresentamos os mesmos dados numa aplicação completamente independente da original, e da mesma forma poderíamos utilizá-los para muitos outros fins. Este relatório visa clarificar elementos do trabalho realizado e justificar as decisões tomadas na abordagem de resolução do problema apresentado.

Informação Recolhida

A aplicação escolhida para a extracção de informação foi a secção de eBooks em português do site da Livraria Bertrand “<http://www.bertrand.pt>”, no qual “navegamos” pela várias categorias e extraímos a informação da primeira página de cada uma delas. Para isso acedemos primeiro à primeira página dessa secção, onde se encontram os links para as várias categorias, guardamos esses mesmos links numa colecção e de seguida percorremo-los e adicionamos os vários livros presentes em cada página, armazenados num objecto do tipo `Ebook`, a uma outra colecção.

Finalmente inserimos essa informação num ficheiro XML, recorrendo ao JDOM.

Screen-Scrapping

O *Web-Scrapping* é uma forma de *screen-scraping* que consiste na técnica de extrair informação a partir de uma interface gráfico, neste caso uma página web, e extrair a informação nela contida, simulando a recolha de informação manual. Após a recolha do conteúdo integral de cada página, procede-se ao seu *parsing* com recurso ao uso de expressões regulares. Os elementos de cada livro são isolados através da utilização de expressões regulares na `Pattern compile`, na função `getDadosEbook` da classe `RegEx`. De cada livro foi extraído: título, capa, autor, ISBN, categoria, subcategoria, formato, ano de edição, editora, pontos Bertrand, preço e moeda do preço.

Para a extracção da informação relevante sobre cada ebook teve de ser feita uma primeira análise manual à estrutura do código HTML da página. Assim, conhecendo-se os elementos da página, é possível usar expressões regulares que procurem por elementos capazes de isolar a informação que se pretende extrair. Esses elementos capazes de guiar as expressões regulares à informação pretendida podem ser pedaços de texto, classes de tags de HTML anteriores, pedaços específicos de URLs, entre outros.

Depois de extraída a informação esta é guardada num objecto do tipo `Ebook`, que é guardado em memória numa lista, para depois ser usado na criação do documento XML.

Criação do XML

Na classe `XLMLCreator` utilizamos o JDOM para interpretar objectos das aplicações em JAVA e automaticamente tratar a informação neles contidas e traduzir-la para um XML. O conteúdo de cada `Ebook` é lido e passado no JDOM, para ser estruturado no documento XML, automaticamente. Nesta representação XML dos ebooks há ainda atributos (o ISBN do ebook e a moeda do preço) e elementos com mais elementos filhos (`edicao` que inclui `anoEdicao` e `editor`, e `tema` que inclui `categoria` e `subcategoria`). Com a ajuda do JDOM foi ainda criado o namespace "`http://bertrand.pt`" e atribuído a todos os elementos do XML.

Desta forma é possível de uma forma dinâmica e intuitiva construir um documento XML a partir de objectos da aplicação JAVA.

XML schema

O XML schema (XSD) é um documento onde estão expressas as restrições dos documentos XML, para além das restrições sintácticas do XML por si.

O ficheiro XSD presente neste trabalho foi gerado automaticamente através da utilização da ferramenta Trang. O ficheiro XSD resultante foi depois modificado manualmente para ficar mais de acordo com os dados do XML.

A validade dos dados e da estrutura do XML face ao schema presente foi testada recorrendo à ferramenta online <http://tools.decisionsoft.com/schemaValidate>. O teste foi passado com sucesso, o que significa que tanto o XML como o XSD estão correctos, e que os dados presentes no XML estão de acordo com as restrições apresentadas no XSD.

XSL

Para que a informação contida no ficheiro XML pudesse ser apresentada de uma forma mais agradável para o utilizador, recorreu-se a uma stylesheet de XSL. Com essa stylesheet é possível, num web browser compatível (como por exemplo o Mozilla Firefox), visualizar a informação do ficheiro XML numa página semelhante a uma página HTML.

De modo a expandir a lista de web browsers compatíveis, gerou-se ainda um ficheiro HTML a partir dos ficheiros XML e XSL. Desta forma, é possível visualizar a informação do ficheiro XML com o aspecto definido na stylesheet XSL numa página HTML que pode ser apresentada em qualquer web browser.

No XSL foi utilizado HTML e CSS para dar um aspecto mais agradável à página HTML resultante. O aspecto final da página resultante pode ser visto na imagem seguinte.



Figura 1 - Visualização da página HTML resultante

Considerações Finais

O desenvolvimento da aplicação foi bem sucedido, tendo sido atingidos os objectivos do trabalho. Como medida de precaução descarregámos as páginas do site da Bertrand para o nosso computador, seguindo a sugestão do professor, de forma a garantidamente estar disponível, aquando da apresentação da aplicação e para não sofrer alterações imprevistas até lá. Para não se alterar a funcionalidade de extracção do HTML, colocou-se as páginas descarregadas numa pasta do servidor Apache Tomcat, de forma a se poder aceder ao seu conteúdo em localhost.

Todos os elementos do grupo estiveram envolvidos em todas as partes deste projecto. O elemento Gustavo Fernandes esteve mais envolvido na parte de descarregar o web site para o acesso local. O elemento Ricardo Lopes esteve mais envolvido na parte do descarregamento das páginas HTML, elaboração do XML schema e da extracção da informação usando expressões regulares. O elemento Rui Chicória esteve mais envolvido no XSL e sua conversão para HTML e na navegação pelas várias categorias.