Digital Libraries and Historical Document Processing

Apostolos Antonacopoulos
Simone Marinai

Overview

The tutorial will cover the background issues, challenges and opportunities in the analysis of historical documents and their applications in the Digital Library world. The tutorial is divided in four parts.
The first part proposes a survey of the aims of Digital Libraries and their technical evolution over the years. Potential applications of Document Image Analysis techniques in the field of DLs are pointed out and discussed with the audience. The use of open source software for building simple digital library infrastructures is analyzed as well.
The second part starts with an examination of the different motivation and other institutional factors that influence technical decisions. The types of documents typically encountered are discussed next with the challenges and possibilities they offer for digitization and full-text conversion. Focusing on the needs of major libraries, the remainder of this part presents in detail the different stages in full-text conversion. In each of the stages (scanning, image enhancement, segmentation, OCR and postprocessing) the challenges and possibilities for improvement are examined.
The third part deals with recent advances in Document Image Retrieval (DIR) and potential applications of these techniques in the field of Digital Libraries. Two main paradigms will be described: retrieval by layout similarity and text-based retrieval. These techniques will be explained also through the demonstration of a DIR system specifically developed for the use on DL-related documents.
The fourth part of the tutorial comprises a more technical description of the state-of-the-art in the analysis of historical documents. Major past and current initiatives will be mentioned and individual methods will be described for each stage in the processing, analysis and recognition of historical documents. Finally, as an essential aspect in measuring and making progress, ways of performance evaluation of historical document analysis methods will be presented.

Outline

Introduction
Digitization Approaches. Differences between libraries and archives, showcase vs. mass digitization, preservation vs. searchability.

PART A
1. Libraries, Digital Libraries and the Web.
2. DL user interfaces - the DL from the user perspective.
3. DL architectures - the DL from inside.
4. Information retrieval in DLs.
5. Open source software for DL building: the Greenstone example.

PART B
1. Documents. Differences through the centuries. Handwritten and printed. Manuscripts, books and newspapers.

2. Full-text conversion workflow. Stage-by-stage description of processing, analysis and recognition.

1. Scanning: scanning options, library processes.
2. Image enhancement: different types of artefacts, challenges.
3. Layout analysis / segmentation: challenges.
4. OCR: challenges and approaches.
5. Post- processing: dictionaries, automated and manual correction approaches.

PART C

1. Information retrieval in Digital Libraries.
2. Layout-based document image retrieval.
3. Word-based document image retrieval.
4. The AIDI system.

PART D

1. State-of-the-art in historical document processing.
1. Projects: past and current.
2. Methods and examples for each stage.
3. Performance evaluation.