Explorative Analysis and Visualization of Large Information Spaces

Automatic Information Extraction from Textual Documents

Andreas Stoffel

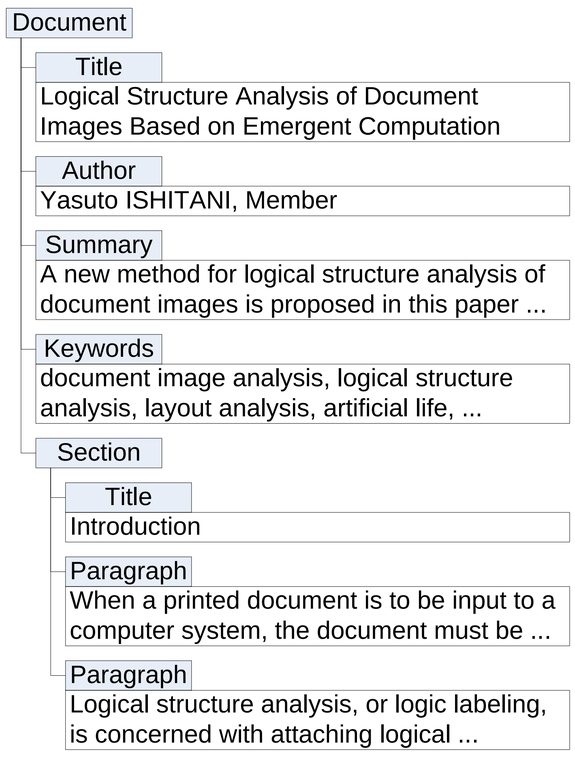
Universität Konstanz

1 Background

Information are usually exchanged with textual documents, e.g. papers, newspapers, books, ...
A special problem is the information extraction from documents like shipping orders or parts lists:

- These type of documents contain similar content like shipper, recipient, items, ...
- The layout of the document differs
- The information is hidden in the structure

(a) The physical layout of a document.



(b) The logical structure of a document.

Figure 1: Example of the physical layout and logical structure [1] of a document. The physical layout describes where a part of a document is placed on the page. The logic structure subdivides a document into logical units. The logical structure can have different granularities, for instance the references in a paper can be labeled as "paragraph" or as "reference".

The challenges of document structure recognition are:

- How to efficiently integrate the user into the learning phase?
- How to identify the best set of features for the structure recognition of a particular document type?
- How to efficiently and effectively visualize the classification results?

2 Research Framework and Approach

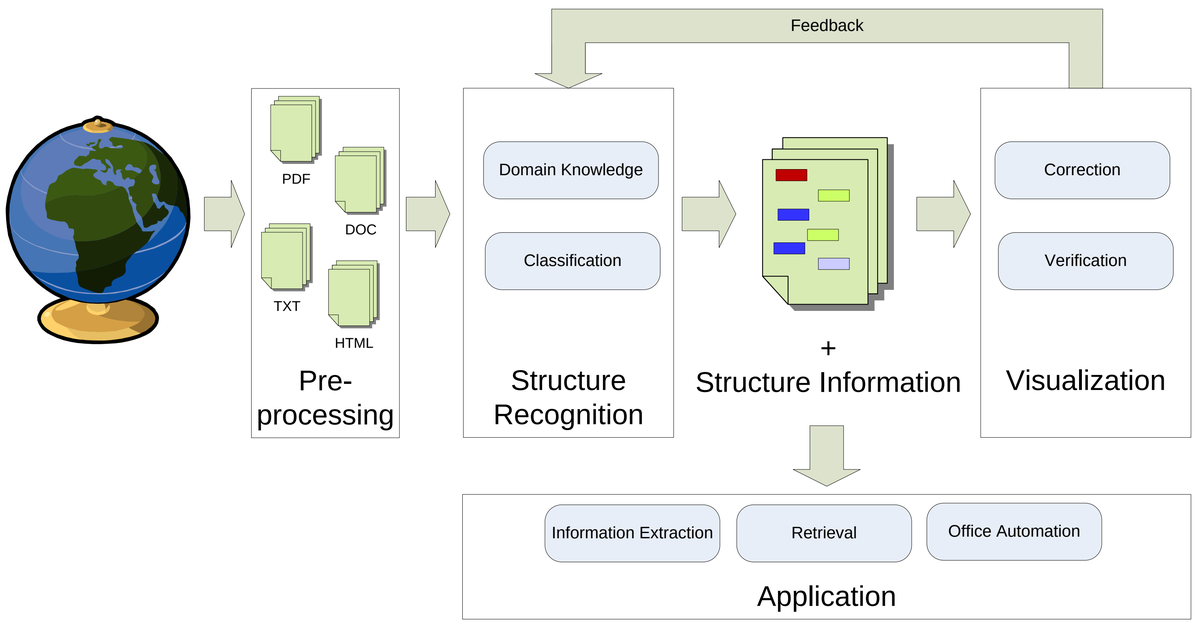
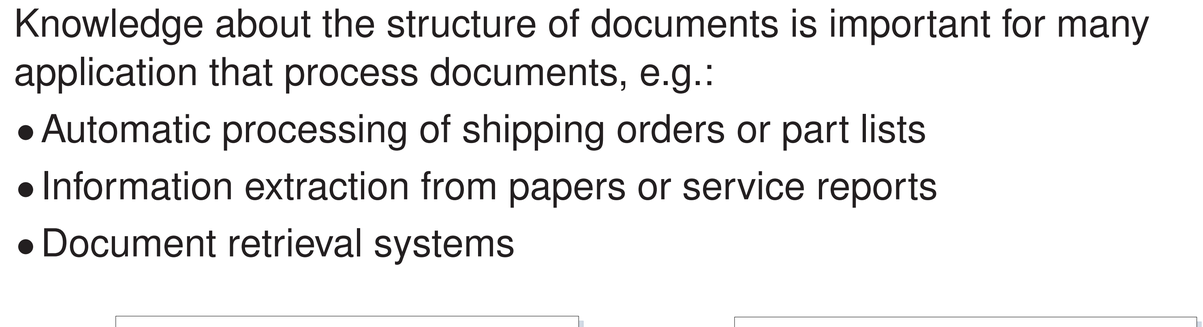


Figure 2: The pipeline of interactive structure recognition.

The structure recognition evaluates keywords, geometry and formatting properties:



(a) Spacing information.

(b) Font information.

Figure 3: Examples of the visual properties used to extract the structural elements from documents. Beside the spacing and font information, also indention from left and right and geometric information, like position on the page, are used to capture the visual impression of the elements.

The challenge of the structure recognition is that the level of detail and the types of structural elements are application specific. A solution for one application problem may not be applicable in other cases.

3 Applications

Knowledge about the structure of documents is important for many application that process documents, e.g.:

- Automatic processing of shipping orders or part lists
- Information extraction from papers or service reports
- Document retrieval systems

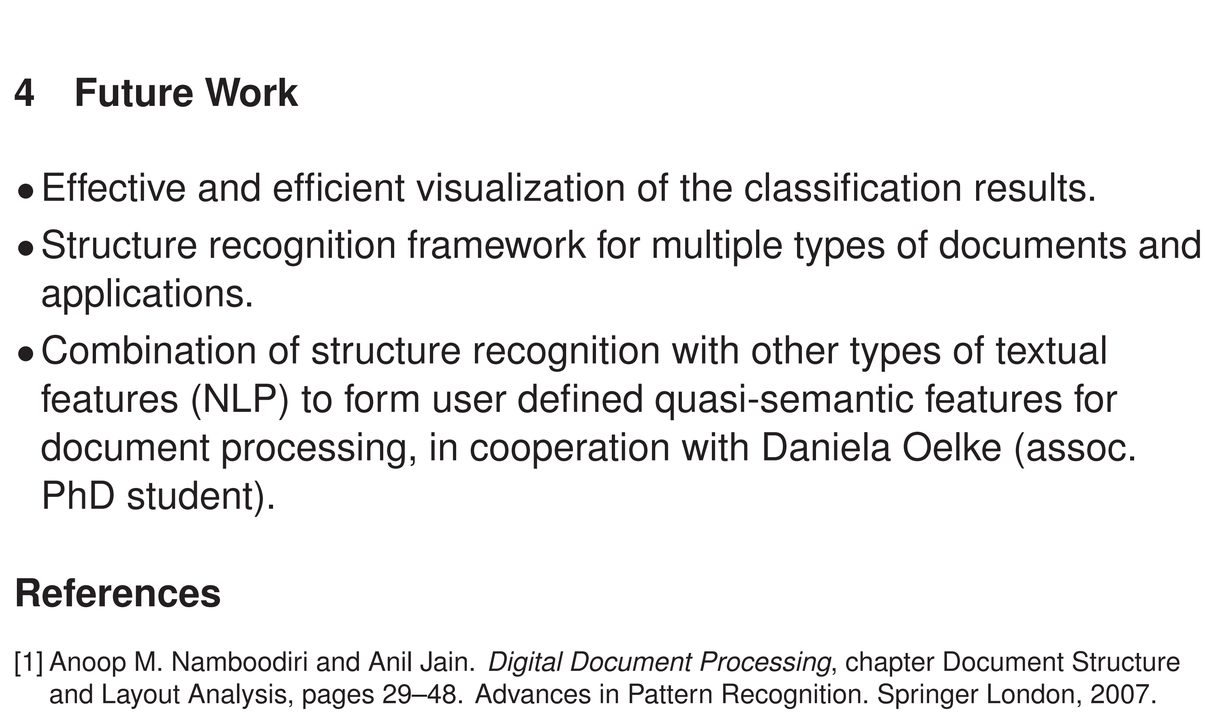


Figure 4: Two pages of a paper with highlighted structural elements. In this example the abstract and the references are recognized, but the introduction and the conclusion are treated as ordinary sections. For applications, where the introduction and the conclusion play an important role, the extraction process must be changed.

4 Future Work

- Effective and efficient visualization of the classification results.
- Structure recognition framework for multiple types of documents and applications.
- Combination of structure recognition with other types of textual features (NLP) to form user defined quasi-semantic features for document processing, in cooperation with Daniela Oelke (assoc. PhD student).

References

[1] Anoop M. Namboodiri and Anil Jain. *Digital Document Processing*, chapter Document Structure and Layout Analysis, pages 29–48. Advances in Pattern Recognition. Springer London, 2007.