The Wikipedia XML Corpus

Ludovic Denoyer and Patrick Gallinari

Laboratoire d’Informatique de Paris 6
8 rue du capitaine Scott
75015 Paris
{ludovic.denoyer,patrick.gallinari}@lip6.fr
http://www-connex.lip6.fr/denoyer/wikipediaXML

Abstract. This article presents the general Wikipedia XML Collection developed for Structured Information Retrieval and Structured Machine Learning. This collection has been built from the Wikipedia Encyclopedia. We detail particularly here which parts of this collection have been used during INEX 2006 for the Ad-hoc track and for the XML Mining track. Note that other tracks of INEX - multimedia track for example - have also been based on this collection.

1 Introduction

Wikipedia\(^1\) is a well known free content, multilingual encyclopedia written collaboratively by contributors around the world. Anybody can edit an article using a wiki markup language that offers a simplified alternative to HTML. This encyclopedia is composed of millions of articles in different languages.

Content-oriented XML retrieval is an area of Information Retrieval (IR) research that is receiving an increasing interest. There already exists a very active community in the IR/ XML domain which started to work on XML search engines and XML textual data. This community is mainly organized since 2002 around the INEX initiative (INitiative for the Evaluation of XML Retrieval) which is funded by the DELOS network of excellence on Digital Libraries.

In this article, we describe a set of XML collections based on Wikipedia. These collections can be used in a large variety of XML IR/Machine Learning tasks like ad-hoc retrieval, categorization, clustering or structure mapping. These corpora are currently used for both, INEX 2006\(^2\) and the XML Document Mining Challenge\(^3\). The article provides a description of the corpus.

The collections are downloadable on the website:


---

\(^{1}\) http://www.wikipedia.org
\(^{2}\) http://inex.is.informatik.uni-duisburg.de/2006
2 Description of the Corpus

The corpus is composed of 8 main collections corresponding to 8 different languages: English, French, German, Dutch, Spanish, Chinese, Arabian and Japanese. Each collection is a set of XML documents built using Wikipedia and encoded in UTF-8. In addition to these 8 collections, we also provide different additional collections for other IR/Machine Learning tasks like categorization and clustering, NLP, machine translation, multimedia IR, entity search, etc.

2.1 Main Collections

The main collections are a set of XML files in 8 different languages. The table gives a detailed description of each collection.

<table>
<thead>
<tr>
<th>Collection name</th>
<th>Language</th>
<th>Number of documents</th>
<th>Size of the collection (MegaBytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>main-english</td>
<td>English</td>
<td>659,388</td>
<td>≈ 4,600</td>
</tr>
<tr>
<td>20060130_french</td>
<td>French</td>
<td>110,838</td>
<td>≈ 730</td>
</tr>
<tr>
<td>20060123_german</td>
<td>German</td>
<td>305,099</td>
<td>≈ 2,079</td>
</tr>
<tr>
<td>20060227_dutch</td>
<td>Dutch</td>
<td>125,004</td>
<td>≈ 607</td>
</tr>
<tr>
<td>20060130_spanish</td>
<td>Spanish</td>
<td>79,236</td>
<td>≈ 504</td>
</tr>
<tr>
<td>20060303_chinese</td>
<td>Chinese</td>
<td>56,661</td>
<td>≈ 360</td>
</tr>
<tr>
<td>20060326_arabian</td>
<td>Arabian</td>
<td>11,637</td>
<td>≈ 53</td>
</tr>
<tr>
<td>20060303_japanese</td>
<td>Japanese</td>
<td>187,492</td>
<td>≈ 1,425</td>
</tr>
</tbody>
</table>

Each collection contains a set of documents where each filename is a number corresponding to the id of the file (for example: 15243.xml). Each id is unique and each file corresponds to an article of Wikipedia. We only kept articles and removed all the wikipedia pages corresponding to “‘Talks’”, “‘Template’”, etc. Each file is an UTF-8 document which is created from the wikitext of the original article. Figure gives an example of an English article extracted from the corpus.

Tag labels. We introduced different tags in order to represent the different parts of a document. We distinguish two types of tags:

- The general tags (article, section, paragraph, etc.) that do not depend on the language of the collection. These tags correspond to the structural information contained in the wikitext format (for example: == Main part == is transformed into <title>Main part</title>)
- The template tags (template, infobox, etc.) represent the information contained into the wikipedia templates. Wikipedia templates are used to represent a repetitive type of information. For example, each country described into wikipedia starts with a table containing its population, language, size, etc.

\[^4\] Some additional languages will be added during the next months.