XRCE Participation to the 2009 Book Structure Task

Hervé Déjean and Jean-Luc Meunier

Xerox Research Centre Europe
6 Chemin de Maupertuis, F-38240 Meylan
Firstname.lastname@xrce.xerox.com

Abstract. We present here the XRCE participation to the Structure Extraction task of the INEX Book track 2009. After briefly explaining the four methods used for detecting the book structure in the book body, we explain how we composed them to address the book structure task. We then discuss the Inex evaluation method and propose another measure together with the corresponding software. We then report on each individual method. Finally, we report on our evaluation of the results of all participants.

1 Introduction

We present in this paper our participation to the Structure Extraction task of the INEX Book 2009. Our objective was to experiment with the use of multiple unsupervised methods to realize the task. This article will therefore briefly describe each of them before focusing on the evaluation of our results as well as those of the other participants. We use here the metric we proposed in 2008, whose software implementation is now available at: http://users.info.unicaen.fr/~doucet/StructureExtraction2009/AlternativeResults.html.


This work is supported by the Large Scale Integrating Project SHAMAN, cofunded under the EU 7th Framework Programme (http://shaman-ip.eu/shaman/).

2 Pre-processing

The first step simply consists in reformatting the XML INEX format into our internal format, mostly renaming elements and adding some attributes (such as unique IDs). This was performed using XSLT technology.

A second step consists in detecting page headers and footers that often introduce noise for our table of contents detector (see [1]).

A third step consists in recognizing the page numbering of the document (see [3]), in order to associate each physical page with zero or one logical page number, the latter being a piece of text. This is again an unsupervised method.
3 The Four Methods

The first method aims at parsing the ToC page so as to segment it into entries, each entry being formed by a label and a reference to a page number. Our second method is dedicated at parsing an index page. The third method is our classical Toc detection method (see [2]). The fourth method uses some characteristic page layout in the document body.

None of these methods aims at determining the entry level, so it is arbitrarily set to 1.

3.1 Parsing the ToC Pages

Parsing the ToC pages involves first finding them. For this purpose, we tried first with a simple heuristic that consists in looking for the keyword ‘contents’ in the few first lines of each page, under the hypothesis of the presence of one ToC, possibly split over consecutive pages at the beginning or end of the book. We look for a series of pages containing this word and tolerate a certain number of misses.

Somehow to our own surprise, this method led to a F1, in term of ToC page retrieval task, in the range 90-95% over the 2008 INEX dataset. However, retrieving the page of the ToC of a document is not enough.

We then need to parse the contiguous pages deemed to be the ToC. The segmentation into paragraphs is unfortunately not adequate, since a valid paragraph may both correspond to multiple entries, as shown below, or to part of one entry. In the latter case, one could argue that the paragraph segmentation is wrong but it is so.

![Fig. 1. Excerpt from the 2008 book #3 (0008D0D7B18E65AD)](image)

We decided to use the reference to page numbers as ToC entry segmenter, as already mentioned in the literature [4, 5, 6]. Once the ToC is segmented, the page that is referred to is considered as the target of the entry and the entry text becomes the title in the result. Here we take benefit of the recognition of the page numbering, performed in the pre-processing step, and which associates each physical page with zero or one logical page number.

3.2 Parsing the Index Pages

In a similar way, it is possible to look for some index pages, using the ‘index’ keyword, and to segment it based on the appearance of a page number at the end of the lines. The text in-between two page numbers is deemed to be the title of some part of the document, starting at the indicated page number. This step is performed due to