UvA-DARE (Digital Academic Repository)

Internet access to scanned paper documents

Worring, M.; Smeulders, A.W.M.

Publication date
1998

Published in
Proceedings of the 3rd ACM International Conference on Digital Libraries, June 23-26, 1998, Pittsburgh, PA, USA

[Link to publication](#)

Citation for published version (APA):

Worring, M., & Smeulders, A. W. M. (1998). Internet access to scanned paper documents. In I. Witten, R. Akscyn, & F. M. Shipman (Eds.), *Proceedings of the 3rd ACM International Conference on Digital Libraries, June 23-26, 1998, Pittsburgh, PA, USA* (pp. 313-314). ACM.

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Internet Access to Scanned Paper Documents

Marcel Worring Arnold W.M. Smeulders
Intelligent Sensory Information Systems
University of Amsterdam, Faculty WINS
Kruislaan 403, 1098 SJ Amsterdam
The Netherlands
E-mail: worring@wins.uva.nl
<http://carol.wins.uva.nl/~worryng/>

ABSTRACT

In this contribution we identify the different structures to encounter in a hyperdocument. Methods are described for deriving those structures from scanned paper originals. The content *and* structure of the document is then made available in a form suited for an Internet browser. It provides convenient access to the scanned paper document

KEYWORDS Document access, document understanding, hypertext structure, hypertext understanding

INTRODUCTION

With the advent of the Internet, remote access to archives of digitized paper documents is now feasible. This is of great importance for archived scientific journals, design archives of airplanes, patent applications, and manuals. Access to such documents is mostly done by scanning the document, showing the facsimile for visual inspection. However, such documents exhibit a number of identifiable structures. When the content and full structure of the document would be available, access could be based on those. To derive the structure and content of documents, document image understanding tools (see e.g. [3]) are in common use. Finding hypertext structure is mostly limited to text. However, figures often play a major role in communicating the message of the document. They have been included by the author to highlight and explain the most important parts. We therefore consider links between the text and the contents of figures. For vector based graphics, impressive work in this direction is presented in [1]. We focus on manual like documents (see figure 1) where the document is available as a bitmap only [4].

HYPERDOCUMENT STRUCTURES

For a paper document models for the layout, dealing with how things have been printed and models for logical structure, deal-

Figure1: Example document pages.

ing with meaning, are in common use. A document which in addition has hypertext structure is called a hyperdocument in this paper. In a hyperdocument we distinguish six different structures. They are exemplified in figure 2. *Hierarchical structure* is a tree based grouping of objects. *Linear structure* is reading order and any other type specific linear ordering like subsequent figures. An *index structure* is indeed an index. A *side-loop structure* is any structure where one can jump to a node, and then only go back to the link source. *Cross-group structure* is a set of relations between exactly two nodes. These nodes are each others scope. Finally, any remaining relations form the *cross-reference structure*.

DOCUMENT UNDERSTANDING

In [4] methods are presented for deriving the hypertext structure of a scanned document. The first analysis step is to take the scanned image and find its constituent basic objects like text objects, figures, and lines. From there its layout structure is derived, mainly concerned with column structure. Basic objects with label text are processed by a commercial OCR system directly. The content of figures is analyzed to locate all textual labels. These are also processed by the OCR system. Section identifiers in the text are the basis for finding the logical structure of the text. The text analysis also provides the *scope* of each figure that is the part of the text in which a specific figure is described.

For deriving the six hypertext structures, note that the layout structure is needed to find the logical structure, but after that

Figure2: Different document structures.

has become irrelevant. Hence, *hierarchical and linear structure* are derived completely from the logical structure. An *index structure* can be based on important keywords in the text or on the collection of all figure labels found. Most prominent *side-loops* are footnotes, which are identified by searching for specific markers in the text. To find *cross-group structures* each label in a figure is matched to the text in its derived scope. As the text is the result of an OCR process care has been taken to be robust against recognition errors. *Cross-reference links* have not been considered.

INTERNET ACCESS

Let us now consider the presentation of the document and its structure to the user. First, we consider the design in terms of abstract presentation specifications [2]. For the presentation of the structure derived in the previous section, high level presentation specifications are needed, in fact equivalent to an identification of one of the six structures. Each of these should be compiled into a set of low level presentation specifications which in the reference are defined as logical channels. Furthermore, link context is specified, indicating which part of a displayed component is replaced when the link is followed.

In our case two content channels are defined. One for the text objects and one for figures. In addition we define navigation channels, where the navigation controls are placed. Figures are displayed with their caption, one at a time, in the figure content channel. The *linear structure*, ordering the figures, is made available in the navigation channel. For text, all subsequent pieces of text are placed in the content channel. Access to the *linear structure* is via a scrollbar. The *hierarchical text structure*, is placed in the text navigation channel. When an *index structure* is created it is placed in the navigation chan-

nel, as it provides an alternative way of going to a specific place in the document. For *cross-group structure* between text and figure, link context plays an important role. When the links point from and to the same figure, the link context is restricted to the affected anchors whenever it remains within the group. Activation of a *side-loop* in the text, should reuse the navigation channel as the system must always return to its previous state whenever the user is finished with reading the information. Finally, we have the *cross-reference* links, but these are so broad that we can not make any specific remarks on the presentation specifications.

The design given above is realized in HTML using frames to implement channels. Java applets are used to implement the cross-group structures.

Figure3: Screenshot of the Internet interface.

In conclusion having identified and analyzed the structures in a hyperdocument, allows convenient Internet access to the content *and* the structure of the paper document; a major improvement over facsimili presentation. The document interface (see figure 3) can be accessed via the demo link on our homepage.

REFERENCES

1. L.S. Baum, J.H. Boose, and R.J. Kelley. Graphics recognition for a large-scale airplane information system. In *Proceedings of second IAPR workshop on Graphics Recognition*, pages 62–69, 1997.
2. L. Hardman, D.C.A. Bulterman, and G. v. Rossum. The Amsterdam hypermedia model: Adding time and context to the Dexter model. *Communications of the ACM*, 37(2):50–62, 1994.
3. L. O’Gorman and R. Kasturi. *Document Image Analysis*. IEEE Computer Society Press, 1995.
4. M. Worrington and A.W.M. Smeulders. From linear to non-linear reading: a case study to provide internet access to paper documents. In *Fourth International Conference on Document Analysis and Recognition, Ulm*, pages 273–277. IEEE computer society, 1997.