Análise e Preparação de Dados da BBC News

Clara Sacramento
Faculdade de Engenharia da Universidade do Porto,
R. Dr. Roberto Frias, 4200-465 Porto
ei09090@fe.up.pt

Diogo Teixeira
Faculdade de Engenharia da Universidade do Porto,
R. Dr. Roberto Frias, 4200-465 Porto
ei09086@fe.up.pt

9 de Outubro de 2013

Resumo

A *BBC* é uma das maiores cadeias de media a nível mundial. Entre as suas variadas divisões encontra-se a *BBC News*, uma das mais importantes e prestigiadas agências noticiosas do mundo. No entanto, os serviços de pesquisa sobre os seus vastos repositórios de notícias são simplistas, e pouco personalizáveis. Pretende-se assim fazer um estudo sobre novas funcionalidades que possam ser adicionadas ao serviço, de forma a melhorar a experiência dos utilizadores. Consequentemente será também construído um repositório *offline* de notícias.

1 Introdução

Este documento descreve de uma forma sintética o trabalho de escolha, análise e caracterização de um conjunto de dados, feito até ao momento da sua escrita. Neste trabalho foram analisadas grandes quantidades de notícias, extraídas a partir das plataformas *web* da *BBC News*. Pretende-se que com base nesta análise seja possível desenvolver novas funcionalidades de pesquisa e apresentação de conteúdos, melhorando assim a experiência dos utilizadores destas plataformas.

A Secção 2 está reservada a uma breve exposição dos fatores considerados para garantir a qualidade do conjunto de dados utilizados. Na Secção 3 é apresentado o modelo conceptual que representa a organização dos dados obtidos. Mais à frente, nas Secções 4 e 5, é feita uma análise mais detalhada e técnica sobre a obtenção, processamento e armazenamento de dados. Na Secção 6 são analisadas possíveis tarefas de pesquisa a efetuar. Por fim, na Secção 7, são apresentadas as conclusões deste documento.

2 Qualidade dos Dados

O conjunto de dados usado será diretamente extraído de páginas *web* sob a alçada da *BBC News*. Uma vez que o conteúdo destas páginas é controlado na totalidade por esta última agência noticiosa, a qualidade, bem como a fiabilidade dos dados está assegurada.

É ainda importante referir a qualidade dos suportes nos quais os dados são disponibilizados, uma vez que a obtenção destes dados é feita sobretudo através de *screen scraping* de páginas *web*. Nos casos analisados até à altura de escrita deste relatório, verifica-se que os suportes disponibilizados pela *BBC News* são consistentes, e principalmente bem estruturados. Isto significa que o processo de extração de dados pode ser facilmente automatizado definindo um número reduzido de regras de extração.

3 Modelo Conceptual de Domínio

A unidade de dados usada no domínio deste serviço é a notícia. Uma notícia inclui os seguintes atributos: *id*, título, descrição, o corpo da notícia, a data de publicação, e o endereço de consulta na página da *BBC News*. Existem outros atributos adicionais de uma notícia, nomeadamente o número de vezes que a notícia foi partilhada nas redes sociais e a indicação se a notícia contém conteúdo áudio e/ou vídeo. É possível que em versões posteriores as notícias possam incluir novos atributos (ver Figura 1).

O modelo conceptual de domínio conta ainda com a noção de tópico. Um tópico representa um aglomerado de notícias que partilham uma temática. Assim sendo, um tópico pode ter inúmeras notícias a si associadas. Uma notícia pode também estar associada a mais do que um tópico.

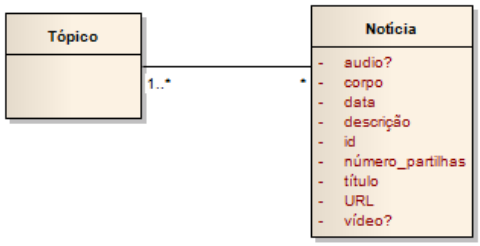

Figura 1: *Modelo conceptual de domínio.*

4 Obtenção e Processamento dos Dados

A obtenção de notícias da *BBC News* divide-se em três passos principais: interrogação à *API* disponibilizada, obtenção das páginas *web* de cada notícia, e processamento dessas páginas. A *BBC News* disponibiliza uma *API* de acesso às notícias mais recentes. É uma *API* simples, contando apenas com dois tipos de pedido *REST*, maioritariamente usada como suporte a serviços *RSS*. A aplicação de extração de dados desenvolvida obtém, através desta *API*, a lista de todas as categorias de notícia disponíveis, em formato *JSON*. Obtida e interpretada esta lista de categorias, é então possível executar um novo pedido à *API*, que retorna um vetor de notícias. Cada notícia contém título, uma breve descrição, endereço de consulta, data de publicação e imagem *thumbnail*.

A informação obtida através da *API* é importante, mas muito incompleta. Como solução a este problema, é obtida a página *web* correspondente a cada notícia no vetor previamente mencionado. Devido ao elevado número de pedidos e ao consequentemente elevado tempo de resposta, esta porção do processo foi paralelizada, de forma a reduzir o tempo de espera na obtenção dos dados.

Por fim, cada uma das páginas *web* é processada, recorrendo a várias tecnologias de pesquisa sobre documentos (seletores *CSS*, seletores *XPath* e expressões regulares). Este processamento permite não só obter o corpo de cada notícia, como também enriquecer cada uma com novas informações. Como exemplo de informações adicionais temos o grau de popularidade de cada notícia, ou a sinalização de conteúdos vídeo/áudio.

5 Armazenamento dos Dados

O processamento e obtenção de dados da *BBC News*, realizado com frequência diária, produz grandes quantidades de informação. A técnica de *screen scraping* exige uma calibração cuidadosa nesta fase inicial do trabalho, muitas vezes conseguida através da análise manual dos resultados obtidos. Assim sendo, torna-se importante armazenar a informação num formato estruturado, que possa ser facilmente entendido e analisado, tendo-se optado neste caso pelo formato *XML* (ver Figura 2).

Em versões posteriores da aplicação de extração de dados estes serão armazenados em bases de dados relacionais. Esta mudança tem como objetivos facilitar a consulta da informação através de interrogações em *SQL*, e proporcionar o armazenamento da informação de uma forma centralizada. Será também implementado um conversor que permitirá transferir toda a informação armazenada nos ficheiros *XML* para uma base de dados relacional, de forma automática.

6 Tarefas de Pesquisa Consideradas

Para este trabalho foram consideradas quatro grandes vertentes de pesquisa sobre o conjunto de dados estudado: pesquisa textual simples, pesquisa por imagem, refinamento de pesquisa e organização de resultados.

A pesquisa textual consiste simplesmente na busca de uma ou várias palavras no texto indexado, nomeadamente no título, descrição ou corpo da notícia. Esta é a pesquisa mais comum, e está já disponível através das plataformas da *BBC News*.

Foi também estudada a possibilidade de utilizar pesquisa por imagem, quer por comparação direta entre imagens, quer por indexação de atributos de imagens através do modelo *bag-of-words* [1]. Esta opção foi no entanto posta de parte, uma vez que requer um grande esforço de implementação, muito para além do domínio deste trabalho. Um dos pontos principais do trabalho é a possibilidade de refinar a pesquisa textual através de vários atributos. Até este momento foi estudada a filtragem por data de publicação, por tópico, por exclusão de campos usados para pesquisa textual e por artigos com conteúdos áudio e/ou vídeo.

```
<?xml version="1.0" encoding="UTF-8"?>
<bulletin>
   <topic id="topic">
      <news id="id" video="true" audio="false">
         <title>Simple Title</title>
         <description>Informative description.</description>
         <url>http://www.bbc.co.uk/news/topic-example-id</url>
         <date>1381217048</date>
         <body>Article's full body.</body>
      </news>
   </topic>
</bulletin>
```

Figura 2: *Estrutura típica de armazenamento dos dados.*

Por fim, foram ainda estudadas possibilidades de ordenação de resultados. Para além da ordenação por data, será possível a organização por "popularidade", sendo que um artigo é tanto mais popular quantas mais vezes for partilhado nas redes sociais.

Mais tarde, com posterior análise às páginas web obtidas, poderá ser possível encontrar novos métodos de organização ou refinamento de pesquisas.

7 Conclusões

Espera-se que a ferramenta produzida em seguimento deste estudo seja realmente útil, e que enriqueça a experiência dos utilizadores das plataformas web de informação da *BBC News*. Para além das funcionalidades já descritas neste documento, existe ainda espaço para acrescentar novos métodos de pesquisa e refinamento, à medida que forem identificadas novas propriedades do conjunto de dados analisado.

Referências

[1] Wikipédia. Bag-of-words model in computer vision, 2013. [Online; acedido a 8 de Outubro de 2013].