Discovering Logical Structures in Digital Documents

Floriana Esposito, Stefano Ferilli, Teresa M.A. Basile, and Nicola Di Mauro

Università di Bari, Dipartimento di Informatica
via E. Orabona, 4 I-70126 Bari, Italia
{esposito, ferilli, basile, nicodimauro}@di.uniba.it

Abstract. Document management is critical for the distribution and preservation of knowledge. The aim is discovering, in a database of documents in paper, electronic and Web pages format, significant knowledge to be used as meta-information for their content-based retrieval and management. This paper proposes processing solutions that are suitable for application in the three cases, all of them exploiting symbolic (first-order) learning techniques for automatically classifying the documents and their layout components according to their semantics. This will allow to properly tag the documents in a Semantic Web development perspective.

1 Introduction

Document management has always been a fundamental issue for the distribution of knowledge and for its preservation in time. Up to a few decades ago, most of the available material consisted in printed paper documents, often in few copies stored in archives and libraries, which represented a serious obstacle to wide access and distribution of the information content they bear. More recently, the great diffusion of computers on one side, and of the Internet on the other, caused a significant migration of document formats from the paper support to the digital one, which greatly facilitates copy, distribution and accessibility of the document themselves. Many such documents are spread throughout the World Wide Web in the most diverse websites, and a whole research area centered on principles and techniques for setting up and managing document collections in the form of Digital Libraries has started and suddenly developed. Moreover, not only typeset documents in digital format are available, but the Internet pages can be considered documents that convey interesting and important information to be collected and managed, as well.

The question is, given databases containing documents of the above mentioned kinds (scanned images of paper documents, documents in electronic format and Web pages), how it is possible to discover among them useful knowledge to be used as meta-information for their content-based retrieval and management. The three kinds of material referred to above pose, of course, different problems for their effective handling, which in turn require the development of different approaches and techniques that are able to deal with their peculiarities and characteristics.

In the following, a number of approaches will be proposed for these kinds of documents. Section 2 presents techniques devoted to obtain a digital counterpart of the original paper documents, enriched with information related to their geometrical and logical structure. Section 3 shifts to the case of documents already available in digital format with the aim of obtaining information related to their layout structure. Then, Section 4 discusses a proposal on how the specific case of Web pages can be handled. Finally, Section 5 concludes the paper and outlines future work directions.

2 Discovering Logical Structures in Digitized Paper Documents

The large amount of legacy documents available in paper format explains the significant portion of research devoted to them [8]. Tools for automated intelligent processing of paper documents have been developed at the LACAM laboratory of the University of Bari in the last decade. To this purpose, different kinds of paper documents were taken into account, ranging from commercial letters to scientific papers and, lately, cultural heritage material. The design and implementation of components for the automatic detection of the layout structure of documents is a preliminary step to their automatic interpretation on the ground of such a structure, in order to annotate each document component according to its logical/semantic meaning. It involves a number of related activities, such as designing a proper representation of the document structure and successively training the component for document analysis by means of sample documents.

Layout analysis is carried out by WISDOM++ [3], a knowledge-based system which is characterized by an extensive use of Machine Learning methods applied to infer automatically the rules for performing the various processing steps. Preprocessing starts from scanned color images of the documents, transforms them into black-and-white 300dpi images (much less computationally demanding), evaluates and removes skew, and estimates a complexity factor of the layout. Then, the global analysis determines possible areas containing sections, figures, tables etc. Specifically, the preprocessed image is segmented, the segments (blocks) are classified in order to separate text from graphics, the layout is automatically analyzed and the result can be manually corrected. Finally, the local analysis groups together blocks that may fall within the same area. The result of the local analysis depends on the quality of the global analysis, and in turn strongly influences the result of the next steps' accuracy.

Once the layout structure has been found for a set of training documents, learning tools can be exploited to induce rules for the automatic *classification* of documents on the ground of spatial and perceptual factors. After that, the logical components of the document could be identified. Such components can be arranged in another hierarchical structure, called logical structure, on the basis of the human-perceptible meaning of the content. In the logical structure, the leaves are the basic logical components, composite logical components are internal nodes, and the root is the document class. 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 (document *understanding*). This way, the low-level image feature space (based on geometrical and textural features) can be replaced with a higher-level semantic space. Since the kind of logical components that can be found in a document depends on the class of the document at hand, document classification precedes document understanding.

The problem of learning and incrementally refining the set of first-order logic rules for document classification and understanding when new observations are available is in charge of the learning system INTHELEX [2], that embeds efficient and effective operators for refining the theories in its search space. Indeed, incremental learning is particularly suitable for environments that are subject to continuous updating, as a document collection is. A set of documents of interest must be previously preprocessed and segmented for obtaining their layout structure, and then tagged by experts as regards both their class and the meaning of their significant layout components for use in the training phase. A suitable language has been developed in order to describe the documents according to their size and the type, size and relative position of their layout blocks.

A peculiar challenge of paper documents is the possibly considerable amount of noise in their description. The layout quality can be affected by manual annotations, stamps that overlap to sensible components, ink specks, etc. As to the layout standard, many documents may be typewritten sheets, that consist of all equally spaced lines in ancient typeface. The multistrategy operators embedded in INTHELEX may be of help. Deduction can be exploited to fill observations with information that is implicit in their description, thus improving their representation. Abduction aims at completing possibly partial information in the examples (adding more details), and thus can make the system more flexible in the absence of particular layout components due to the typist's style. Abstraction removes superfluous details from the description of both the examples and the theory and hence can help in focusing on more meaningful layout patterns.

The average predictive accuracy of theories learned by INTHELEX when applied to various real-world document image classification and understanding problems, and specifically to documents from historical archives[1], is almost always above 90%, with peaks of 99,17%. A comparison to another learning system was encouraging. Due to lack of space, the experimental set-

[1] EU IST-1999-20882 project COLLATE (Collaboratory for Annotation, Indexing and Retrieval of Digitized Historical Archive Material), URL: `http://www.collate.de`

ting and results cannot be reported here in detail. A complete discussion can be found in [2,6,4].

As a side effect of document understanding, OCR can be applied only to interesting textual components whose content type has already been determined, thus avoiding useless effort, and XML (eXtensible Markup Language) tags can be associated to meaningful components, to represent their semantic role. The XML version of the documents can be stored in a metadata repository, and used to perform higher level queries for retrieval purposes. Moreover it can be exploited to render the documents in the form of HTML pages that are faithful with respect to the original and that can be accessed via a web-browser.

3 Discovering Logical Structures in Documents Available in Digital Format

Transposing the results obtained with paper documents to the case of documents that are already available in digital format has of course a straightforward motivation. This section presents a novel technique for extracting the layout structure from documents in digital format. A prototype system embedding such a technique was implemented, and is under optimization for effectiveness and efficiency improvement. In the following, we will refer to *digital documents* as to documents that are available in some standard typesetting format, such as PostScript (PS) or Portable Document Format (PDF). This obviously encompasses formats of other proprietary word processing tools, such as Microsoft Word (.doc), due to the existence of translators that can easily and immediately export them in the above formats. On the contrary, images of documents represented in graphical formats will be excluded, since they don't carry explicitly any actual document-related information, but consist of just a set of pixels. In these cases, the techniques presented in the previous section are more suitable.

What changes when handling digital documents, with respect to paper ones, is obviously the transition from their original representation to a first-order logic description of their layout structure at different levels. Then, the next steps of document classification and understanding can be carried out in exactly the same way as before. Starting from an initial format that already embeds information related to the document organization clearly makes the pre-processing step easier, since no noise is present (introduced by the scanning phase or due to bad preservation of the original). On the other hand, one has the opportunity to focus on devising and applying new algorithms in order to go beyond the representation of the layout components as just isolated rectangles, and are able to identify also a number of shapes and layout that are frequently used in typesetting styles, such as pictures surrounded by text. In such a case, forcing the layout analyzer to cast each block as a rectangle with white space all around would yield one that includes both the

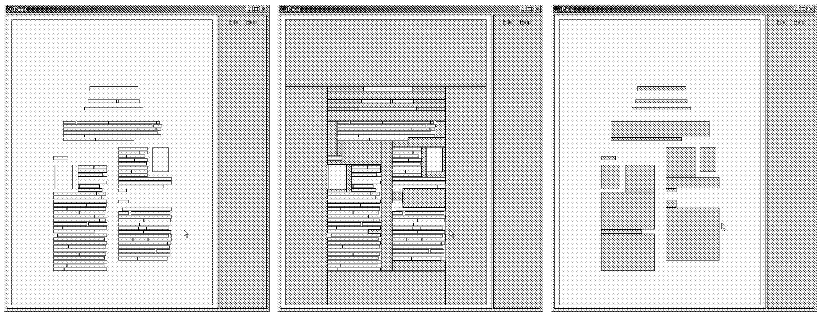

Fig. 1. Stages in electronic document processing

text and the picture. Conversely, a more flexible representation would enter into more details and split complex shapes into smaller (rectangular) adjacent blocks, each of a single type, expressing their relationships within the larger (isolated) rectangle.

Figure 1 shows the processing stages carried out by this new approach on a document in which text surrounds two pictures. Here it is clear that, unless the white spaces separating the figures from the text are identified, there is no way of splitting the two components from the rectangular frame enclosing both of them. By analyzing the PS or PDF source, it is possible to identify the basic blocks that make up the document. Often such blocks correspond to fragments of words, so that a first aggregation based on their overlapping or adjacency is needed in order to obtain blocks surrounding whole words. A further aggregation could be performed, based on proximity, to have blocks that group words in lines. In either case, the next step must aim at finding significant blocks of content while being able to separate blocks belonging to different flows of information or containing different information, such as the pictures in the present case. This task can be cast as finding all white (background) spaces between these blocks, and then reconstructing the printed areas surrounded by such spaces. To identify background pieces inside the document, a variant of the algorithm reported in [1] can be exploited, with a threshold on the dimensions and/or area of the spaces to be found that avoids extraction of non-significant backgrounds (such as spaces between lines). Then, given the background, it is possible to compute its complement, thus obtaining the desired output.

When computing the complement, two levels of description are generated. The former refers to single blocks filled with the same kind of content, the latter consists in rectangular frames each of which may be made up of many blocks of the former type. Thus, the overall description of the document to be exploited by the learning system includes both kinds of objects, plus information on which frames include which blocks and on the actual spatial relations between frames and between blocks in the same frame (e.g., above, touches, etc.). This allows to maintain both levels of abstraction independently, until some evidence forces to drop any of them. As an example, a fragment describing the highlighted block in the right screenshot of Figure 1 could be (omitting the attributes of each object and focusing on the relationships among them only):

. . . , part_of(doc,frame1), part_of(doc,block11), part_of(doc,block12), part_of(doc,block13), part_of(doc,block14), contains(frame1,block11), contains(frame1,block12), contains(frame1,block13), contains(frame1,block14), above(block11,block12), touches(block11,block12), above(block12,block13), touches(block12,block13), above(block14,block12), to_right(block14,block11), . . .

Again, a Machine Learning system can be applied to a set of tagged descriptions obtained in this way, in order to infer classification and understanding rules. Then, the text contained in significant components can be extracted directly from the digital source. Since such a text is split into a number of small blocks, this obviously requires collecting all such blocks that fall in the given components, finding their correct sequence in order to reconstruct the text flow, and appending all of them in a single text according to such an ordering.

4 Is it Possible to Discover Logical Structure in Web Pages?

The problem may be cast as a learning problem, i.e.: is it possible to discover from different instances of Web pages referred to the same field or application area a relation between content and spatial organization? This involves an easier pre-processing task, since the layout structure of the document is already explicit in the HTML tags contained in the page code, and can be recognized by suitable parsers. A proposal for representing Web pages by means of clauses is the following:

class(Page) :- title(Page,Title), *title_description*(Title), *page_description*(Page)
where predicates

`class(P)` Page P belongs to class *class*
`title(P,T)` T is the title of page P

are fixed (in fact, *`class`* is a meta-predicate). The title description can be obtained as a sequence of predicates associated to the stemmed words appearing in the title (a complete parsing of the sentence is not done, since in many cases the title is just a sequence of keywords, or is expressed in a telegraphic style that does not comply to any grammatical rule). The description of the page body involves predicates expressing it at the proper (i.e., not too detailed nor too general) grain-size.

Page structure-related predicates :

`adjacent(o1,o2)` object $o1$ is adjacent to object $o2$
(can be specialized into `above`, `left`, `right`, `below`)

`contains(o1,o2)` object $o1$ contains object $o2$

Style-related predicates :

`emphasized(o)` object o is emphasized
(can be specialized into `italic`, `bold`, `underlined`, ...)

`link(l)` object l is a link

Typology-related predicates :

`frame(f)` object f is a frame

`heading(h)` object h is a heading
(can be specialized into `level1_heading`, ..., `level6_heading`)

`paragraph(p)` object p is a paragraph

`sentence(s)` object s is a sentence

`image(i)` object i is an image

`table(t)` object t is a table
`cell(c,t)` object c is a cell of a table t

`list(l)` object l is a list
(can be specialized in `bulleted_list`, `numbered_list`, `description_list`)
`item(o,l)` object o is an item of a list l

Sentences can be described according to their grammatical structure. Since advanced techniques for understanding images content are not available yet, text comments to the images can be exploited, if any, in a way similar to what was done for the page title. Ontological information on layout objects might help: the various spatial relations are all instances of *adjacent*, the different kinds of typographical emphasis are all instances of *emphasized*, the six levels of heading are all instances of *heading*, the various kinds of lists are all instances of *list*; moreover, tables and frames can be seen as specific cases of *space_partition*, tables and lists can be considered as instances of a concept *information_organization*, and so on.

A distinction is worth between so-called *semi-structured* web pages [9], reporting information as a tabular sequence of items (e.g., results of a search engine, on-line catalogues, etc.), and web pages in general. Indeed, in the former case (typically pages generated automatically as a results of some database query), learning rules for recognizing the items structure is much easier, due to the same style being used for each component in all items of pages from a given site. On the other hand, differently from printed documents belonging to series, generic Web pages show an enormous variability and lack of layout standards: creative and eccentric pages are considered an added value for a site. This means that the learning phase cannot rely on layout considerations only, but must be supported by text-oriented techniques (see [7]), by ontologies that can shift the reasoning level to a higher semantic level, and by the information contained in pages that are related by hyperlinks to the one under processing.

Sample descriptions of academic pages (shown in Figure 2) are:

academic(p) :- title(p,t), universita(t), studi(t), bari(t), contains(p,f1),
contains(p,f2), contains(p,f3), frame(f1), frame(f2), frame(f3),
adjacent(f1,f2), adjacent(f2,f3), adjacent(f1,f3), ..., contains(f3,t),

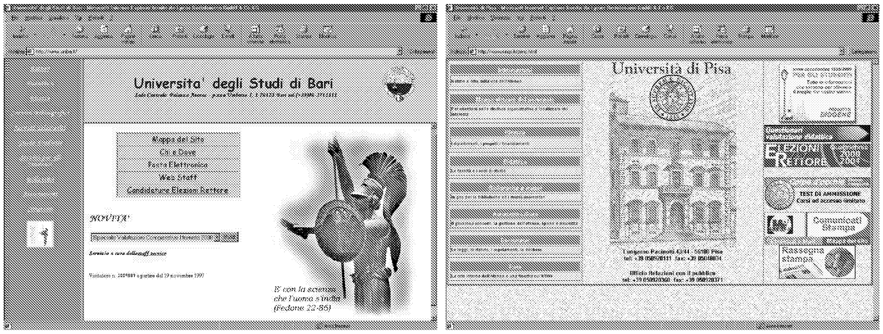
Universita' degli Studi di Bari
Mappa del Sito
Chi e Dove
Posta Elettronica
Web Staff
Candidature Elezioni Rettore
NOVITA'
Servizio a cura dello staff tecnico
Visitatore n. 1009887 a partire dal 19 novembre 1997
E' con la scienza
che l'uomo s'india
(Fedone 22-86)
Università di Pisa
Lungarno Pacinotti 43/44 - 56100 Pisa
tel: +39 050920111 fax: +39 05048034
Ufficio Relazioni con il pubblico
tel: +39 050920360 fax: +39 050920371
ELEZIONI RETTORE
TEST DI AMMISSIONE
Corsi ad accesso limitato
Comunicati Stampa
Rassegna stampa

Fig. 2. Academic homepages

table(t), cell(c1,t), contains(c1,s1), sentence(s1), link(s1), *sentence_description(s1)*, . . . , contains(f3,i), image(i), adjacent(t,i), . . .

academic(p) :- title(p,t), universita(t), pisa(t), contains(p,t), table(t), cell(c1,t), cell(c2,t), . . . , contains(c1,t1), table(t1), cell(c11,t1), . . . , contains(c11,t11), table(t11), cell(c111,t11), contains(c111,s111), sentence(s111), link(s111), *sentence_description(s111)*, cell(c112,t11), contains(c112,s112), sentence(s112), *sentence_description(s112)*, . . . , contains(c2,i), image(i), . . .

Interesting and promising preliminary results have been obtained on the induction of rules for classification of this kind of Web pages, and on the exploitation of their textual content for identifying the topic under consideration [5].

5 Conclusions and Future Work

Document management is critical for the distribution and preservation of knowledge. This paper proposes a general technique to discover significant knowledge in a database of documents in paper, electronic and Web format, to be used as meta-information for their content-based retrieval and management. Such a technique is based on symbolic (first-order) learning for automatically classifying the documents and their layout components according to their semantics, whose performance is encouraging. In a Semantic Web development perspective, this will allow to represent the documents in XML format, enclosing proper tags that express the role of their components. A system implementing these ideas for paper documents is already working, while one for electronic documents is currently under development, at prototype stage. The next step will concern the extension to web pages.

References

1. T.M. Breuel. Two geometric algorithms for layout analysis. In *Workshop on Document Analysis Systems*, 2002.
2. F. Esposito, S. Ferilli, N. Fanizzi, T.M.A. Basile, and N. Di Mauro. Incremental multistrategy learning for document processing. *Applied Artificial Intelligence*, 17(8/9):859–883, 2003.
3. F. Esposito, D. Malerba, and F.A. Lisi. Machine learning for intelligent processing of printed documents. *Journal of Intelligent Information Systems*, 14(2/3):175–198, 2000.
4. S. Ferilli, F. Esposito, T.M.A. Basile, and N. Di Mauro. Automatic induction of rules for classification and interpretation of cultural heritage material. In T. Koch and I.T. Sølvberg, editors, *Research and Advanced Technology for Digital Libraries*, number 2769 in Lecture Notes in Computer Science, pages 152–163. Springer, 2003.
5. S. Ferilli, N. Fanizzi, and G. Semeraro. Learning logic models for automated text categorization. In F. Esposito, editor, *AI*IA 2001: Advances in Artificial Intelligence*, number 2175 in Lecture Notes in Artificial Intelligence, pages 81–86. Springer, 2001.
6. S. Ferilli, N. Di Mauro, T.M.A. Basile, and F. Esposito. Incremental induction of rules for document image understanding. In A. Cappelli and F. Turini, editors, *AI*IA 2003: Advances in Artificial Intelligence*, number 2829 in Lecture Notes in Artificial Intelligence, pages 176–188. Springer, 2003.
7. D. Freitag. Information extraction from HTML: Application of a general machine learning approach. In *AAAI/IAAI*, pages 517–523, 1998.
8. G. Nagy. Twenty years of document image analysis in PAMI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):38–62, 2000.
9. S. Soderland. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34:233–272, 1999.