A mixed approach to auto-detection of page body

Conference Paper *in* Proceedings of SPIE - The International Society for Optical Engineering · January 2008

DOI: 10.1117/12.765886 · Source: DBLP

Liangcai Gao
Peking University
82 PUBLICATIONS 882 CITATIONS

Zhi Tang
Sichuan University
179 PUBLICATIONS 1,385 CITATIONS

A mixed approach to auto-detection of page body

Liangcai Gao, Zhi Tang, Ruiheng Qiu
Institute of Computer Science & Technology of Peking University

ABSTRACT

Page body holds the central information of a page in most documents. This paper addresses the problem of automatically detecting page body area in digital books or journals. A novel method based on font expansion and header and footer elimination is detailed. This method extracts body text font (BFont) and headers and footers from a document first, and then draws two page body bounding boxes for each page, one by analyzing the distribution of BFont in pages and the other by removing headers and footers from pages. Finally, the two bounding boxes are combined to obtain the resultant page body bounding box. The test results demonstrate very high recognition rate: up to 99.49% in precision.

Keywords: document analysis, header and footer, page body area, body text font

1. INTRODUCTION

A large amount of useful information can be extracted from electronic documents through document processing, which can be divided into two stages: document analysis and document understanding. Document analysis refers to the extraction of the layout structure from documents, and document understanding refers to mapping the layout structure into a logical structure after the former is extracted [1]. One important task of document analysis is to segment out the page body area (referred to simply as "BArea" henceforth) from a page, because that area carries the main information of a page. BArea is usually the remainder area after removing the area of header and footer from a page, as shown in Picture 1. The text in BArea is more informative than that in other parts of a document, so the extraction of BArea is often a preprocessing step of document information retrieval. Another major application of BArea extraction is in handheld devices. As mobile technology advances, handheld devices have rapidly gained popularity, and digital documents are more and more used in those devices. Handheld devices have relatively small display screens, so the documents often need to be simplified or reflowed prior to being loaded in order to make people read them comfortably. A common and reasonable simplification is to display only the content in the BArea on the small screens. Page headers and footers, especially page numbers, are omitted to retain the continuity of the text and to retrench the content of a document.

Unfortunately, currently available document viewers in handheld devices have not the capacity to filter the redundant information contained in headers and footers, and many popular file formats, such as pdf, ps and rtf, do not distinguish BArea from other areas. Various typesetting styles increase the difficulty of BArea extraction. For example, some documents have horizontal layouts but others have vertical ones. Especially, background pictures often confuse the delimitation of BArea. Although human operators can cut out the BArea from pages, the manual processing tends to be slow and expensive, and can not meet the need of batching or real-time document processing. Our research is thus motivated by the need of a system that enables the computers to automatically detect the BArea of digital documents. This work is also called as page body delimitation in the literature.

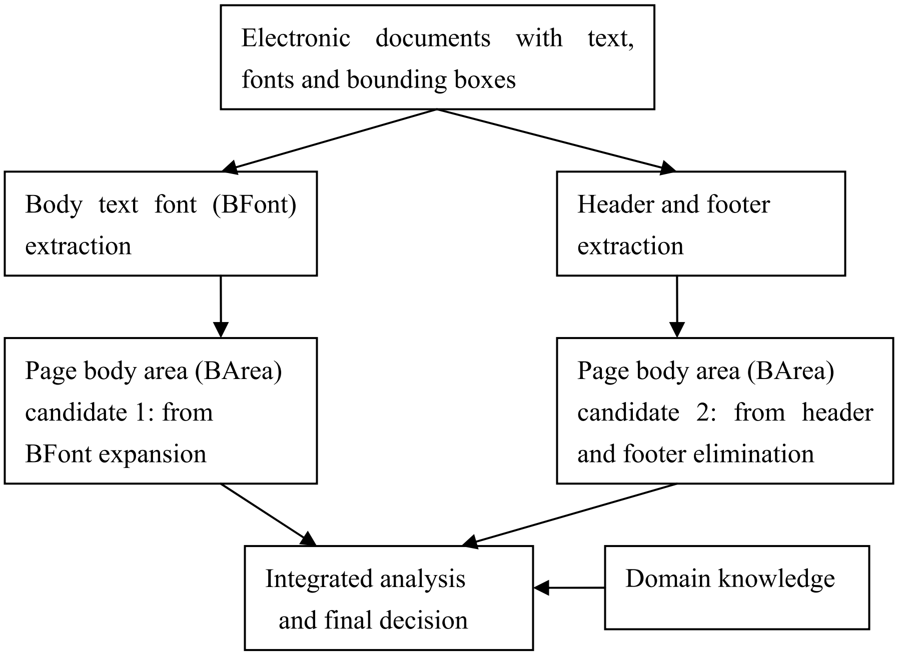
Document Recognition and Retrieval XV, edited by Berrin A. Yanikoglu, Kathrin Berkner,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 6815, 68150T, © 2008 SPIE-IS&T · 0277-786X/08/$18

The problem of judging the importance of different parts in a web page has attracted a lot of attention recently [2, 3, 4, 5], but the specific study on page body delimitation of book or journal documents has not received much attention. A relevant research is on header and footer extraction [6, 7]. Lin [6] and Déjean et al. [7] respectively proposed similar methods for detecting the header/footer by page association. The former measures the similarity scores between the top/bottom lines of successive pages in geometric position and string matching, and the later computes the textual variability of the multiple pages on the same position.

Section 2 illustrates our approach to page body delimitation, and section 3 shows the experimental results. We conclude and discuss future work in section 4.

2. PAGE BODY DELIMITATION

We develop a set of techniques for page body delimitation combined in a collaborative manner which aim at the book or journal documents (In the following part, for convenience, we name them book documents uniformly if not explicitly stated) with the text, font and bounding box information (the coordinates of the four corners). Figure 1 illustrates the proposed approach.


Electronic documents with text, fonts and bounding boxes

Body text font (BFont) extraction

Header and footer extraction

Page body area (BArea) candidate 1: from BFont expansion

Page body area (BArea) candidate 2: from header and footer elimination

Integrated analysis and final decision

Domain knowledge

Figure 1: Process for page body delimitation

The proposed approach exploits the following properties of a generic page body:

A. BArea usually occupies the most space of a page, and its size and position in a page is invariable through a whole document.

B. The text in BArea has large variability between different pages, while the text in headers and footers has fixed positions and similar content.

C. Fonts used in BArea are often different from fonts in header and footer, and the most frequently used font in a document usually is a kind of fonts employed in BArea.

D. The BAreas of odd pages and even pages are symmetrical horizontally, so once we obtain the BArea of odd pages, the BArea of even pages can be computed out using this symmetry.

To exploit these generic properties of page body, the following process is used.

2.1 Extraction of BFont

In this phase, we randomly draw several sample pages from a book document, check the font of every character or word (In the following part, for convenience, we name them character uniformly if not explicitly stated) in the sample pages, collect all the fonts adopted in the pages and count the number of characters using each font. We select the most frequently used font as BFont, since the number of characters in BArea is far larger than the number of characters in other parts of a book. Two questions are important here: how many sample pages to draw, and how to check the fonts. When determining the sample size, we draw out the total pages of a book if the book has no more than ten pages; otherwise ten pages are enough to get a good result according to the experimental result shown in figure 3. When checking all the fonts used in a book, two fonts are considered as the same if and only if they match in font type, color, size and style.

2.2 Header-footer extraction

Headers and footers are common formatting elements in book documents, and they surround outside BArea, as shown in Picture 1. The most stable feature of headers and footers is that they repeat in neighboring pages, and we utilize this feature to extract them.

The core algorithm in our proposal is similar to the method presented by Lin in [6], where a similarity score in text matching and geometric position between the top/bottom lines on neighboring pages is calculated and header and footer lines are determined according to the score. We modify his algorithm in how to select the compared pages, how to measure similarity, and how to process the case of vertical type headers/footers. Firstly, in many books the text of the headers and footers on neighboring odd (even) pages is usually similar but the text of the headers and footers on an odd page and its contiguous even page is different. So when page ***i*** is selected, we select page ***i*** + ***2*** and ***i*** – ***2*** rather than page ***k*** (***i-Win*** <= ***k*** <= ***i+Win***) adopted in [6] to compare their corresponding lines. Parameter ***Win*** is to control the number of neighboring pages. Secondly, we add a new similarity score in computing the similarity of the corresponding lines in neighboring pages to increase the measuring accuracy. That is, we count the number of matched characters in font, and give a score to each compared line according to the count. The score is higher with more matched characters in font. Thirdly, we deal with the case of vertical type headers/footers by rotate the pages by 90 degrees before segmenting pages into lines.

It is worthy to note that some kinds of files (for example, the files produced by Microsoft Word or Excel) contain the information of header and footer in their description parts, and thus the information of header and footer can be directly obtained from them without executing the above procedure of header-footer extraction.

Two different mechanisms to detect BArea are discussed in the following sections.

2.3 Method 1: BFont expansion

This method uses the fact that BFont is the most frequently used font in a document, as mentioned above in Property C of BArea. We employ a bottom-up approach to obtain a candidate of BArea. Figure 2 shows the whole process.

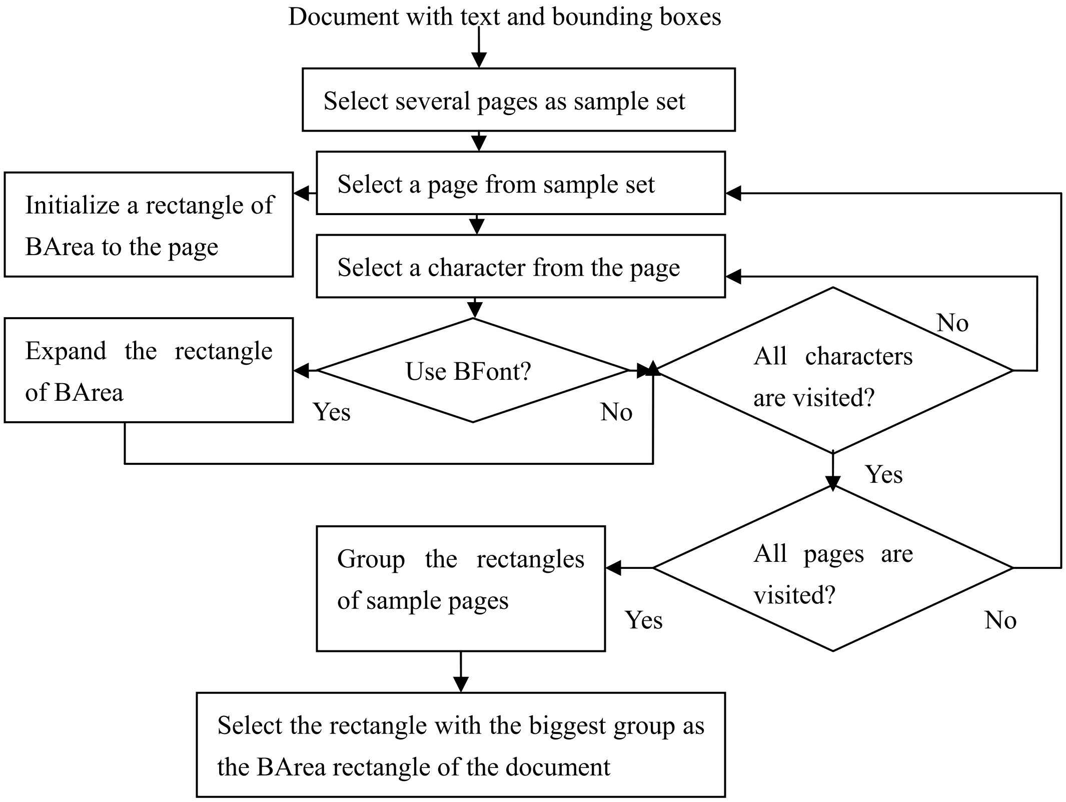

Document with text and bounding boxes
Select several pages as sample set
Select a page from sample set
Initialize a rectangle of BArea to the page
Select a character from the page
Expand the rectangle of BArea
Use BFont?
Yes
No
All characters are visited?
No
Yes
Group the rectangles of sample pages
All pages are visited?
Yes
No
Select the rectangle with the biggest group as the BArea rectangle of the document

Figure 2: Workflow of method 1 (BFont expansion)

Firstly, we randomly extract several odd (even) pages from a book document. For each page, we initialize a small rectangle of BArea in the page center, and then examine the font information of every character. Once we visit a character using BFont, we expand the rectangle to a new size so that it can contain the bounding box of the character. When all the characters in a page are visited, the resulting rectangle is the smallest rectangle that can cover all the bounding boxes of the characters using BFont in the page, and we call it the BArea rectangle. When all the selected pages are processed, each extracted page has one BArea rectangle. After that, we check the BArea rectangle of each extracted page. If two rectangles have very small differences (we set the distance as a quarter of the BFont size in the experiment) in their left, right, top and bottom borders, we consider the two rectangles as the same size rectangles. Then the BArea rectangles of the extracted pages are grouped by size, and the rectangles in every group are similar in size. We select a rectangle in the biggest group as the BArea candidate of the document.

2.4 Method 2: Header and footer elimination

This method uses the fact that a page is mainly composed of page body, header and footer in most books. Thus, BArea can be obtained in the way to remove the header and footer from a page.

Using header and footer extraction procedure described in section 2.2, we identify and remove the header and footer lines from a page and select the remainder area as another BArea candidate.

2.5 Combiner

In this phase, we analyze the two BArea candidates extracted from the above two methods. When headers or footers of a book document adopt BFont, the BArea from method 1 covers their zones. When a book document has no header or footer, the BArea from method 2 expands to page border in the directions without header or footer. In a few books, headers/footers are so close to page bodies that they can not be segmented away from the page bodies. So we combine the two BArea candidates from the above two methods to determine the final BArea.

We compare each border of the two BArea candidates from method 1 and method 2 in left, right, top and bottom directions, and choose the border closer to page center as the border of the final BArea.

Finally, the BArea of a book document is determined after considering some document style heuristics such as: BArea's size is less than page size, while BArea occupies most space of a page in every direction, etc.

3. EXPERIMENT AND ANALYSIS

In this section, we present preliminary experimental results demonstrating that our method can do a reasonably good job in detecting page bodies in digital books. The experiment data is composed of 820 books with various layout styles. The books are PDF files from which the information about fonts, characters and coordinate data in pages can be extracted.

We first test method 1 and method 2 on the experiment data. And then we test the combining method. We provide the precision (the percentage of BAreas extracted that are in fact true) of each method in detecting BArea. Table 1 shows the results.

Table 1: Precision for detecting page body

Methods	Precision
Method 1: BFont expansion	85.41%
Method 2: header-footer elimination	72.64%
Combining method	99.49%

In our experiments, a book file is the input, and a bounding rectangle of BArea is the output. If the bounding rectangle of a book is extracted exactly, we consider the extracted BArea of the book is true. In addition, it is natural that every book has a true bounding rectangle of BArea. The total number of output rectangles from each method is equal to the number of books, so the recall (the percentage of true BAreas that are extracted) is equal to the precision in this paper.

We track the executing procedure of our approach and analyze the errors in our experiment. The main causes of errors can be summarized as follows:

- Method 1: In some documents, headers/footers adopt the same font as BFont. As a result, the BArea from the method expands to the area of headers/footers. The layout of some commercial journals varies across pages, and BFont only appears in small and variable parts of pages. As a result, the BAreas extracted from all pages are different and small, no BArea is qualified as the final BArea candidate of a document.
- Method 2: Some books have no header/footer, and the BArea from the method expands to page border in the direction without header/footer. In a few books, headers/footers are so close to page bodies that they can not be segmented away from the page bodies.

- Combining Method: Method 1 and method 2 mutually support each other, and thus the combining method outperforms either single method. It hardly makes mistake except the documents with very complex and variable layout, such as some commercial journals, which have little similarity between headers and footers of neighboring pages and have too much image in most pages.

We also research the relation between the precision rate and the number of extracted pages on 100 books with header and footer, and the experimental result is shown in Figure 3. As we can see from Figure 3, the precision rates about BFont extraction, method 1 and method 2 generally increase with more extracted pages. The precision of BFont extraction is nearly 100% when the number of extracted pages is 10, so we select 10 pages to extract BFont. The precision rate of method 1 reaches the maximum when 14 pages are extracted. When the number of extracted pages goes above 16, the precision rate of method 2 hardly increases anymore. Thus we select 14 pages for method 1 and 16 pages for method 2 in the final combining method, and the execution time for a book is around 5 seconds on a personal computer (CPU 2GHZ, RAM 512MB). The separate running time of BFont extraction, method 1 or method 2 is nearly linear with the number of extracted pages. The most running time is consumed in accessing the information of text, fonts and coordinates of every extracted page.

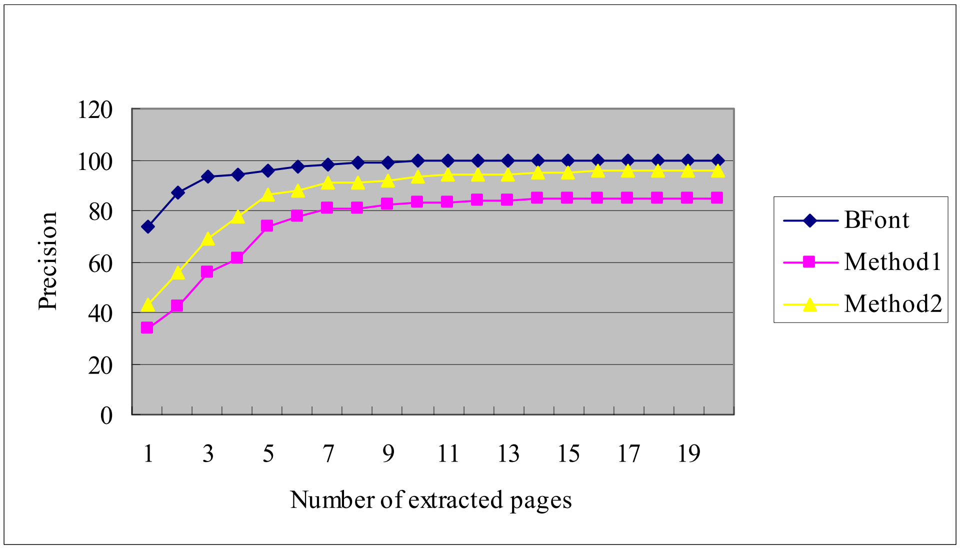


Figure 3: The relation between the precision rates and the number of test pages

4. CONCLUSION

In this paper, we describe a novel approach to automatically detecting page body after our research motivation and some relevant research work are introduced. The approach consists of two methods: BFont expansion, header and footer elimination, and the combination of both. A preliminary experiment demonstrates the effectiveness of this approach.

One future work is to expand our method to enable it to process documents in image format through defining similarity in pixel level or employing OCR software before using our method. And we will do some researches on document reflowing and document transplanting from computers to handheld devices, basing on the information of BArea.

5. ACKNOWLEDGEMENTS

This work is supported by National Project of Scientific and Technical Supporting Programs Funded by Ministry of Science & Technology of China (NO. 2006BAH02A21).

REFERENCES

[1] Y.Y. Tang, C.D. Yan and C.Y. Suen, "Document Processing for Automatic Knowledge Acquisition", IEEE Transactions on Knowledge and Data Engineering, **vol. 6**, pp. 3-21, 1994.

[2] S. Gupta, G. Kaiser, D. Neistadt and P. Grimm, "DOM-based content extraction of HTML documents", *Proc. 12th International World Wide Web Conference (WWW2003)*, pp. 207-214, Budapest, Hungary, 2003.

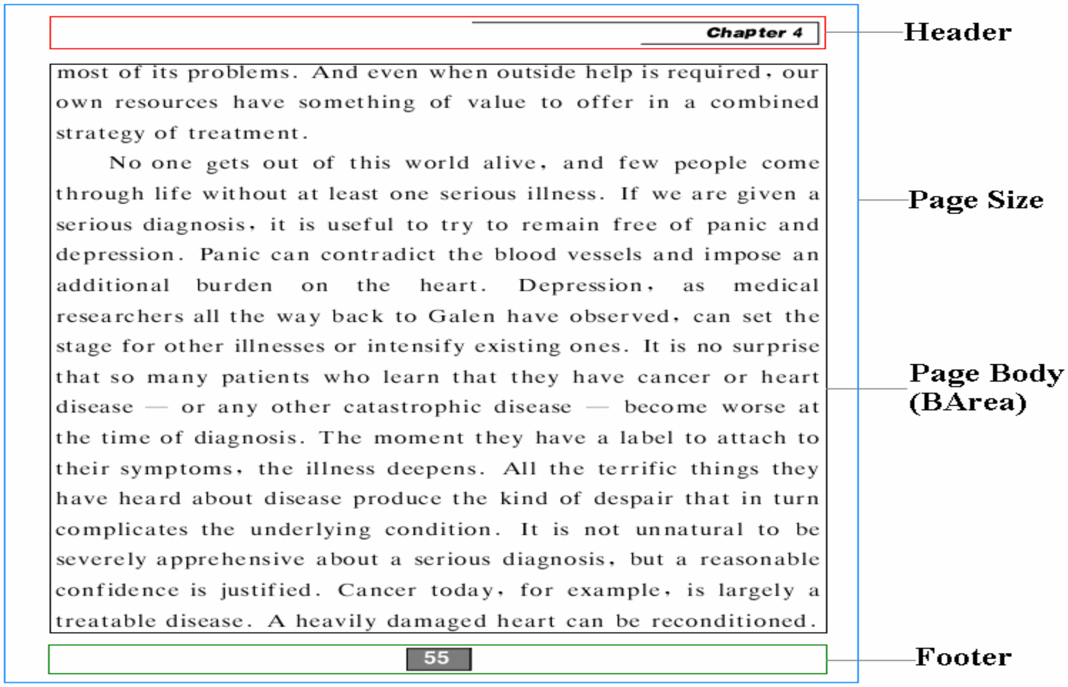
[3] M. Kovacevic, M. Diligenti, M. Gori and V. Milutinovic, "Recognition of common areas in a web page using visual information: a possible application in a page Classification", *the 2002 IEEE International Conference on Data Mining (ICDM 2002)*, pp. 250-257, Maebashi City, Japan, 2002.

[4] S.H. Lin and J.M. Ho, "Discovering informative content blocks from web documents", *Proc. 8th ACM SIGKDD*, pp. 588-593, Edmonton, Canada, 2002.

[5] R. Song, H. Liu, J.R. Wen and W.Y. Ma, "Learning block importance models for web pages", *Proc. 13th International World Wide Web Conference (WWW 2004)*, pp. 203-211, New York, USA, 2004.

[6] Xiaofan Lin, "Header and Footer Extraction by Page-Association", *Proc. SPIE Conference on Document Recognition and Retrieval X*, pp. 164-171, Santa Clara, USA, 2003.

[7] H Déjean and JL Meunier, "A System for Converting PDF Documents into Structured XML Format", *Proc. DAS'06*, pp. 129-140, Nelson, New Zealand, 2006.



Chapter 4

most of its problems. And even when outside help is required, our own resources have something of value to offer in a combined strategy of treatment.

No one gets out of this world alive, and few people come through life without at least one serious illness. If we are given a serious diagnosis, it is useful to try to remain free of panic and depression. Panic can contradict the blood vessels and impose an additional burden on the heart. Depression, as medical researchers all the way back to Galen have observed, can set the stage for other illnesses or intensify existing ones. It is no surprise that so many patients who learn that they have cancer or heart disease — or any other catastrophic disease — become worse at the time of diagnosis. The moment they have a label to attach to their symptoms, the illness deepens. All the terrific things they have heard about disease produce the kind of despair that in turn complicates the underlying condition. It is not unnatural to be severely apprehensive about a serious diagnosis, but a reasonable confidence is justified. Cancer today, for example, is largely a treatable disease. A heavily damaged heart can be reconditioned.

55

Picture 1: Page body area (BArea) illustration