(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 April 2003 (17.04.2003)

PCT

(10) International Publication Number
WO 03/032202 A2

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number: PCT/US02/32422

(22) International Filing Date: 9 October 2002 (09.10.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/972,055 9 October 2001 (09.10.2001) US

(71) Applicant: HEWLETT-PACKARD COMPANY [US/US]; 3000 Hanover Street, M/S 1051, Palo Alto, CA 94304-1112 (US).

(72) Inventors: CHAO, Hui; 1026 Craig Drive, San Jose, CA 95129 (US). SANG, Henry, W., Jr.; 21975 Hyannisport, Cupertino, CA 95014 (US).

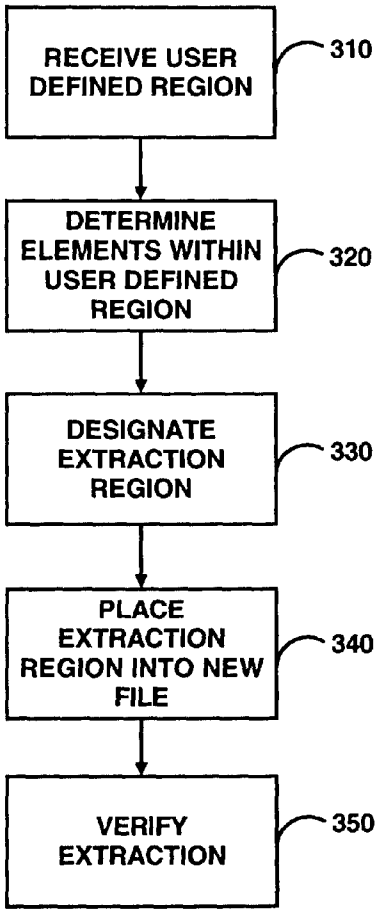
(74) Agent: HEMINGER, Susan, E.; Hewlett-Packard Company, IP Administration, P.O. Box 272400, Ft. Collins, CO 80527-2400 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK,

[Continued on next page]

(54) Title: SECTION EXTRACTION TOOL FOR PDF DOCUMENTS



(57) Abstract: A method of extracting a section of a page from a portable document format file ("pdf"): The method includes receiving (310) indication of a user-defined region (450a) on a pdf file page (200), designating (330) an extraction region (450b) including all elements (451-454) determined to be within the user-defined region, and placing (340) the extraction region into a new file. The method may also include determining (320) if one or more elements (210, 220, 230) on the pdf page are within the user-defined region (450a) by applying inclusion rules based on whether an element's bounding box (211, 221, 231, 241) is within or intersects the user-defined region (450a) in the original pdf document and the extracted region to bitmap images and comparing the two bitmap images, bit by bit.

WO 03/032202 A2

TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

SECTION EXTRACTION TOOL FOR PDF DOCUMENTS

FIELD OF THE INVENTION

The invention is generally related to electronic data files. More particularly, the invention is related to extraction of a section of a portable document format document.

BACKGROUND OF THE INVENTION

Electronic files may be created using a variety of techniques. Thus, it may be desirable to store data from an electronic file in a format that is independent of the process used to create it so that it may be accessible to a range of users. One format that allows such access is the portable document format. The portable document format ("pdf") is a file format for representing documents in a manner independent of the application software, hardware, and operating system used to create the documents and independent of the output device on which they are displayed or printed.

A PDF workflow assumes a one-way production process where the PDF file contains a rendition that is laid out for final presentation, i.e., no logical structural information is preserved. Consequently, one problem with storing documents in a pdf format is that it is difficult to reuse parts of documents because elements with semantic affinity are not stored as one logical group of elements. Although it is possible to store the original editable document as an attribute in the PDF file, this is not generally done, since the original program for creating the pdf document is unavailable anyway, or because this introduces a vulnerability for computer viruses. Without the original editable document, removing a portion of the pdf document for use in another document or file is not easily accomplished. For example, it may be desirable for a

user to insert a graph or chart from a pdf document into a document of the user's own creation or make a slide presentation with the graph or chart. The PDF specification makes an allowance to include structural information, however, very few pdf documents are created with such structural information due to size constraints and/or creation processes. Thus, most pdf documents do not generally support sharing or repurposing the content of the document and it is generally not possible to extract a figure, an illustration or a paragraph from a chapter as an integrated object from PDF.

There are a few techniques available for reusing pdf document content. However, some of these processes are complicated and require extensive user interaction, while others extract a raster rendition of the selected document portion from the display bitmap, thereby losing all original document structure and attribute information, as well as resolution, which is usually limited to the 72 dpi screen resolution.

SUMMARY OF THE INVENTION

An aspect of an embodiment of the invention is to provide a method for extracting a section of a portable document format ("pdf") document.

In one embodiment, the method may include receiving indication of a user defined region on a pdf file page, determining if each element on the pdf page is within the user defined region, designating an extraction region including all elements determined to be within the user defined region, and placing the extraction region into a new pdf file.

Those skilled in the art will appreciate these and other advantages and benefits of various embodiments of the invention upon reading the following detailed description of preferred embodiments with reference to the below-listed drawings.

Another aspect of the invention includes checking the extracted region for accuracy. In one embodiment, both the extracted region and the region in the original document may be converted to bitmap images and compared bit by bit.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention is illustrated by way of example and not limitation in the accompanying figures in which like numeral references refer to like elements, and wherein:

Fig. 1 is a block diagram illustrating one embodiment of an extraction tool;

Fig. 2 illustrates an example of the structure of a portable document format document;

Fig. 3 is a flow diagram illustrating an exemplary embodiment of a method for extracting a section of a portable document format page; and

Fig. 4 is a block diagram illustrating an example of an extraction region determination process.

DETAILED DESCRIPTION OF THE INVENTION

In the following detailed description, numerous specific details are set forth in order to provide a thorough understanding of the invention. However, it will be apparent to one of ordinary skill in the art that these specific details need not be used to practice the invention. In other instances, well known structures, interfaces, and processes have not been shown in detail in order not to obscure unnecessarily the invention.

Fig. 1 is a block diagram illustrating one embodiment of an extraction tool. Extraction tool 100 may include an input/output module 110, a section determination module 120, a memory module 130, a document generation module 140, a verification module 150 and

processing module 160. The modules 110-160 are shown to be located within extraction tool 100 for conceptual purposes only. In other embodiments, one or more of the modules 110-160 may reside outside of the extraction tool 100 and may be called upon by the extraction tool 100 as needed.

The input/output module 110 may accept instructions from a user, such as instructions for extracting a section of a portable document format file. These instructions may include the user drawing a box or other shape to outline the section of the pdf file the user would like to extract to a new document, such as a new pdf file. The input/output module 110 may also present the user with instructions or messages regarding the performance of the extraction, such as, for example presenting the user with a message regarding the accuracy of the extraction, as described below with regard to Fig. 3.

The section determination module 120 may determine what elements of the pdf file should be included in the new document. For example, if the user defined region includes parts of elements, the section determination module 120 may apply rules of inclusion to determine if the element should be included in the region to be extracted to the new document.

The memory module 130 may be used to store image information, data, instructions or any other information usable for extracting a section of a pdf file. For example, the memory may be used to store the user defined region while the section determination module 120 determines what elements will be included in the extraction region.

The document generation module 140 may generate the new document by extracting the elements in the region determined by the section determination module 120 into the new document. In one embodiment, the new document generation module 140 may extract the elements in the extraction region into a new pdf file.

The verification module 150 may verify the accuracy of the extracted region in the new document generated by document generation module 140. In one embodiment, the verification module 150 may convert the original document and the new document generated by document generation module 140 into bitmap images for comparison, as described below with respect to Fig. 3.

The processing module 160 may execute the processes described with respect to Fig. 3 below, using instructions received from modules 110, 120, 140 and 150. For example, the processing module 160 may increase the size of the user defined region based on rules of inclusion received from the section determination module 120. An example of an inclusion rule is to fully include all elements that intersect the user defined region.

Fig. 2 illustrates an example of the structure of a pdf document. A pdf document 200 may include a text element(s) 210, a graphic element(s) 220 and image element(s) 230. The text element(s) 210consist of text runs, which are runs of characters with the same attribute. A text run element 240 is a representation of a text run. Graphic elements 220 are arbitrary shapes made up of a sequence of straight lines, rectangles and cubic Bezier curves. Image elements 230 are sequences of pixels obtained by scanning the image arrow in row or column. Each element 210, 220, 230 may exist within its corresponding bounding box 211, 221, 231.

Bounding boxes are rectangles which surround objects in a document, and may refer to the smallest rectangle which entirely encloses the object on a page. The bounding box location and size for each element may be obtained, for example, through ADOBE's ACROBAT™ Software Development Tool Kit Application Programmer Interface, where a bounding box is guaranteed to encompass the element, but is not necessarily the smallest box that contains the element. To achieve higher accuracy of extraction result, a bounding box may be modified to be

the smallest bounding box containing the element. For example, for a rectangular shaped graphic element, the bounding box may be modified to be the outline of the rectangle itself.

Bounding boxes are invisible to a viewer of a document. In the exemplary pdf document 200, text element 210 exists within bounding box 211, each of the graphic elements 220 exists within its associated bounding box 221 and each image elements in 230 exists within its associated bounding boxes 231. Fig. 3 is a flow diagram illustrating an exemplary embodiment of a method for extracting a section of a portable document format ("pdf") page. It will be appreciated that the process described with regard to Fig. 3 does not require all of the steps described and the order of the steps may vary depending on design.

At step 310, the extraction tool 100 receives an indication of a user defined region of a pdf page for extraction. In one embodiment, the user may draw a rectangle or other shape around a region of interest to the user to identify the region for extraction. Such a rectangle or shape is referred to as a selection marquee. In one embodiment, the user may use an object recognition tool to identify the region for extraction. In one embodiment, a user may use a graphic select tool that is available in ADOBE ACROBAT™ to draw the region of interest. The user may then click on the extraction processing icon for module 160 of the extraction tool 100 from a menu or toolbar. In one embodiment, the object recognition tool is a part of the input/output module 110.

At step 320, the extraction tool 100 may determine what elements of the original pdf page are within the user defined region for extraction received through input/output module 110. In one embodiment, the section determination module 120 determines what elements of the original pdf are within the user defined region for extraction. Since the bounding boxes 211, 221, 231 of elements are not visible to the user and a bounding box may be bigger than the actual element the

region of interest chosen by the user may not include the all of the element's bounding box. Thus, the section determination module 120 may apply inclusion (or alternatively, exclusion) rules to determine which elements should be extracted based on the user defined region of interest.

In one embodiment, the inclusion rules may be based on the type of element. For example, a graphic or image element 220, 230 may be determined to be within the extraction region only if its entire bounding box 221, 231 is within the user defined region. Thus, if the bounding box 221, 231 of a graphic or image element 220, 230 intersects with the user defined region, but is not completely within the user defined region, the graphic or image element 220, 230 will not be included in the extraction process.

In one embodiment, a text element 210 or part of the text element 210 may be included in the extraction region if all or a part of its bounding box 211 intersects with the user defined region of interest. In one embodiment, if the bounding box 211 of the text element 210 intersects the user defined region, the section determination module 120 may evaluate whether the sub-elements, or text-run elements 240, of the text element 210 are within the user defined region. If a text-run element's bounding box 241 is completely within the user defined region, or if any part of the text-run element's bounding box intersects the user defined region, the user defined region of interest may be expanded to include the entire bounding box of the text-run element in the region for extraction.

Since bounding boxes of text run elements 240 are sometimes much larger than the text itself, the user defined region may not include the entire bounding box of the text run element 240. Thus, including any text run element 240 that intersects the user defined region of interest would help to include all of the elements chosen by the user for extraction.

At step 330, the extraction tool 100 may designate an extraction region. At the end of the determination step 320, the extraction region may be defined to include all of the elements determined to be included in the extraction.

At step 340, the extraction tool 100 may place the extraction region into a new file. In one embodiment, the document generation module 140 may create a second pdf document and insert the extracted region into the second pdf. In another embodiment, the document generation module 140 may insert the extraction region into an already existing second pdf, or a desktop publishing software document, such as, for example, a ADOBE FRAMEMAKER™ or ADOBE INDESIGN™ document. In one embodiment, the user may be asked to choose a file into which the extraction region may be inserted when the user requests the extraction of a selected region.

At step 350, the extraction tool 100 may check the extracted region in the second pdf document for differences from the user defined region in the original file. In one embodiment, the verification module 150 may verify the accuracy of the second pdf document by converting the original document extraction region, defined at step 330, to a first bitmap image and the second pdf document's extraction region to a second bitmap image. After aligning the two bitmaps, the verification module 150 may then compare the second bitmap image to the first bitmap image, bit by bit.

If there are differences between the two images, the extraction tool 100 may inform the user of the differences by presenting the user with a message through input/output module 110. For example, the extraction tool 100 could attach a verification message to the second pdf document letting the user know that there are differences between the extracted image placed in the second pdf document and the extraction region defined in the original pdf document.

Fig. 4 is a block diagram illustrating an example of an extraction region determination process. A document 401 may include graphic or image elements 452-454 and text element 451. After the user indicates a user defined region 450a, the user defined region is input 410 into the section determination module 420. The section determination module 420 determines which elements of document 401 should be included in the user defined region. As shown, the user defined region 450a, is expanded to include a text run element 451 when the extraction region 450b is designated 330, although text run element 451 only intersected the user defined region 450a.

The method for extracting a user defined region described allows a user to select a region in a pdf document and select the option of extracting the region. The extraction tool 100 needs no further interaction from the user. The tool 100 allows a user to reuse selected content of a pdf document without having to learn or perform complicated processes.

Steps 310 - 350, described above, may be compiled into computer programs. These computer programs can exist in a variety of forms both active and inactive. For example, the computer program can exist as software comprised of program instructions or statements in source code, object code, executable code or other formats. Any of the above can be embodied on a computer readable medium, which include storage devices and signals, in compressed or uncompressed form. Exemplary computer readable storage devices include conventional computer system RAM (random access memory), ROM (read only memory), EPROM (erasable, programmable ROM), EEPROM (electrically erasable, programmable ROM), and magnetic or optical or magneto-optical disks or tapes. Exemplary computer readable signals, whether modulated using a carrier or not, are signals that a computer system hosting or running the computer program can be configured to access, including signals downloaded through the

Internet or other networks. Concrete examples of the foregoing include distribution of executable software program(s) of the computer program on a CD ROM or via Internet download. In a sense, the Internet itself, as an abstract entity, is a computer readable medium. The same is true of computer networks in general.

While this invention has been described in conjunction with the specific embodiments thereof, it is evident that many alternatives, modifications and variations will be apparent to those skilled in the art. These changes and others may be made without departing from the spirit and scope of the invention.

What is claimed is:

1. A method of extracting a section of a page (200) from a portable document format file ("pdf") comprising:

receiving (310) indication of a user-defined region (450a) on a pdf file page (200);

determining (320) if one or more elements (210, 220, 230) on the pdf page are within the user-defined region;

designating (330) an extraction region (450b) including all elements (451-454) determined to be within the user-defined region (450a); and

placing (340) the extraction region into a new file.

2. The method of claim 1, wherein determining if one or more elements (210, 220, 230) are within the user-defined region comprises applying extraction determination rules to each element based on element type.

3. The method of claim 2, wherein the element type comprises at least one of graphic element (220), image element (230) and text element (210).

4. The method of claim 3, wherein applying the extraction determination rules comprises:

including a graphic element (220) within the extraction region if a bounding box (221, 452-454) of the graphic element is within the user-defined region (450a);

including an image element (230) within the extraction region if a bounding box (231, 452-454) of the image element is within the user-defined region (450a);

including a text element (210) within the extraction region if a bounding box (211) of the text element is within the user-defined region (450a);

evaluating if sub-elements (240) of the text element (210) are within the user-defined region (450a) if the bounding box (211) of the text element intersects the user-defined region (450a);

including a sub-element (240) of the text element (210) if a bounding box (451, 241) of the sub-element is within the user-defined region (450a); and

expanding the user-defined region to include a sub-element (240) of the text element (210) if the bounding box (241, 451) of the sub-element of the text element intersects the user-defined region.

5. The method of claim 1, further comprising verifying (350) the accuracy of the extracted user-defined region (450a) in the new file.

6. The method of claim 5, wherein verifying (350) the accuracy of the extracted user-defined region in the new file comprises converting the pdf file page (200) into a first bitmap image and the extracted user-defined region in the new file into a second bitmap image and comparing the first bitmap image to the second bitmap image bit by bit to confirm the accuracy of the extraction.

7. The method of claim 1, wherein receiving the indication of the user-defined region on the pdf file page comprises one of receiving an input of a user-defined region drawn on the pdf file page and receiving an user selection of a button on the pdf screen after the user draws the user-defined region on the pdf file page.

8. An apparatus extracting a section of a page of a portable document format file comprising:

a processor (160) configured to perform the steps of:

receiving (310) indication of a user-defined region (450a) on a pdf file page (200);

determining (320) if one or more elements (210, 220, 230) on the pdf page are within the user-defined region;

designating (330) an extraction region (450b) including all elements (451-454) determined to be within the user-defined region (450a); and

placing (340) the extraction region into a new file.

9. A computer readable medium containing executable instructions which, when executed in a processing system, cause the system to perform a method comprising:

receiving (310) indication of a user-defined region (450a) on a pdf file page (200);

determining (320) if one or more elements (210, 220, 230) on the pdf page are within the user-defined region;

designating (330) an extraction region (450b) including all elements (451-454) determined to be within the user-defined region (450a); and

placing (340) the extraction region into a new file.

10. The computer readable medium of claim 9 wherein the method further comprises verifying (350) the accuracy of the extracted user-defined region in the new file.

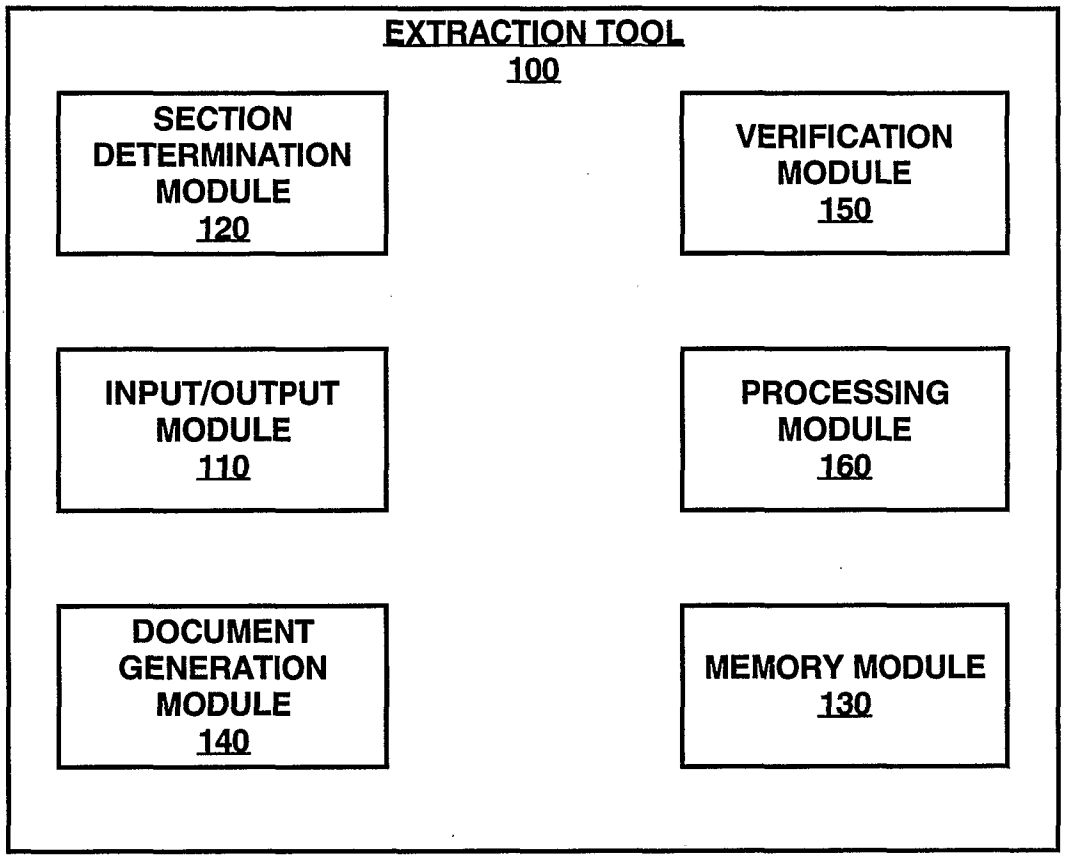

EXTRACTION TOOL
100
SECTION DETERMINATION MODULE
120
VERIFICATION MODULE
150
INPUT/OUTPUT MODULE
110
PROCESSING MODULE
160
DOCUMENT GENERATION MODULE
140
MEMORY MODULE
130

FIG. 1

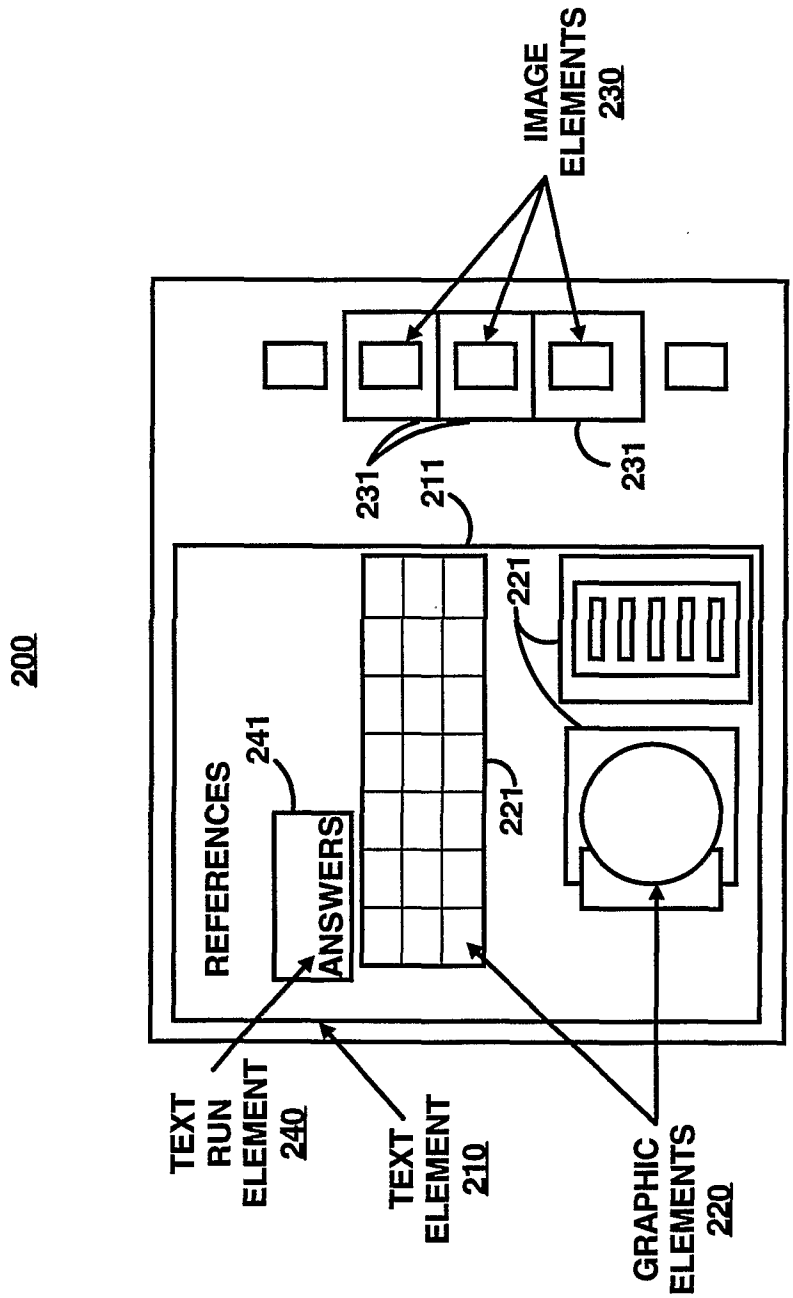

200
IMAGE ELEMENTS 230
231
211
231
221
221
241
REFERENCES
ANSWERS
TEXT RUN ELEMENT 240
TEXT ELEMENT 210
GRAPHIC ELEMENTS 220

FIG. 2

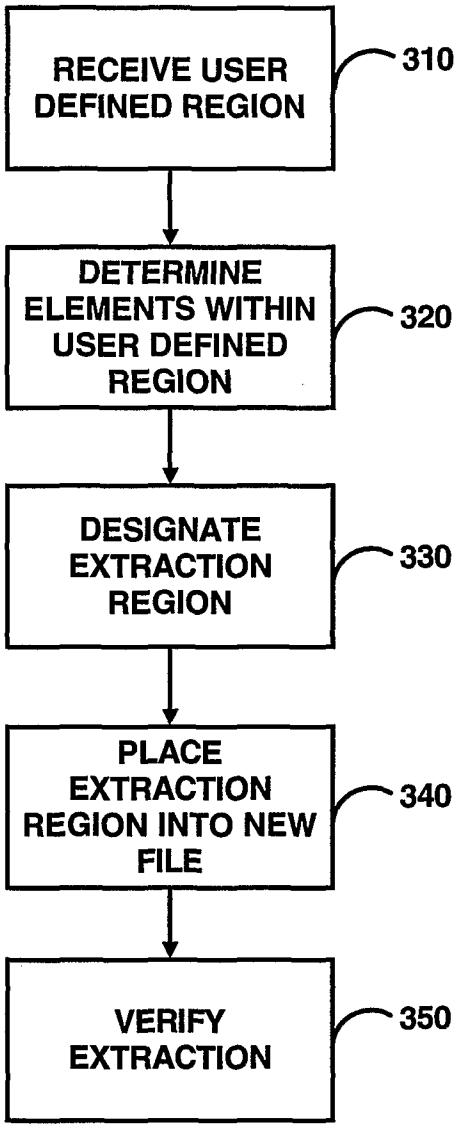

RECEIVE USER DEFINED REGION
310
DETERMINE ELEMENTS WITHIN USER DEFINED REGION
320
DESIGNATE EXTRACTION REGION
330
PLACE EXTRACTION REGION INTO NEW FILE
340
VERIFY EXTRACTION
350

FIG. 3

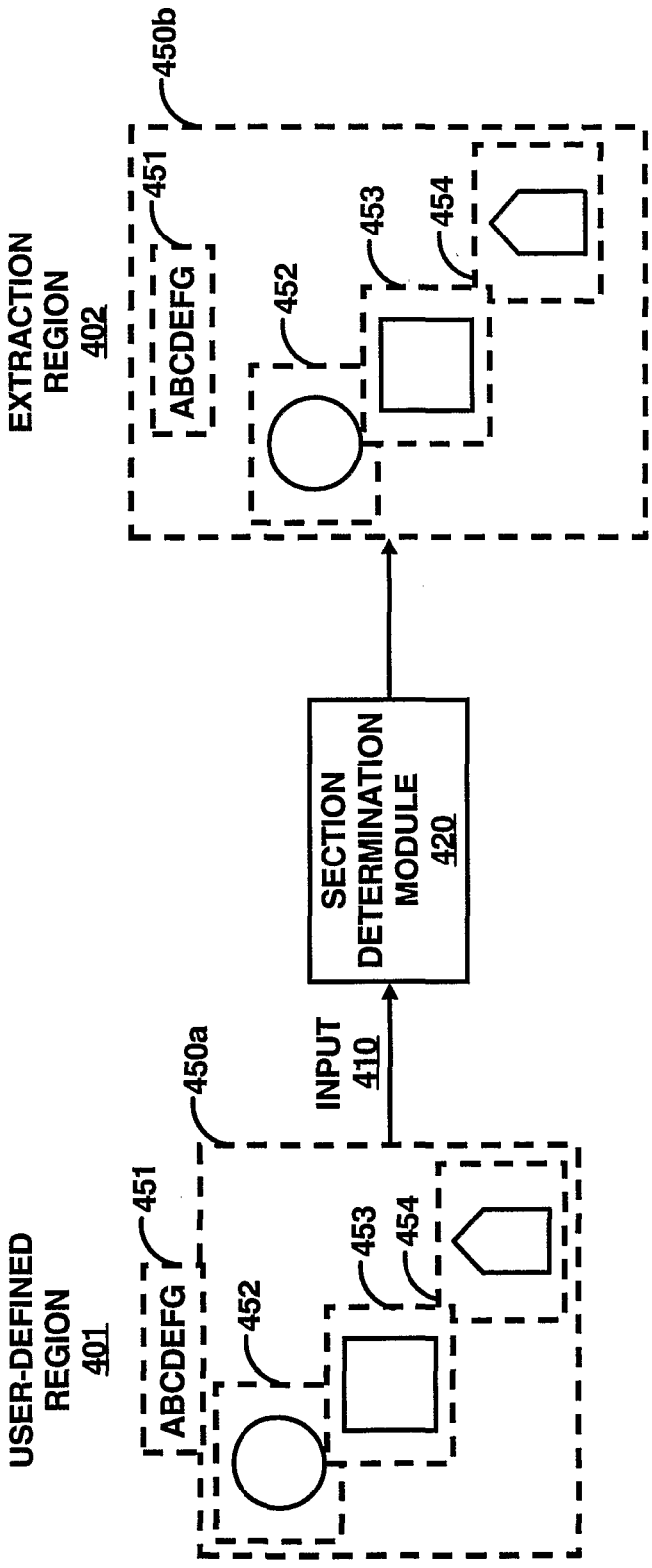

USER-DEFINED REGION 401
ABCDEFG
451
452
453
454
450a
INPUT 410
SECTION DETERMINATION MODULE 420
EXTRACTION REGION 402
ABCDEFG
451
452
453
454
450b

FIG. 4