Header and Footer Extraction by Page-Association

Xiaofan Lin
Information Infrastructure Laboratory
HP Laboratories Palo Alto

HPL-2002-129
May 6th , 2002*

E-mail: xiaofan_lin@hp.com

document structure analysis, optical character recognition, header/footer extraction, digit content re-mastering

This paper introduces a robust algorithm to extract headers and footers from a variety of electronic documents such as image files, Adobe PDF files and files generated by Optical Character Recognition (OCR). Compared with the conventional methods based on page-level layout and format, the proposed novel strategy considers a page in the context of neighboring pages. Through such page-association, the headers and footers on a variety of documents can be automatically detected without human interference. In addition, the application of fuzzy string match also make the method resistant against OCR errors.

Header and Footer Extraction by Page-Association

Xiaofan Lin
Hewlett-Packard Laboratories, 1501 Page Mill Road, MS 1126, Palo Alto, CA 94304
E-mail: xiaofan_lin@hp.com

Abstract

This paper introduces a robust algorithm to extract headers and footers from a variety of electronic documents such as image files, Adobe PDF files, and files generated OCR. Compared with the conventional methods based on the page-level layout and format, the proposed strategy considers a page in the context of neighboring pages. Through such page-association, the headers and footers on a variety of documents can be automatically detected without human interference or individual templates . In addition, the application of fuzzy string match make the method resistant against OCR errors.

Keywords: document structure analysis, Optical Character Recognition, header/footer extraction, digit content re-mastering

1. Introduction

Headers and footers are common formatting elements in all kinds of documents. Besides reiterating key archival information such as author names, publication titles, page numbers, and release dates, they also serve decoration purpose by making the page layout more balanced and more visually appealing. With all the publishing and word-processing software, it is very straightforward to add headers and footers to a document. However, the reverse engineering procedure of header/footer generation --- the extraction of headers and footers --- poses a great challenge, which is the subject of this paper.

Header/footer extraction can benefit a number of downstream applications in digital content understanding and re-mastering:

1) **Natural language processing (NLP)**

Because headers and footers do not belong to the body text, they can fragment the normal text flow if not extracted. Let us examine the document shown in Figure 1. Without separating the header from the body text, the sentence across the two pages will read as "we simply asked Distinction between Mental, Physical Phenomena 49 whether a second person could see a first person's mental entity (a thought about a dog)". Obviously, this makes the entire sentence very difficult, if ever possible, to understand. Consequently, the performance of computerized NLP systems can be significantly affected at all levels (see Table 1).

Table 1: Impact of headers and footers on different NLP applications

NLP Applications	Impacts of Headers and Footers
Part-of-speech (POS) tagging	The intruding header/footers change the context of the neighboring words.
Grammatical parsing	The sentences becomes grammatically incorrect.
Keyword extraction and information retrieval	Repeats of headers and footers can distort the statistics of the text.
Text summarization	This application usually depends on POS tagging, parsing and keyword extraction.

2) **Document re-purposing**

A common purpose of content understanding is to re-use the contents. In this type of re-purposing applications, the detection of headers and footers is of great value. For example, when rendering the multi-page document as a complete HTML page, it is desirable to have only continuous text flow without any page breaks. In Print-on-demand (POD) applications, it is often required to customize the headers and footers.

On the other hand, manual extraction of headers and footers can be time and labor consuming because they can appear on every page of a document. That is the motivation behind our exploration of automatic header/footer extraction methods. In some electronic documents such as Microsoft Word files, headers and footers are stored in dedicated sections or marked with special tags, and it is a trivial job to directly locate the headers and footers. However, the original documents usually do not explicitly reveal where the headers and footers are:

1) **Electronic documents derived from scanned paper documents**

A large amount of electronic documents are created by scanning paper documents into computers. They are kept as raw raster images or are further processed by OCR software. In the first case, there exist no clues about the headers and footers at all. In the second case, the OCR software can recognize and convert the text to ASCII/UNICODE format and can also output certain formatting information such as font size, font style, and paragraph alignment. A few OCR software packages even try to detect the headers and footers --- but with limitations described later.

2) **Electronically originated documents**

Header/footer information can be unavailable even for documents originated electronically. For example, due to the ubiquity of Adobe PDF files, it is now a common practice to convert electronic documents from alternative formats such as Microsoft Word and HTML to PDF format before the final distribution by using the so-called "Virtual Printer Driver". Although the resulting PDF versions keep the same look-and-feel as the original documents, much internal information, including the header and footer tags, usually is lost during the conversion.

Although it seems easy for humans to locate the headers and footers, it is technically challenging to build "intelligent" computer programs with similar capabilities:

1) **Headers and footers exist in all kinds of formats.**

Some documents have both footers and headers, some only have headers or footers, and some have neither of them. Besides, the headers/footers can contain the same text such as journal or book tiles on all the pages, or various text such as page numbers and current article titles on different pages.

2) **OCR text errors can make things worse.**

For electronic documents scanned from paper versions, it is quite routine to apply OCR software for the retrieval of the text information. The recognition errors introduced by OCR add to the complexities.

Although much research has been carried out in the general area of document logical structure analysis (DLSA) [1]-[7], there is no published work dedicated to header/footer extraction. A few DLSA systems have limited capabilities extracting headers/footers. Most existing approaches utilize page-level heuristics about the layout and formatting: For example, there should be a large gap between the header and the body text, and the font size of headers/footers should be smaller than that of body text. This kind of heuristics can be based on rules [1]-[3] or statistics [5]. But as mentioned earlier, headers and footers come in different forms and it is almost impossible to find a set of common parameters or rules applicable to most documents (see Figure 2). These type of methods only work well if we can tune the parameters and rules for each type of documents and store them in templates which can be then applied at runtime. Unfortunately, this strategy is only meaningful in limited applications such as processing many back issues of the same journal title.

In the following sections, we present a page-association based method to automatically and robustly detect headers and footers in a variety of electronic documents. Section 2 describes the algorithm and Section 3 shows the experimental results . In Section 4, we summarize the method and discuss directions for future research.

2. Page-Association Based Header/footer Extraction

Although it is difficult to find stable page-level features that can be used to extract headers and footers, there does exist a relatively stable characteristic if we look beyond individual pages. Usually a document contains multiple pages, whose headers and footers are related to each other. The page-association based header/footer extraction is such an observation: Headers/footers are text lines on the top/bottom of the pages with the same/similar counterparts in the neighboring pages. So instead of concentrating on individual pages, we inspect one page's relationship with its neighbors. In fact, this idea is in accordance with the way headers and footers are generated: The publishing or word-processing software allows the user to define rules to generate the headers and footers of continuous pages.

2.1 Workflow of the method

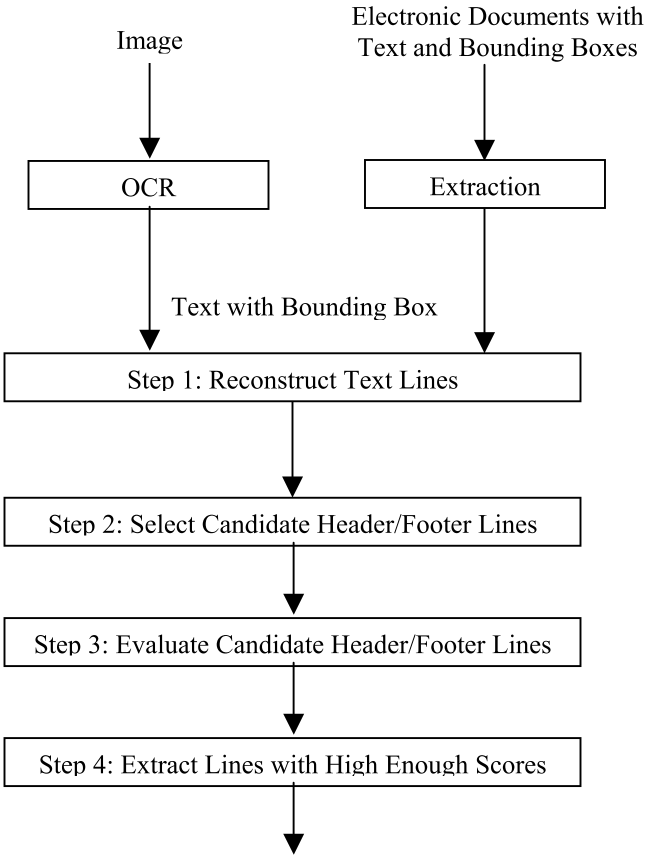


Image
Electronic Documents with Text and Bounding Boxes
OCR
Extraction
Text with Bounding Box
Step 1: Reconstruct Text Lines
Step 2: Select Candidate Header/Footer Lines
Step 3: Evaluate Candidate Header/Footer Lines
Step 4: Extract Lines with High Enough Scores

Figure 3: Workflow of the page-association based header/footer extraction

The process operates on the text information of all the pages. It requires both the text and the bounding box information (the coordinates of the four corners) as the input. If the document already contains such information (for example, text PDF files), we can directly extract text and bounding box from the document. If the electronic document is in raster image format, any OCR software can be employed to recognize the image and generate the required information. Figure 3 shows the whole process:

Step 1: Reconstruct text lines

Even if the input is already in text lines, we still rebuild the text lines on top of the bounding box information because the existing text line information may not be appropriate for header/footer extraction. For example, when there is a big gap between the page number and other part of the header/footer, they can be considered as two logical lines by OCR software. In the header/footer extraction, they should be treated as a single line. The text line construction works as follows:

1. Empty the line buffer;
2. Examine each word in the input document. If its height overlaps more than 50% with any existing line, the word will be added to that line. If it does not belong to any line, a new line will be created.
3. Sort all the lines on the vertical coordinates (from top to down).
4. Sort the words in each line on the horizontal coordinates (from left to right).

Step 2: Select candidate header and footer text lines

The purpose of this step is to reduce the search space and thus increase the speed. Currently the top five lines are selected as the header candidates and bottom three lines are chosen as the footer candidates.

Step 3: Evaluate candidate text lines

This is the central part of the proposed method. Each candidate line will be quantitatively evaluated as to how well it qualifies as a header or footer. We will describe this step further in the next subsection.

Step 4: Make decision
With all the evaluation done in Step 3, the candidate lines with the enough confidence scores are selected as headers or footers.

2.2 Page-association based header/footer evaluation

As mentioned earlier, the most stable feature of headers and footers is that they will repeat in neighboring pages. The problem is how to quantitatively measure such repeats. A couple of issues have to be addressed. First, OCR errors can make the headers/footers not exactly the same on different pages. Second, headers/footers can appear in numerous patterns. For example, odd pages have the journal's title as the headers and even pages have the titles of individual articles as the headers. It is also possible that there is no header on the first page of each article. Besides, page/chapter numbers are usually part of headers/footers and are different from page to page.

The following algorithm is designed to solve the above problems: The i th candidate line on Page j (Line[j][i]) is compared with the i th candidate line on Page k (max (j-WIN,1) <=k <= min (j+WIN, PageNum). Parameter WIN is to control the number of neighboring pages. and it is set to be 8 in our experiment. PageNum is the total number of pages in the document. The max and min operations are to restrain the pages within the legitimate range. For each comparison between Line[j][i] and Line[k][i], a similarity score is calculated as Similarity(Line[j][i],Line[k][i]). The weighted sum of all the scores indicates the page's likelihood to be a header/footer:

$$Score(Line[j][i]) = \sum_{k=\max(j-WIN,1)}^{k\le\min(j+WIN,PageNum)} weights[i] * Similarity(Line[j][i], Line[k][i]) \qquad (1)$$

Weights[i] reflects the fact that different lines have different *a priori* probabilities to be headers/footers. For example, the first line is more likely to be a header than the second line. Here the weights are chosen to be 1.0, 0.75, 0.5, 0.5 and 0.5 for the five header candidates and 0.5, 0.5, 0.5, 0.75 and 1.0 for the five footer candidates. By considering the neighboring pages within the window, all kinds of header/footer "patterns" are implicitly included. The net effect is that the score is high only if similar lines exist within the page window, no matter what pattern the headers/footers are following.

Text match is the basis in computing the similarity between two lines. The text strings contained in the two lines are matched using dynamic programming (DP) to minimize the total edit cost. Then a score is assigned as follows:

$$BaseSimilarity(Line[j][i], Line[k][j]) = \frac{\text{Number of matched characrters}}{\text{Larger of the numbers of characters in the two lines}} \qquad (2)$$

The fuzzy string match tries to reduce the impact of OCR errors. Because page/chapter numbers are very common in headers and footers and can be different from page to page, all the digits are replaced with a special character '@' before the match. For example, the string "48 Chapter 2" becomes "@@ Chapter @" and "50 Diapter 2" becomes "@@ Diapter @" ("Diapter" is due to OCR errors). In this case, the DP match gets a score of 0.69.

In addition, a geometry based similarity score GeometrySimilarity(Line[j][i], Line[k][i]) is also obtained. It is based on comparison of the bounding boxes of the two lines. This measurement is to intended to eliminate accidental good text matches between a header/footer and a normal body text line.

The final similarity is defined as the product of the two components:

$$\begin{aligned} &Similarity(Line[j][i], Line[k][j]) = \\ &BaseSimilarity(Line[j][i], Line[k][j]) * GeometrySimilarity(Line[j][i], Line[k][j]) \end{aligned} \qquad (3)$$

3.Experimental results

The proposed method has been tested on 9 documents with different styles, including 7 periodicals and 2 books. All the 1156 pages are scanned and recognized using a commercial OCR engine. The header/footer extraction system then runs on the generated text and bounding box information. Table 2 shows the results. "false negative" means that system misses a header/footer lines and "false positive" refers to errors in which it incorrectly adds an extra header/footer line. The precision rate and recall rate are defined as follows:

$$\text{Precision Rate} = \frac{\text{Number of Correctly Extracted Header/footer Lines}}{\text{Total Number of Extracted Header/footer Lines}} \times 100\% \quad (4)$$

$$\text{Recall Rate} = \frac{\text{Number of Correctly Extracted Header/footer Lines}}{\text{Total Number of True Header/footer Lines}} \times 100\% \quad (5)$$

The precision rate is 98.00% and the recall rate is 92.7%. This result is satisfactory considering the fact that the extraction is automatically done on a variety of documents with no document-dependent templates or parameters. Because the context of a header/footer plays a critical role in the decision, the method is to prone to errors if there are too many unique headers and footers. For example, the third and eighth documents are composed of mostly short articles and the headers reflect the current article names. Consequently, page-association cannot give enough confidence scores to many headers because they only appear in one or two pages.

Table 2: Statistics on 9 documents

No	Type	Number of Pages	Headers			Footers		
			Header Lines	False Negative	False Positive	Footer Lines	False Negative	False Positive
1	Journal	119	94	4	0	102	0	0
2		124	108	6	0	10	6	0
3		200	179	30	4	0	0	0
4		176	154	1	1	4	0	0
5		90	0	0	19	79	2	0
6		112	80	1	0	87	4	0
7		76	60	2	0	65	0	0
8	Book	113	100	25	0	13	2	0
9		146	126	9	0	0	0	0
Total		1156	901	78	24	360	14	0
Precision Rate =98.00%					Recall Rate=92.70%			

Figure 4 also displays the detailed results on a document of 76 pages. Each curve represents a specific candidate header/footer lines across all the pages. The Y axis corresponds to scores as calculated by (3). It can be seen that the first lines ("Top 1") and the last lines ("Bottom 5") enjoy high scores on most pages. This observation indicates that the first lines are headers and the last lines are footers on those pages. The deep gaps on the two curves reflect the "abnormal" pages (for example, the title pages) without headers/footers.

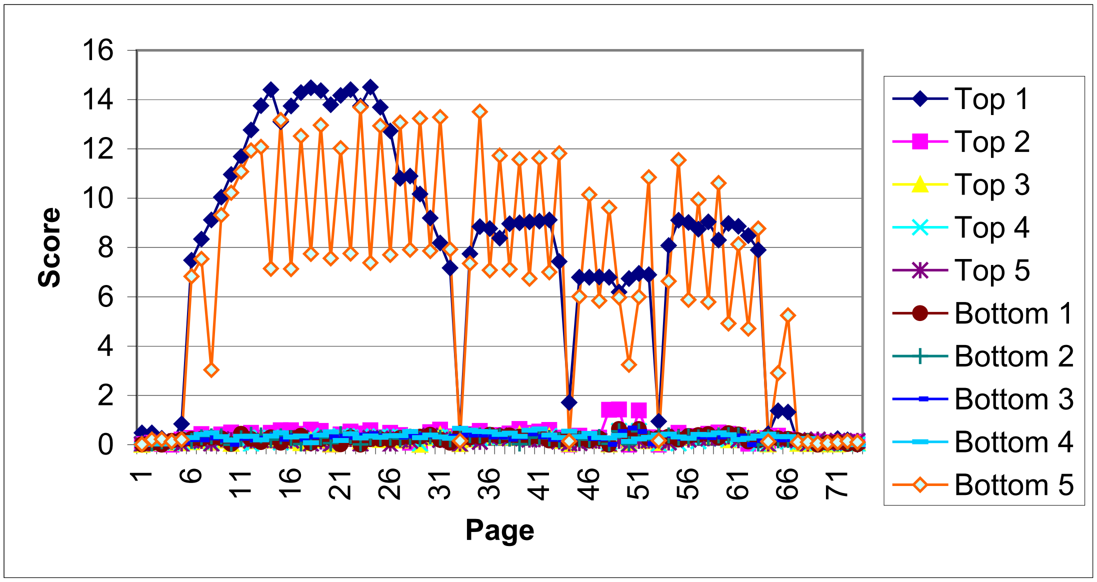


Figure 4: Confidence scores for one document

4.Conclusions

In this paper we analyze the significances and challenges of header/footer extraction. We introduce a page-association based approach to the problem. The experimental results show that the method provides a robust solution directly applicable to a wide range of documents. Furthermore, the method is based on statistics and hand-crafted rules are avoided. The decision is also in terms of confidence scores and gives the user the opportunity to review suspect headers and footers.

Traditionally, most research in document understanding concentrates on the page-scope analysis. This work goes beyond individual pages and tries to solve this seemingly page-scope problem on the document level. We believe that document-level understanding represents an important trend as the page-level processing becomes mature and leaves less room for further improvement.

5. References

[1] George Nagy, Sharad Seth, Mahesh Viswanathan, "A Prototype Document Image Analysis System for Technical Journals", Computer, vol 25 no 7, pp 10-22, July 1992.

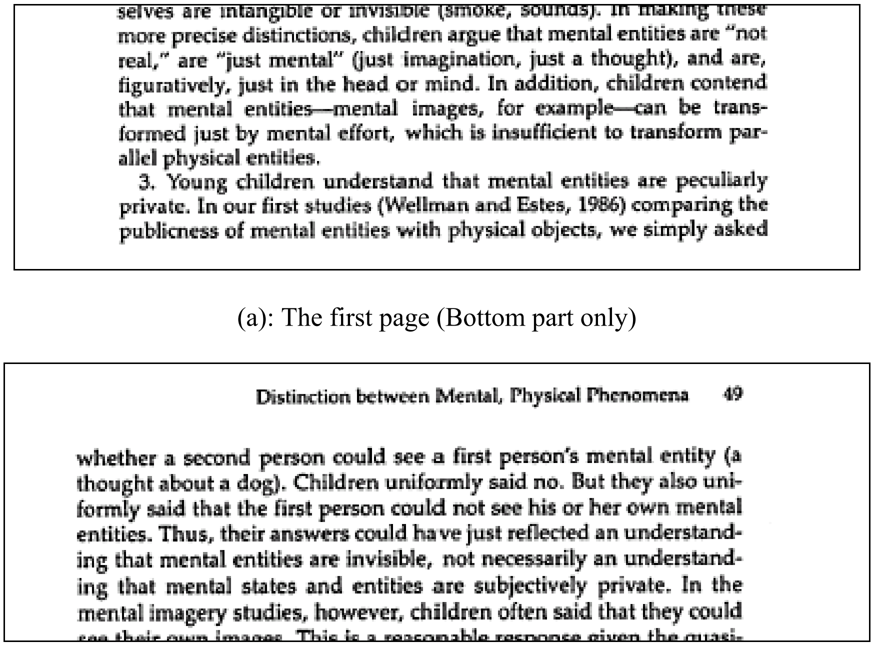
[2] Jürgen Schürmann, Norbert Bartneck, Thomas Bayer, et al, "Document Analysis-From Pixels to Contents", Proceedings of IEEE, pp 1101-1119, July 1992.

[3] ChunChen Lin, Yosihiro Niwa, Seinosuke Narita, "Logical Structure Analysis of Book Document Images Using Contents Information", Proceedings of 4th International Conference on Document Analysis and Recognition, Ulm, Germany, pp 1048-1054, Aug 1997.

[4] Anjo Anjewierden, "AIDAS: Incremental Logical Structure Discovery in PDF Documents", Proceedings of 6th International Conference on Document Analysis and Recognition, Seattle, USA, pp 374-377, Aug 2001.

[5] Souad Souafi-Bensafi, Marc Parizeau, Franck Lebourgeois, Hubert Emptoz, "Logical Labeling using Bayesien Networks", Proceedings of 6th International Conference on Document Analysis and Recognition, Seattle, USA, pp 832-836, Aug 2001.

[6] A. Belaïd, "Recognition of Table of Contents for Electronic Library Consulting", International Journal on Document Analysis and Recognition, in press.



selves are intangible or invisible (smoke, sounds). In making these more precise distinctions, children argue that mental entities are "not real," are "just mental" (just imagination, just a thought), and are, figuratively, just in the head or mind. In addition, children contend that mental entities—mental images, for example—can be transformed just by mental effort, which is insufficient to transform parallel physical entities.

3. Young children understand that mental entities are peculiarly private. In our first studies (Wellman and Estes, 1986) comparing the publicness of mental entities with physical objects, we simply asked

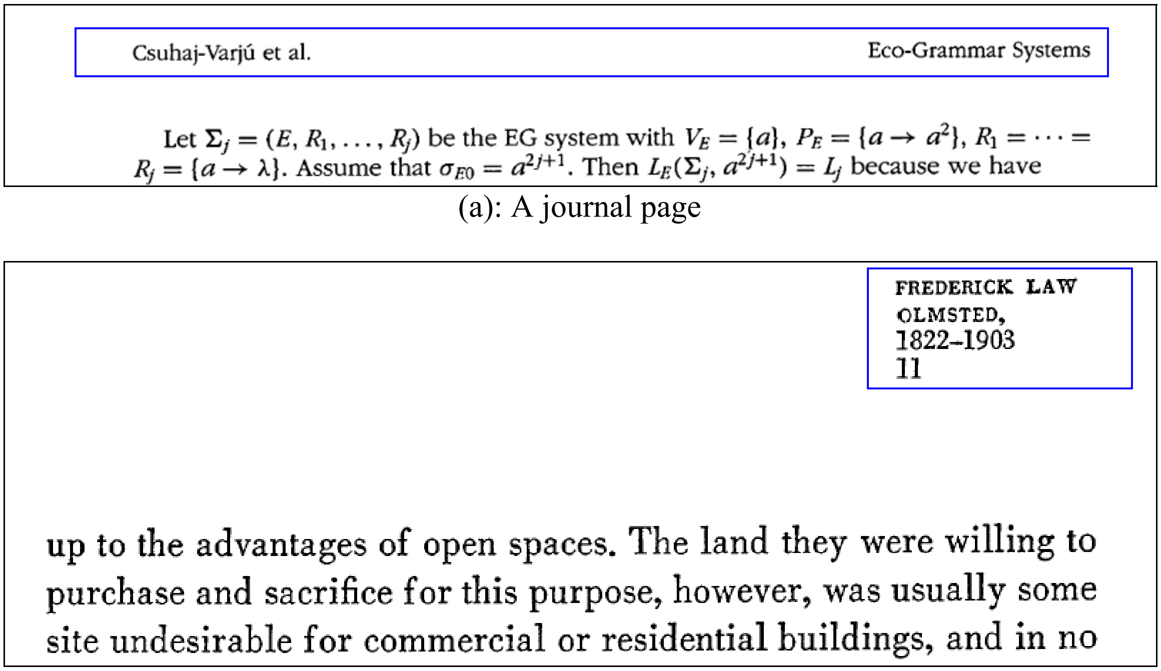
(a): The first page (Bottom part only)

Distinction between Mental, Physical Phenomena 49

whether a second person could see a first person's mental entity (a thought about a dog). Children uniformly said no. But they also uniformly said that the first person could not see his or her own mental entities. Thus, their answers could have just reflected an understanding that mental entities are invisible, not necessarily an understanding that mental states and entities are subjectively private. In the mental imagery studies, however, children often said that they could see their own images. This is a reasonable response given the quasi-

(b): The second page (Top part only)

Figure 1: Two neighboring pages



Csuhaj-Varjú et al. Eco-Grammar Systems

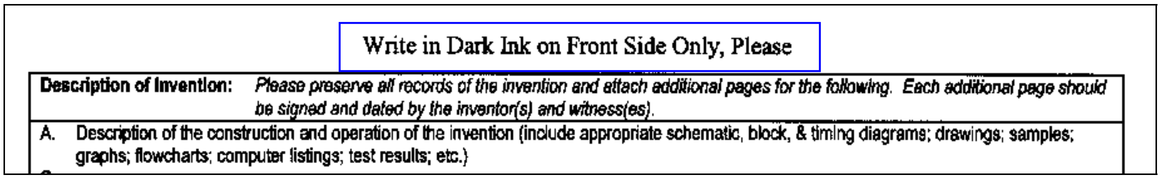
Let $\Sigma_j = (E, R_1, \ldots, R_j)$ be the EG system with $V_E = \{a\}$, $P_E = \{a \to a^2\}$, $R_1 = \cdots = R_j = \{a \to \lambda\}$. Assume that $\sigma_{E0} = a^{2j+1}$. Then $L_E(\Sigma_j, a^{2j+1}) = L_j$ because we have

(a): A journal page

FREDERICK LAW
OLMSTED,
1822–1903
11

up to the advantages of open spaces. The land they were willing to purchase and sacrifice for this purpose, however, was usually some site undesirable for commercial or residential buildings, and in no

(b): A book page

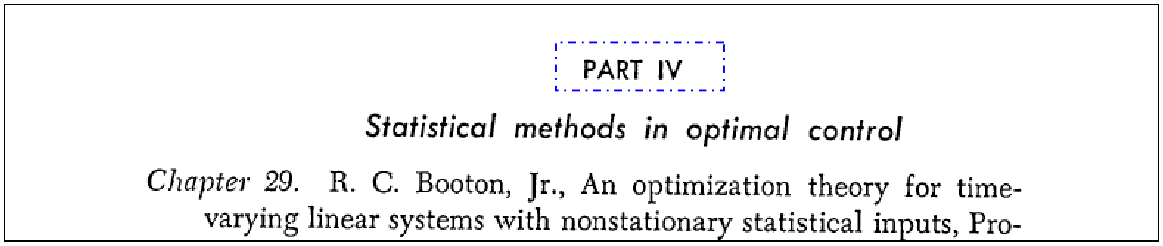


Write in Dark Ink on Front Side Only, Please

Description of Invention: *Please preserve all records of the invention and attach additional pages for the following. Each additional page should be signed and dated by the inventor(s) and witness(es).*

A. Description of the construction and operation of the invention (include appropriate schematic, block, & timing diagrams; drawings; samples; graphs; flowcharts; computer listings; test results; etc.)

(c): A form page



PART IV

Statistical methods in optimal control

Chapter 29. R. C. Booton, Jr., An optimization theory for time-varying linear systems with nonstationary statistical inputs, Pro-

(d): A book page

Figure 2: Different patterns of headers (Marked in solid rectangles in a-c) and potential false positives (Marked in dashed rectangles in d) (Only the top of each page is shown to save space.)