AUTOMATIC INDEXING OF TEXT AND GRAPHICS IN TECHNICAL MANUALS

M. Worring [1], *B. Wielinga*[2], *A. Anjewierden*[2]

[1]Intelligent Sensory Information Systems
University of Amsterdam
kruislaan 403, 1098 SJ Amsterdam
The Netherlands
worring@science.uva.nl
www.science.uva.nl/˜worring

F. Verster[1] *L. Todoran*[1], *S. Kabel*[2] *R. de Hoog*[2]

[2]Department of Social Science Informatics
University of Amsterdam
Roetersstraat 15, 1018 WB Amsterdam
The Netherlands
wielinga@swi.psy.uva.nl
www.swi.psy.uva.nl

ABSTRACT

Goal driven authoring of training material from existing technical manuals requires automatic indexing of the content of the manual. In this contribution we consider the different representation levels and document knowledge required to do the task. On that basis we have developed tools for automatic indexing in diverse domains.

1. INTRODUCTION

Digital media handling tools are drastically changing the way technical manuals are being employed in teaching people operating instruction and maintenance.

Instead of a teacher using paper material as the basis for a course, the use of digital media is now becoming commonplace. Internet allows for one-line courses [2] and the availability of cheap laptops and handheld devices opens the way to on-the-job-training.

Which parts of the manual are used for training, depends on the user goal, which is situation dependent. It is undoable to write from scratch new material for all those different ways of using the technical information. Needed is a decomposition of existing manuals into basic fragments with proper index terms. These fragments can then be reused for goal driven authoring of training material.

Technical manuals contain *multiple* media in the form of text and graphics. The use of multiple media is especially important for teaching novice users [5]. The text provides the basic information. The illustrations are important in understanding the material, for both novice and expert users, as images convey visual information which is hard to describe in words. So, both the text and the graphical pictures should be decomposed and properly indexed.

This project is sponsored by the European commitee through the ESPRIT project Integrated Manuals and Training (IMAT)

In general, indexing of graphics is impossible due to the "semantic gap": the difference between the interpretation a user has when seeing the document, and the index terms that can be derived automatically e.g. the ones reviewed in [3]. Technical manuals are, however, highly structured, adhere to style rules, and have a limited vocabulary. This makes the task doable.

Boeing [1] was one of the first to build a system for automatically indexing technical manuals consisting of paper or vector graphics based illustrations. They treat graphics and text in different ways. In this paper we present our approach to the problem which aims at treating the textbody and the graphical drawings on equal footing, exploiting the structure and the limited vocabulary.

In the following, we first derive the different levels of representation that we have in a document and explicify the knowledge about the document and the domain. Then we present the automatic analysis of these documents. Finally, we give some implementation details.

2. DOCUMENT PRIMITIVES AND THEIR STRUCTURE

Proper definitions on what is in the document, have to be made before indexing can take place. We first define, for the purpose of indexing, the atomic components of the document. For text and graphics alike these are called *logical primitives* here.

- logical primitives: the smallest components in the document that can be given an interpretation.

Examples are electrical components and wires in the drawing, and paragraphs, captions, and emphasized text primitives in the text body.

As the basis for document analysis is the original document without its interpretation, we also define the atomic

0-7695-1198-8/01/$10.00 (C) 2001 IEEE

components with respect to layout. These are the

- layout primitives: the smallest components in the document with consistent visual representation.

For graphical drawings the layout primitives correspond to the basic shapes like lines, strings, rectangles, icons, circles, etc. Primitives for text are pieces of text with consistent fonttype and indentation style. Layout primitives can be parameterized. For graphical primitives, attributes are for example scale or color. In most cases, each set of parameter values found in a given document are associated with a different logical primitive. For example, a red circle is not the same as a green circle as they usually have a different interpretation. For text example parameters are font size and amount of indentation.

To be consistent with documents in digital format, layout primitives are defined atomic for the analysis task. However, for scanned documents the layout primitives can in turn be decomposed into *low-level primitives*. These are the (binary) image objects that make up the layout primitives.

For text primitives (either in a drawing or as part of the textbody) the text itself forms the *content* of the primitive.

Having properly defined the notion of logical and layout primitives, we define three levels of representation for the structure of the document:

- layout structure: the relations between the layout primitives based on their spatial relations in the document.
- logical structure: the relations between the logical primitives based on the roles they have in the document.
- semantic structure: the relations between the interpretations of logical primitives.

The layout structure is induced by the spatial relations between the layout primitives. In the logical structure, among other things, the reading order in the document is captured. Semantic structure can only be considered by employing domain knowledge which is the topic of the next section.

3. DOCUMENT KNOWLEDGE

The limited vocabulary used in technical manuals is captured in two ontologies [4]. These provide a closed conceptual vocabulary based on a sub/super class hierarchy. The concepts are developed semi-automatically by considering the authoring process, words in the text, and the parts-list associated with a technical manual. As visual concepts often cannot be described in words, the legenda is used. In the legenda, pairs of a part and its visual representations are present. These are combined into a concept in the ontology.

The first ontology is the

- domain ontology : concepts to describe what the manual is about.

These are concepts for describing the technical equipment, its functioning, and its use. For each new domain this ontology has to be adapted to concisely desribe the domain. It is not dependent on the carrier i.e. a concept could be described in words or visualized in a picture.

The second ontology is the

- fragment ontology : domain independent concepts for describing individual layout/logical primitives.

The concepts here are mainly for describing the carrier of the fragment and its general classification based on surface characteristics. Furthermore, It determines the set of allowed logical and layout primitives for the manual under consideration. This includes the set of symbols used which are represented as small binary pictures.

In structured technical manuals there is a strong relation between the terms in the domain ontology and the fragments. Drawing and style conventions assure that the manual is consistent and easy to read. For example, names of electrical components have standard style italics in the text and in the illustrations such components are always drawn as solid boxes with text inside. The associated information is important for the indexing process and is captured in two sets of style rules:

- semantic style rule: the mapping from a concept in the domain ontology to one or more concepts in the fragment ontology.
- layout style rule: the mapping from logical primitives to layout primitives.

These style rules form the basis for document analysis. An overview of the decomposition of the document into the different representation levels and the role of knowledge therein is shown in figure 1.

4. DOCUMENT ANALYSIS

4.1. Layout analysis

We now consider the process of decomposing and indexing an existing manual based on the document knowledge derived above.

When the manual is available in digital form created with a word processor and a drawing program, the layout primitives are relatively easy to derive. Each command in the vector file corresponds to a primitive and each style change indicates a new primitive in the text. In most cases, however, we only have the low level representation and we have to devise algorithms to combine or decompose the low level information into layout primitives.

0-7695-1198-8/01/$10.00 (C) 2001 IEEE

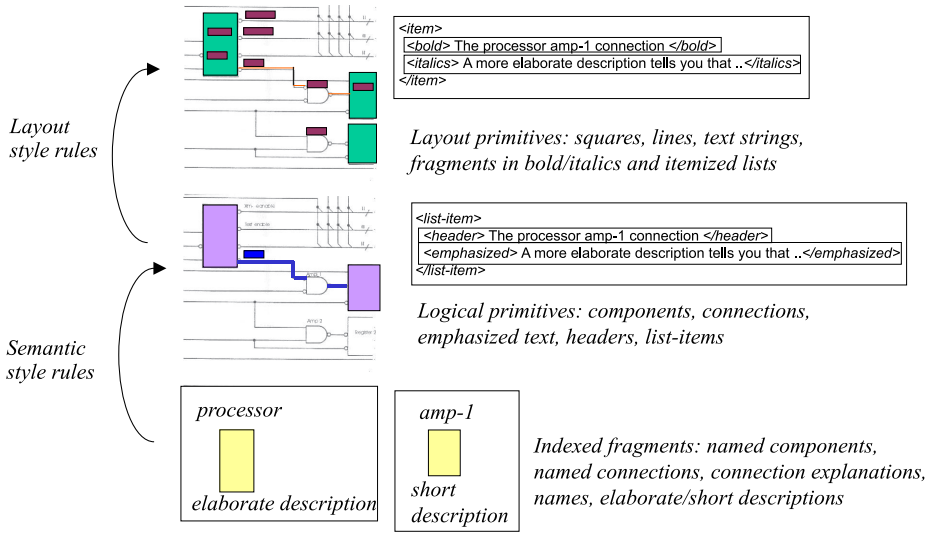

Figure 1: *Overview of the different representation levels for a piece of an example document.*

For graphical drawings we have algorithms for detecting horizontal/vertical lines, rectangles, and strings as they account for the major part of the drawings. Furthermore, we have developed an algorithm which matches a given symbol at every position in the drawing. Significant local maxima of the matchfunction correspond to instantiations of the symbol at that position [6].

OCR (optical character recognition) is applied to convert the strings in the drawings to ascii text. The same tools are applied to the textpages when documents are not available in digital format.

Layout structure of the document follows directly from the extracted fragments by considering their spatial relations.

After layout analysis all documents are brought into standardized format by using XML. For graphical drawings the XML-based Scalable Vector Graphics (SVG) format is used. Note, that after the layout analysis the only essential difference between scanned or digital documents is that the system cannot detect fragments with absolute certainty.

Finally, at this point there is furthermore no real difference between text and graphics, both are represented as XML tagged fragments in the document, and both have associated layout structure.

4.2. Logical analysis

For logical analyis, the layout style rules have to be inverted. In practice, however, as style rules can be many-to-one mappings, the style rules cannot be inverted in a unique way. The *inverted layout style rules* are captured in the form of rules which can be applied whenever the preconditions on the attributes hold. For every layout element this leads to a set of possible roles in the document. In a top-down grammar driven fashion the genuine role of each element is determined. When the inverted style rules are applied, style based XML tags are replaced by tags defining the role of each primitive in the document.

The result of the above analysis is the set of logical primitives and their structure. Currently it identifies the chapter and section headers, tables, items list, figure captions, etc. Reading order is extracted by asking the user to provide some basic information on the use of columns in the document. Components in the drawings are identified as boxes with centered text, current flows are detected as lines with arrows, and named connections are identified as lines with associated text labels.

4.3. Semantic analysis

Similar to the use of layout rules, a set of *inverted semantic style rules* can be derived to annotate fragments with concepts from the domain ontology. In this way text paragraphs are for example automatically indexed as general, short or elaborate descriptions of the associated component. Graphical drawings are classified as overview or detailed and pictures that are the exploded view of others are connected to their origin. When a visual representation derived from the legenda is present, the semantic style rules is a one-to-one mapping and hence trivial to invert.

Having established the relation between fragments and concepts in the ontology we still have to fill in the attributes e.g. the name in the concept to make the fragment an instantiation of that concept. This is always done on the basis of text. The text can be the text of a paragraph in the textbody, or a textstring in the drawings. In both cases the text

0-7695-1198-8/01/$10.00 (C) 2001 IEEE

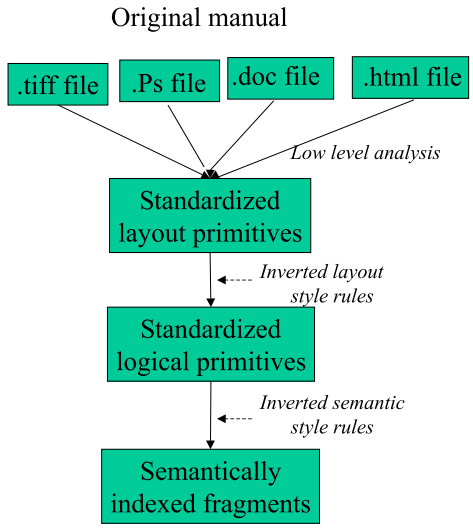

Figure 2: *The different steps that are performed when analyzing and indexing a document.*

is analyzed to find keywords and names that can be related to the attributes in the domain ontology.

At this point in the analysis, document structure does not have a purpose anymore as the *semantic structure* is captured by the relations in the domain ontology. So, all fragments are stored in the database with a link to the concept they are derived from.

An overview of the full analysis process is given in figure 2.

5. IMPLEMENTATION

We have implemented a set of tools for performing the analysis as stated above. There are tools for converting documents in various forms into the standard format used in the system, image analysis tools implemented in C/C++ and Prolog based tools for handling the ontology and style rules. OCR is performed using the commercial TextBridge API from ScanSoft. The system we are developing is currently being applied to air force data, data from a railway company, and manuals from car repair.

As the whole system is XML and SVG based the visualization of (intermediate) results can be done using standard tools. Because OCR results for complex text strings in drawings are often inaccurate they are used for indexing purposes only. When the text is required the corresponding image part is shown. In the future, we aim to extend and use the ontology to generate a lexicon for each domain to drastically improve the OCR performance.

Fragments are stored in the object oriented Jasmine database. For authoring any standard windows based package can be used to retrieve the stored fragments.

6. CONCLUSIONS

In this paper we have considered the automatic document indexing problem by employing a multi-level document representation and explicit document knowledge. This makes the otherwise underdetermined problem tractable. It also allows to treat text and graphics on equal footing as it is clear at what point they can be considered equal and where they are different. Furthermore, it makes clear how to integrate documents in different formats. They have to be brought into the appropriate representation level depending on the richness of the original document. In practice this requires some analysis steps or a simple format conversion.

The analysis process leads to a full decomposition of the document into individually indexed fragments stored in a database. These stored fragments can be retrieved by the human author through the index created, when constructing new teaching material. Ultimately, when the system could analyze a training goal, the system could generate the presentation on-the-fly.

A weakness of the current system is that it assumes that the result of each step is sufficiently accurate to be used in the next phase of indexing. Further consistency checking and error recovery will be investigated in future research.

The approach presented is not restricted to text/graphics documents. An interesting challenge is to follow the same approach for automatic indexing of video training material.

7. 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] F. Bota, L. Farinetti, and A. Rarau. An educational-oriented framework for building on-line courses using XML. In *IEEE International Conference on Multimedia and Expo*, 2000.

[3] D. Doermann. The indexing and retrieval of document images: a survey. *CVGIP: Image Understanding*, 70(3):278–298, 1998.

[4] S.C. Kabel, B.J. Wielinga, and R. de Hoog. Ontologies for indexing technical manuals for instruction. In *Proceedings of the AI-ED*, 1999.

[5] F. Vetere and S. Howard. Prior knowledge and redundant multimedia. In *IEEE International Conference on Multimedia and Expo*, 2000.

[6] M. Worring and A.W.M. Smeulders. Content based internet access to scanned documents. *International Journal on Document Analysis and Recognition*, 1(4), 1999.