Document Reverse Engineering: From Paper to XML

Kyong-Ho Lee¹, Yoon-Chul Choy², Sung-Bae Cho², Xiao Tang¹, and
Victor McCrary¹

¹ National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg,
MD 20889, USA
{kyongho, xiao.tang, victor.mccrary@nist.gov

² Dept. Computer Science, Yonsei Univ., 134 Shinchon-dong, Seodaemun-ku, Seoul, 120-
749, Korea
{ycchoy, sbcho}@cs.yonsei.ac.kr

1 Introduction

Since XML has the advantage of embedding logical structure information into documents, it is widely used as the universal format for structured documents on the Web. This makes it attractive to convert paper-based documents with logical hierarchy into XML representations automatically. Document image analysis and understanding [1] consists of two phases: geometric and logical structure analysis. Because the two phases take different kinds of data as input, it may not be desirable to apply the same method to them. Targeting technical journal document with multiple pages, we present a hybridization of knowledge-based and syntactic methods for geometric and logical structure analysis of document images.

Sophisticated geometric structure analysis is a prerequisite for logical structure analysis that recognizes logical components and their hierarchy. The geometric properties of document images are different according to the type of the document. Even the documents of the same type can differ from each other. The formalism to reflect geometric characteristics that are publication-specific as well as class-specific is very important. For geometric structure analysis, we present a knowledge-based method that can handle more sophisticated problems than previous works.

Logical components such as headers or paragraphs are identified from document images based on their geometric characteristics of the corresponding text regions. The logical components that are directly identifiable from geometric characteristics of text regions are called primary structures and the ones such as *Section* that can be identified through grouping components together are called secondary structures [2]. For XML documents with logical hierarchy, secondary structures as well as primary structures should be extracted from multi-page documents. We present a syntactic method for logical structure analysis. Normally, a text region composing a document image functions as a header or a body, which corresponds to a title or a paragraph, respectively. Headers and bodies may be classified into various kinds according to the geometric characteristics of the corresponding text region. We define headers and bodies as functional components, and a tree with nodes of headers and bodies as a functional structure tree.

Fig. 1. A flow diagram for the proposed document analysis process

To improve in the processing speed of logical structure analysis, compared with previous related works of which the basic units are text lines, the proposed method takes a functional structure tree.

2 Document Structure Analysis

As shown in Fig. 1(a), the geometric structure method is composed of two stages: region segmentation and identification. The knowledge base has been constructed using the characteristics of a technical journal publication. First, it categorizes the types of document images into title and body pages, and examines the characteristics of each type. Regions of the document image are divided into complex layout objects such as text lines, equations, images, drawings, and tables. The knowledge base is expressed with 91 rules, based on a careful examination of geometric characteristics of layout objects. The result of segmentation process does not usually have a one-to-one matching with composite layout components. A figure object contains many small regions that correspond to image or drawing regions. Based on the hybrid of top-down and bottom-up techniques, the proposed method identifies non-text objects such as images, drawings and tables, as well as text objects such as text lines and equations by splitting and grouping segmented regions into composite components.

(a) Document model

(b) Algorithm for a functional structure tree

Fig. 2. The proposed document model and algorithm for creating a functional structure tree

The proposed method for logical structure analysis takes as input a set of text lines from geometric analysis of each multi-page document as shown in Fig 1 (b). A document model is proposed as shown in Fig. 2 (a). For secondary structures, the model expresses such information as their names and frequency, and ordering information among them in the form of regular expressions. The model defines the required and optional contents of secondary structures, specified by sub-element names. Additionally, it defines order among the contents and their occurrence. As geometric conditions that primary structures should satisfy, the model describes geometric characteristics such as column type, the number and height of text lines, the space before and after a text region, the density distribution of a black pixel, and alignment.

As shown in Fig. 1(b), the logical analysis method is composed of three phases. First is the creation step of a functional structure tree where text lines are merged into a sequence of headers and bodies, and by splitting the sequence repeatedly a hierarchical tree is created. A sequence of header and body is created by merging adjacent text lines that have similar geometric characteristics based on the three principles, which are proximity, similarity and contiguity, from Gestalt psychology [3]. There are different types of header and body that are differentiated from each other by the geometric characteristics.

The method next creates a functional structure tree from the sequence of different header and body based on repetitive characteristics of their labels. Normally, a document contains quite a few sections and each section is composed of lower level sections with hierarchically nested structures. Specifically, each of the section structure is identified by the section title and the title of the sub-section is placed after the title of the section that it is contained in. Therefore, the proposed method creates a functional

structure tree in top-down manner by splitting a sequence repeatedly with the repetitive header as the base as shown in Fig. 2(b). If the sequence contains a single header that corresponds to the title of a lower level section, it is split using this as the base.

Secondly, the parsing method applies a document model on a functional structure tree and creates a logical structure tree. By doing depth-first search on a functional structure tree, the method tests whether the hierarchical structure of interior nodes and the geometric characteristics of leaf nodes are acceptable under the grammars of the document model. The method responds appropriately to interior nodes including a root node according to whether the child of the node in question is a leaf node or an interior node. If the child node is a leaf, it is labeled with an element name that satisfies its geometric characteristics. In case of an interior node, one of permissible element names is given and whether the corresponding content model of the element is appropriate or not is tested recursively. If the child node is invalid, the proposed method backtracks to an alternative label for the parent node.

As a result, the proposed parsing method completes a logical structure tree by assigning a label to each node of the functional structure tree. Finally, the method generates an XML document as the result of logical structure analysis by doing depth-first search on a logical structure tree. Particularly, whenever a leaf node is encountered, its corresponding OCR result of the text region is added.

3 Performance Analysis

To evaluate the performance of the proposed methods, we have experimented with 26 journal papers. In terms of geometric structure analysis, experimental results show an accuracy of 99.3% in average. Particularly, to succeed in logical structure analysis, the exact identification of functional components from text lines is very important. In the case of the identification rate of header and body, the proposed method performed in an accuracy of 98.9% in average. Because the parsing method takes a functional structure tree, it is faster than the conventional methods, which the basic units for parsing are either pixels or text lines. Particularly, the proposed method generates XML documents as the result of structural analysis, so that it enhances the reusability of documents and independence of platform.

References

1. Nagy, G.: Twenty Years of Document Image Analysis in PAMI. *IEEE Trans. Pattern Analysis and Machine Intelligence*. 1 (2000) 38–62
2. Summers, K.M.: Toward a Taxonomy of Logical Document Structures. *Proc. Dartmouth Institute for Advanced Graduate Studies (DAGS'95)*, Boston (1995) 124–133
3. Koffka, K.: *Principles of Gestalt Psychology*. Harcourt, Brace and World, New York (1935)