Project D.A.M.A.: Document Acquisition, Management and Archiving

Michelangelo Ceci, Corrado Loglisci, Stefano Ferilli, and Donato Malerba

Department of Computer Science, University of Bari "Aldo Moro"
{ceci,loglisci,ferilli,malerba}@di.uniba.it

Abstract. A paper document processing system is an information system component which transforms information on printed or handwritten documents into a computer-revisable form. In intelligent systems for paper document processing this information capture process is based on knowledge of the specific layout and logical structures of the documents. In this project we design a framework which combines technologies for the acquisition and storage of printed documents with knowledge-based techniques to represent and understand the information they contain. The innovative aspects of this work strengthen its applicability to tools that have been developed for building digital libraries.

1 Introduction and Motivation

The large and increasing amount of paper documents to be processed daily demands new document management systems that are able to catalogue and organize them automatically based on the semantics of their contents. Personal document processing systems that can provide functional capabilities of classifying, storing, retrieving, and reproducing documents, as well as extracting, browsing and synthesizing information from a variety of documents are in ever-growing demand. However, they operate on electronic documents and not on paper ones. This is the focus of the Document Image Analysis (DIA) area, which investigates the theory and practice of recovering the symbol structure of digital images scanned from paper or produced by computer, and hence results in the conversion of document images to symbolic form for modification, storage, retrieval, reuse and transmission. This conversion is a complex process articulated into automatic and semi-automatic stages which we have explored in the project "D.A.M.A. Document Acquisition, Management and Archiving", funded by the Data Service S.p.A. company (Mantova, Italy).

2 Scientific Challenges

Document Image Analysis deals with the recognition of logically and semantically relevant components in the layout extracted from a document image. This opens several challenges. The representation of recognized and extracted information into some common data format is a key issue. A solution to this problem

M. Agosti et al. (Eds.): IRCDL 2011, CCIS 249, pp. 115–118, 2011.

can come from the XML technology. XML has been proposed as a data representation format in general, but it was originally developed to represent (semi-) structured documents, therefore it is a natural choice for the representation of the output of DIA systems. XML is also an Internet language, a feature that can be profitably exploited to make information drawn from paper documents more quickly web-accessible and retrievable than distributing the bitmaps of document images on a Web server. Moreover, it is possible to define hypertext structures which improve document reading. Finally, in the XML document, additional information on the semantics of the text can be stored in order to improve the effectiveness of the retrieval. This is a way to reduce the so-called semantic gap in document retrieval, which corresponds to the mismatch between a user's request and the way automated search engines try to satisfy it.

The extraction of semantics from the document image requires knowledge-based technologies, which offer various solutions to the knowledge representation problem and automated reasoning, as well as to the knowledge acquisition problem by means of machine learning techniques. The importance of knowledge technologies has led to the proliferation of machine learning and data mining methods which, especially with classification approaches, provide suitable tools for the recognition of components and understanding of the content.

The representation formalism used in these classification approaches is another issue constantly discussed. The spatial dimension of page layout makes formalisms used in inductive logic programming and multi-relational data mining the most suitable candidate for modeling documents in DIA.

3 Contribution by the Research Group

In the project, we have designed a framework of tools to integrate information based on the understanding of the document content. This framework offers functionalities to digitize paper documents, acquire them as document images and interpret them with sophisticated and intelligent techniques. Tools for understanding the content of documents allow to integrate structured, semi-structured and unstructured information stored in different repositories. Among the acquisition functionalities, the possibility to extract atomic pieces of information and the mechanisms of contextualization permit to simplify the data entry operation and to reduce human intervention. Here we describe some of the functionalities of the framework which are available in the prototype system IDIS [1,3].

Document image analysis is performed through a process composed by the preprocessing of the raster image of a scanned paper document, the segmentation of the preprocessed raster image into basic layout components, the classification of basic layout components according to the type of content (e.g., text, graphics), the identification of a more abstract representation of the document layout (layout analysis), the classification of the document on the ground of its layout and content, the identification of semantically relevant layout components, the application of OCR only to textual components and the storing in XML format providing additional information on the semantics of the text (Figure 1).

In particular, initial processing steps include binarization, skew detection and noise filtering. Before the actual interpretation of text data takes place, graphic data that are present in the digitized document must be separated from the text so that subsequent processing stages may operate exclusively on textual information. The separation of text from graphics is performed in two steps: image segmentation and block classification. The former is the identification of rectangular blocks enclosing content portions while the latter aims at discriminating blocks enclosing text from blocks enclosing graphics (pictures, drawings, ...). In order to facilitate subsequent document processing steps, it is important to classify these blocks according to the type of content: text block, horizontal line, vertical line, picture (i.e., halftone images) and graphics (e.g., line drawings). The classification of blocks is performed by means of a decision tree automatically built from a set of training examples (blocks) of the pre-defined classes. The result of the segmentation process is a list of classified blocks, corresponding to printed areas in the page image. These blocks are processed in order to detect structures among them by means of layout analysis techniques. In IDIS, we integrate an hybrid approach composed by a global analysis technique, which determines possible areas containing paragraphs, sections, columns, figures and tables, and a local analysis technique, which groups together blocks that possibly fall within the same area, called frames. The result is a hierarchy of abstract representations of the document image, the geometric (or layout) structure. The leaves of the layout tree (lowest level of the abstraction hierarchy) are the blocks, while the root represents the whole document.

After having detected the layout structure, the logical components of the document, such as title, authors, sections of a paper, can be identified. The logical components can be arranged in another hierarchical structure, which is called logical structure. The logical structure is the result of repeatedly dividing the content of a document into increasingly smaller parts, on the basis of the human-perceptible meaning of the content. The leaves of the logical structure are the basic logical components, such as authors and title. The heading of an article encompasses the title and the author and is therefore an example of composite logical component. Composite logical components are internal nodes of the logical structure. The root of the logical structure is the document class (e.g. 'scientific paper', 'letter' or 'censorship card').

The problem of finding the logical structure of a document can be cast as the problem of associating some layout components with a corresponding logical component [2]. In IDIS, this mapping is limited to the association of a page with a document class (document classification) and the association of frames with basic logical components (document understanding). The first stage is that to label the pages with a document model (class). This operation is performed by exploiting an inductive logic programming approach of supervised classification which permits to represent documents and models: the class of a document is identified through a positive matching test between logic formulas. A second stage of classification allows to recognize logical components. This step is generally performed by encoding a knowledge base that defines the mapping from

the layout to the logical structure. In IDIS, we integrate a machine learning approach which creates a knowledge base by means of an inductive process that learns rules in first-order logic from layout information of manually classified documents. The matching between these rules and the description of the document layout determines the recognition of logical components. Another stage of classification is done to understand document images by recognizing semantically layout components (e.g., 'title', 'authors' in a scientific paper). In this case we consider also textual features besides of layout information as done in the previous stage and as classifier we integrate the well established Support Vector Machine. The result of document processing is stored in XML format so to include semantic information extracted in the document analysis and understanding processes and make it accessible via web technologies [2].

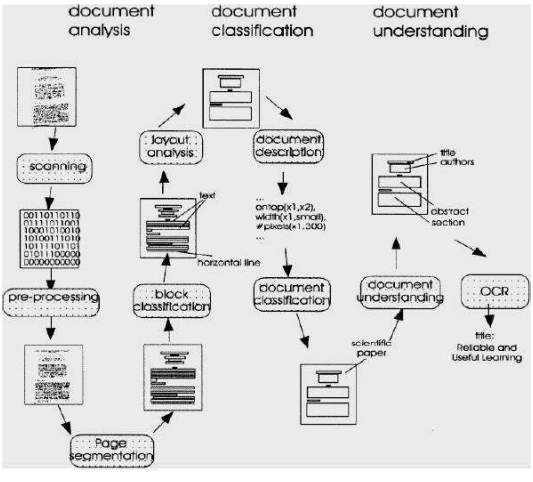

Fig. 1. Intelligent Document Interpretation Framework

References

1. Ceci, M., Berardi, M., Malerba, D.: Relational data mining and ILP for document image understanding. Applied Artificial Intelligence 21(4&5), 317–342 (2007)
2. Esposito, F., Malerba, D., Semeraro, G., Ferilli, S., Altamura, O., Basile, T.M.A., Berardi, M., Ceci, M., Di Mauro, N.: Machine learning methods for automatically processing historical documents: From paper acquisition to XML transformation. In: DIAL, pp. 328–335. IEEE Computer Society (2004)
3. Malerba, D., Ceci, M., Berardi, M.: Machine learning for reading order detection in document image understanding. In: Marinai, S., Fujisawa, H. (eds.) Machine Learning in Document Analysis and Recognition. SCI, vol. 90, pp. 45–69. Springer, Heidelberg (2008)