Title: DOCUMENT INFORMATION MINING TOOL

Abstract: Embodiments of the present invention provide methods, computer-readable media, and systems for extracting referential keys from a document. A document is parsed to identify at least one key, the key being identified from at least one contextual indication. The key is classified according to a key type, the key type being identified from the contextual indication. The key is extracted and then stored in a location in a structured shell with the location corresponding to the key type. As a result, the key can be found by a search seeking one of the key and the key-type allowing a searcher to identify the document from which the key was extracted.
DOCUMENT INFORMATION MINING TOOL

INVENTORS
Brent L. Hadley
Patrick J. Eames
Susan C. Chew

FIELD OF THE INVENTION

This invention relates generally to information extraction and, more specifically, to processing documents to enhance access to information stored in the documents.

BACKGROUND OF THE INVENTION

People often need to access information that is recorded in documents. Such documents may range in length from a single page to many volumes. Certainly, the longer the documents are, the more difficult it is to access the specific information desired. As a result, long or complex documents include multiple tables and indices to facilitate hierarchical and key word searches for the information of interest. These tables and indices are time-consuming to create. Furthermore, although the tables and indices are helpful in locating information, using them is time-consuming since each page referred to in the table or index must be individually examined to find the entry of interest. The foregoing may be repeated numerous times before the substantive entry of interest is found.

The proliferation of computers has revolutionized how information is accessed. Accordingly, computer-readable documents may now be searched using various software routines for terms of interest in the document. In particular, hypertext linking has enabled referencing and cross-referencing of key terms simply by using a pointing device to select a point of interest in a document.

Despite the great advantages that computers permit in accessing and retrieving information, processes for information retrieval still can be improved. For example, creating a hyper-link to a source document involves human intervention to identify the term with which the link will be created and associating it with the link to the related information. Further, to create hyper-linked information, all the documents or at least those documents containing the links need to be computer-readable documents. As a result, a non-computer-readable document may be scanned so that it is computer-viewable, but unless the document is computer-readable such as a
text or graphics document, it is generally not possible to associate links with portions of the
document.

Similarly, even though a referenced document need not be computer-readable to be
accessed from a link, if the reference target is not computer-readable, then a person linking to the
document may be required to manually navigate through the target document to find the
information of interest. Certainly, the task becomes even more complicated if one desires
information in both documents and needs to switch back and forth between the documents. In
such cases, to avoid the difficulty of navigating back and forth, a user desiring portions of such
documents will typically print the needed documents or parts of the documents. When users
print such documents instead of accessing them on the computer clearly undermines one of the
objectives of making such documents accessible by computer.

To avoid the complexities of moving back and forth between documents, one possible is
to extract information from documents that is expected to be relevant. Unfortunately, removing
only the content from the documents may present other problems. For example, some regulatory
agencies require that extracted content be verified as accurately including the content of the
original document before it can be used. This verification is a time-consuming and costly
process. In addition, extraction of content also may obliterate inferential information a user
might otherwise obtain from the document. Such inferential information might exist as an
interrelationship between parts, or as an annotation regarding other parts that might be useful,
and other similarly useful information. Extracting information expected to be relevant thus may
obliterate other useful information.

**SUMMARY OF THE INVENTION**

The various embodiments of the present invention are useful in mining documents to
identify keys that facilitate the creation of references within the documents and that may also
form a basis for links between documents. The various embodiments exploit the fact that that
documents of certain types adhere to formatting standards that may include titles, section
headings, and other parts of the document that are useful for navigating to and within the
document. The various embodiments may also utilize pattern formats that may be present in
documents of various types, such as character patterns that may include a number of
characters having expected separators so that keys may be identified when that pattern appears on
the page. Once these parts of the document representing potentially useful navigational keys
have been identified, the keys are stored in a structured format associated with the original
document to facilitate linking the keys with related documents. In one non-limiting example,
embodiments of the present invention are used to mine a .pdf type document to find textual and
graphical keys in the document, using optical character recognition as needed to extract text from
graphically-presented keys, and to store the keys in an extensible markup language (XML)
document to facilitate creating links to the keys in the original .pdf document.

More particularly, embodiments of the present invention provide methods, computer-
readable media, and systems for extracting referential keys from a document. A document is
parsed to identify at least one key, the key being identified from at least one contextual
indication. The key is classified according to a key type, the key type being identified from the
contextual indication. The key is extracted and then stored in a location in a structured shell with
the location corresponding to the key type. As a result, the key can be found by a search seeking
one of the key and the key-type allowing a searcher to identify the document from which the key
was extracted.

**BRIEF DESCRIPTION OF THE DRAWINGS**

The preferred and alternative embodiments of the present invention are described in
detail below with reference to the following drawings.

FIGURE 1 is a block representation of an embodiment of the present invention.

FIGURE 2 is a document to be mined according to an embodiment of the present
invention.

FIGURE 3 is a flowchart of a routine for mining a document according to an embodiment
of the present invention.

FIGURES 4-7 are examples of portions of documents mined to extract keys into a
structured format.

FIGURE 8 is a structured format collecting a collection of keys mined from a document.

FIGURE 9 is a system according to an embodiment of the present invention.

**DETAILED DESCRIPTION OF THE INVENTION**

The present invention relates generally to a system and method for linking related
documents and, more specifically, to identifying, extracting, and collecting keys used in linking
the related documents. Many specific details of certain embodiments of the invention are set
forth in the following description and in FIGURES 1-9 to provide a thorough understanding of
such embodiments. One skilled in the art, however, will understand that the present invention
may have additional embodiments, or that the present invention may be practiced without several of the details described in the following description.

More particularly, embodiments of the present invention provide methods, computer-readable media, and systems for extracting referential keys from a document. A document is parsed to identify at least one key, the key being identified from at least one contextual indication. The key is classified according to a key type, the key type being identified from the contextual indication. The key is extracted and then stored in a location in a structured shell with the location corresponding to the key type. As a result, the key can be found by a search seeking one of the key and the key-type allowing a searcher to identify the document from which the key was extracted.

Using prior knowledge of the formats employed by authors in creating documents, for a given document format (e.g., .pdf files) and document type (e.g., aircraft maintenance manuals), the document information mining tool reviews the document for elements an author of the document has included that can be used to navigate to and within the document. For example, the document mining tool identifies titles, headers, footers, page numbers, off-page references, annotations to graphics, and other authored elements that a human reader of the document might use to navigate for additional information. The document mining tool identifies these navigational attributes automatically by dissecting the document. The information mined is then stored using a suitable structured format, such as an XML document, allowing the mined information to be used as keys or tags to support linking to other documents. The document mining tool is further suitably equipped with an optical character reading (OCR) capability to analyze text that appears in graphic form in order to mine it from the document.

The document information mining tool thus provides a low-level, literal extraction process by parsing a document, identifying sections of the document that can be used as navigational points to navigate to, from, and throughout the document, and storing the mined information in a useful structure.

FIGURE 1 is a block diagram of a generalized system 100 for mining a source document 110 according to an embodiment of the invention. The document 110 is subjected to examination according to mining rules 120 directed to identify and extract various navigational keys, as will be described in greater detail below. The extracted keys are subsequently stored in a structured format 130. The structured format 130 may include an XML document, although it will be appreciated that the extracted keys may also be stored in a substantive content-based
document. In one particular embodiment, the keys are stored in a structured format used to facilitate navigation to and within the source document. In still another particular embodiment,

The mined information is used to navigate through the existing documents to facilitate access to the complete, underlying documents.

Referring still to FIGURE 1, in another particular embodiment, keys may be created by a user and transferred to the mining rules 120. The source document 110 may then be parsed to locate at least one portion of the document 110 that includes a contextual indication corresponding to the key created by the user. The user may then review the information located within the at least one portion identified during the parsing.

FIGURE 2 shows a sample document 200, which is a page selected from an aircraft maintenance manual (AMM). The document 200 is formatted as a .pdf document and presented on a .pdf reader 210. It will be appreciated that the document 200 could have been generated from a computer-readable document or scanned from a printed document. For purposes of this description, a computer-readable document is one in which characters and figures are generated for display on a computer screen from character and figure representations. In contrast, non-computer-readable documents are documents received in image form because they were scanned into or otherwise stored in a computing system in an image format stored such as rasterized data. The sample document 200 represents a document that will be mined to extract the keys in the document, which may include titles, task designations, and other information, that might be useful in navigating to, from, and within the document 200.

FIGURE 3 is a generalized flowchart of a method 300 for mining a document such as the sample document 200 (FIGURE 2) according to an embodiment of the invention. The routine 300 begins at a block 302. At a block 304, the document to be mined is received. As previously described, the document may be a computer-readable document or non-computer-readable image-type document. At a block 306, a next area of a document is parsed. The exemplary routine 300 parses a document area by area. However, alternatively, the document could be parsed looking for particular features of the document as will be further described below.

At a decision block 310 it is determined if a placement indicator is identified. A placement indicator may indicate presence of a title, a header, footer, a page number, or another key having relevance as indicated by the position of the key on a page. Accordingly, the manner in which the key is detected may be based upon generalized knowledge of the formatting methods used in a document, or upon tribal knowledge of the formatting methods, or other
specific document formatting rules. Generalized knowledge, for example, includes knowledge that indicates that a series of numerals centered or in a corner of a header or footer of a series of pages represent page numbers. General knowledge also may include, for example, that a term appearing at a top of a page generally includes a title. By contrast, tribal knowledge includes knowledge known to a group of persons familiar with a specific type of document. For example, in certain industries or organizations, it may be common practice to place a subject identifier in a header or footer of each page, even though the inclusion of that key may not be indicated by a specific rule. Finally, specific document formatting rules are rules literally dictated by document conventions. For example, it may be an established practice in an industry or organization to position system identifiers on a page of a document in a generally similar location. For example, the format used in legal briefs submitted to courts include titles, page numbers, and other elements placed in generally similar locations according to rules specified by the courts.

Still referring to FIGURE 3, once a placement indicator is identified at the decision block 310, the routine 300 proceeds to a decision block 312 where it is determined if the placement is determinative of what type of key has been identified. One area of a document could include a placement-indicated key, a format-indicated key, and a font-indicated key, such as, for example, if footer includes both a page number and title or other identifier. The position of a numeral alone, such as in a part of a header reserved for a particular identifier may alone be determinative. On the other hand, other aspects of the key may have to be considered. If it is determined at the decision block 312 that placement is determinative, the routine 300 proceeds to a block 314 where the key is extracted and stored in a structured format. If the document is a computer-readable document, characters representing the key are copied. On the other hand, if the document is a non-computer-readable image-type document as previously described, an optical character recognition (OCR) routine is invoked to scan the identified key to extract the characters from the key.

If a placement indicator is not identified or placement alone is not determinative, the routine proceeds to a decision block 316 where it is determined if a format indicator is identified. A format-indicated key may include a string of digits identifiable because of their format. For example, a number including three numbers followed by a separator such as a dash, followed by three more numbers, another dash, and four more numbers might be expected to be a phone number. A number including three numbers followed by a dash, followed by two more numbers followed by another dash, followed by four more numbers might be expected to be a Social
Security number. In a further example, a string of characters followed by an "@" symbol followed by another string of characters including at least one period might be expected to be an e-mail address. Accordingly, a type of key can be determined based upon the format.

Once a format indicator is identified at the decision block 316, the routine 300 proceeds to a decision block 318 where it is determined if the format is determinative of what type of key has been identified by itself or in combination with the placement indicator. If it is determined at the decision block 318 that format alone or with the placement is determinative, the routine 300 proceeds to a block 314 where the key is extracted and stored in a structured format. If the document is a computer-readable document, characters representing the key are copied. On the other hand, if the document is a non-computer-readable image-type document as previously described, an optical character recognition (OCR) routine is invoked to scan the identified key to extract the characters from the key.

If placement or format indicators are not identified, or they are not determinative alone, the routine proceeds to a decision block 320 where it is determined if a font indicator is identified. A font-indicated key may include a key having a type signified by its font or typeface. For example, according to commonly accepted document conventions, information that is underlined, or that appears in bold type, or is presented in an enlarged font, or any combination of the foregoing may represent a title, a part name, a document identifier, or similar information.

Once a font indicator is identified at the decision block 320, the routine 300 proceeds to a decision block 322 where it is determined if the font is determinative of what type of key has been identified by itself or in combination with other indicators. If it is determined at the decision block 322 that format alone or in combination with other indicators is determinative, the routine 300 proceeds to a block 314 where the key is extracted and stored in a structured format. If the document is a computer-readable document, characters representing the key are copied. On the other hand, if the document is a non-computer-readable image-type document as previously described, an optical character recognition (OCR) routine is invoked to scan the identified key to extract the characters from the key.

If indicators are not identified or are not determinative, the routine proceeds to a decision block 326 where it is determined if all areas of the document have been parsed. If not, the routine loops to the block 306 where the next area of the document will be parsed. On the other hand, if all areas of the document have been parsed, the routine 300 ends at a block 328.
Although FIGURE 3 shows a particular sequential ordering of the steps 310, 316 and 320, it will be appreciated that the order of the steps 310, 316 and 320 may be altered to suit a particular application. Moreover, the steps 310, 316 and 320 may also be executed in parallel, so that keys based on placement, format and font may be identified simultaneously.

FIGURES 4-7 illustrate examples of how keys of interest are mined from a document. The examples 400, 500, 600, and 700 from FIGURES 4, 5, 6, and 7, respectively show examples of keys being extracted from a page of an aircraft maintenance manual (AMM). The keys are extracted and then stored in a structured document that, in these examples, are XML documents.

Example 400 (FIGURE 4) shows a key being extracted from a document segment 402. In particular, the key includes a section title, "INTEGRATED DRIVE GENERATOR (IDG) – REMOVAL/INSTALLATION." The key is identified by rules 404 recognizing a combination of indicators using general, tribal, and/or other knowledge. In terms of placement, the potential key begins at a left side of a page and is a first line of a page body. In terms of font, the potential key is all UPPERCASE. The potential key is thus identified as a section title and identified as a key of interest. The key is extracted and stored in a structured format 406, in this case an XML document, where the key is designated as a title so that if can be located for navigation purposes.

Example 500 (FIGURE 5) shows a second key being extracted from a document segment 502. The key includes a task identifier, an expected and sought-after key in an AMM. The task identifier is determined in this case by format indicators. The task identifier includes a string of two digits, two digits, two digits, three digits, and three digits all separated by dashes. Accordingly, a rule 504 identifies this format as a task identifier, which is thus identified as a key. The key is extracted and stored in an XML document 506 where the key is designated as a task number so that it can be located for navigation purposes.

Example 600 (FIGURE 6) shows a third key being extracted from a document segment 602. The key includes a chapter number (24-11-11), which may be another expected and sought-after key in an AMM. The chapter number is determined in this case by placement and format indicators. In the present example, the chapter number is always on a right side of a page, and is always in the form of two digits, two digits, and two digits separated by dashes. The present example also includes a page block (401) and a revision date (Jun 10/01) that follow the formatted chapter number. Again, a rule 604 identifies the placement and format indicators in the document segment 602 to identify the key. The key is then extracted and stored in an XML document 606 where the key is designated as a chapter number so that it can be located for
navigation purposes. It will be appreciated that upon finding a formatted key such as the chapter number, the chapter number maybe further subdivided and stored in its constituent parts, including chapter (24), section (11), subject (11), and page block (401).

Example 700 (FIGURE 7) shows a fourth key being extracted from a document segment 702. The key includes a particular substantive section that, in this case, is "Effectivity" that is used to designate the applicability of a selected portion of the AMM to a particular aircraft. The substantive section is determined in this case by placement and format indicators. In the present example, the substantive section of interest is always on a left side of a page (a placement indicator) and always follows the term "Effectivity." As in the previous example, a rule 704 identifies the potential key as the substantive description "Effectivity" and identifies it as a key of interest. The key is extracted and stored in an XML document 706 where the key is designated as "Effectivity" so that it can be located for navigation purposes.

FIGURE 8 shows an XML document 800 that collects the extracted keys. The XML document 800 shows a page number 802, a chapter number 804 subdivided into its constituent parts, a title 806, a substantive section 808 (Effectivity), and a task identifier 810. The keys 802, 804, 806, 808, and 810 all are in a computer-readable format and are associated with the original document from which the keys 802, 804, 806, 808, and 810 were extracted. Thus, the XML document 800 can be searched or otherwise accessed to find the keys 802, 804, 806, 808, and 810 of interest. The keys 802, 804, 806, 808, and 810 are in turn linked to the document 800 from which they were drawn. Thus, the mining tool facilitates access to the document of interest and the portion of interest.

FIGURE 9 is a block diagram of an exemplary system 900 according to an embodiment of the present invention. The system 900 is presented with a computer-readable or non-computer-readable document 910. The system 900 uses a parser 920 to scan the document for potential keys as previously described. The system also employs an identifier 930 that is programmed to use placement indicators, format indicators, font indicators, and other indicators to determine whether a potential key found by the parser is a key of interest. If it is a key of interest, a classifier 930 receives the key, classifies it according to indicators manifested by the key of interest, and stores the key in a structured format 950, such as in XML data. As previously described in connection with FIGURE 8, the XML data may then be used to navigate to, from, and within the document 910.
While preferred and alternate embodiments of the invention have been illustrated and described, many changes can be made without departing from the spirit and scope of the invention. Accordingly, the scope of the invention is not limited by the disclosure of the preferred and alternate embodiments. Instead, the invention should be determined entirely by reference to the claims that follow.
What is claimed is:

1. A method for extracting referential keys from a document, the method comprising:
   parsing the document to identify at least one key, the key being identified from at least one indicator;
   classifying the key according to a key type, the key type being identified from the at least one indicator;
   extracting the key; and
   storing the key in a location in a structured shell, the location corresponding to the key type such that the key can be found by a search seeking at least one of the key and the key-type allowing a searcher to identify the document from which the key was extracted.

2. The method of Claim 1, wherein the document includes one of a computer-readable content and a non-computer-readable content.

3. The method of Claim 1, wherein the document includes a .pdf document.

4. The method of Claim 1, wherein the key type includes at least one of a title, a header, a footer, a document type, a document identifier, a subject identifier, a section identifier, a chapter identifier, and a page number.

5. The method of Claim 1, wherein the indicator includes at least one of a placement indicator, a format indicator, and a font indicator.

6. The method of Claim 5, wherein the indicator is derived from knowledge including at least one of a general knowledge, a tribal knowledge, and a specific document formatting rules.

7. The method of Claim 5, wherein the placement indicator includes a page position.

8. The method of Claim 5, wherein the format indicator includes a character pattern including at least one of a pattern of digits and a pattern of separators.

9. The method of Claim 5, wherein the font indicator includes at least one of a typeface, a boldface, an underlined text portion, and a font size.
10. The method of Claim 1, wherein extracting the key includes copying the key.

11. The method of Claim 1, further comprising using optical character recognition to extract the key when the document includes a non-computer-readable content.

12. The method of Claim 1, wherein the structured shell includes an extensible markup language (XML) document.

13. A method for extracting referential keys from a document, the method comprising:
   parsing a .pdf document to identify at least one key, the key being identified from at least one indicator, the indicator including at least one of a placement indicator, a format indicator, and a font indicator;
   classifying the key according to a key type, the key type being identified from the at least one indicator;
   extracting the key; and
   storing the key in a location in an extensible markup language (XML) document, the location corresponding to the key type such that the key can be found by a search seeking at least one of the key and the key-type allowing a searcher to identify the document from which the key was extracted.

14. The method of Claim 13, wherein the document includes one of a computer-readable content and a non-computer-readable content.

15. The method of Claim 13, wherein the key type includes at least one of a title, a header, a footer, a document type, a document identifier, a subject identifier, a section identifier, a chapter identifier, and a page number.

16. The method of Claim 13, wherein the indicator is derived from knowledge including at least one of a general knowledge, a tribal knowledge, and a specific document formatting rule.

17. The method of Claim 13, wherein the placement indicator includes a page position.

18. The method of Claim 13, wherein the format indicator includes a character pattern including at least one of a pattern of digits and a pattern of separators.
19. The method of Claim 13, wherein the font indicator includes at least one of a typeface, a boldface, a underlined text portion, and a font size.

20. The method of Claim 13, wherein extracting the key includes copying the key.

21. The method of Claim 13, further comprising using optical character recognition to extract the key when the document includes a non-computer-readable content.

22. A computer-readable medium having stored thereon instructions for extracting referential keys from a document, the computer-readable medium comprising:
   a first computer program portion adapted to parse the document to identify at least one key, the key being identified from at least one indicator;
   a second computer program portion adapted to classify the key according to a key type, the key type being identified from the at least one indicator;
   a third computer program portion adapted to extract the key; and
   a fourth computer program portion adapted to store the key in a location in a structured shell, the location corresponding to the key type such that the key can be found by a search seeking one of the key and the key-type allowing a searcher to identify the document from which the key was extracted.

23. The computer-readable medium of Claim 22, wherein the document includes one of a computer-readable content and a non-computer-readable content.

24. The computer-readable medium of Claim 22, wherein the document includes a .pdf document.

25. The computer-readable medium of Claim 22, wherein the key type includes at least one of a title, a header, a footer, a document type, a document identifier, a subject identifier, a section identifier, a chapter identifier, and a page number.

26. The computer-readable medium of Claim 22, wherein the indicator includes at least one of a placement indicator, a format indicator, and a font indicator.
27. The computer-readable medium of Claim 26, wherein the indicator is derived from knowledge including at least one of a general knowledge, a tribal knowledge, and a specific document formatting rule.

28. The computer-readable medium of Claim 26, wherein the placement indicator includes a page position.

29. The computer-readable medium of Claim 26, wherein the format indicator includes a character pattern including at least one of a pattern of digits and a pattern of separators.

30. The computer-readable medium of Claim 26, wherein the font indicator includes at least one of a typeface, a boldface, a underlined text portion, and a font size.

31. The computer-readable medium of Claim 22, wherein extracting the key includes copying the key.

32. The computer-readable medium of Claim 23, further comprising a fifth computer program portion configured to use optical character recognition to extract the key when the document includes a non-computer-readable content.

33. The computer-readable medium of Claim 22, wherein the structured shell includes an extensible markup language (XML) document.

34. A computer-readable medium for extracting referential keys from a document, the computer-readable medium comprising:

   a first computer program portion adapted to parse a .pdf document to identify at least one key, the key being identified from at least one indicator, the indicator including at least one of a placement indicator, a format indicator, and a font indicator;

   a second computer program portion adapted to classify the key according to a key type, the key type being identified from the at least one indicator;

   a third computer program portion adapted to extract the key; and

   a fourth computer program portion adapted to store the key in a location in an extensible markup language (XML) document, the location corresponding to the key type such that the key can be found by a search seeking one of the key and the

- 14 -
key-type allowing a searcher to identify the document from which the key was extracted.

35. The computer-readable medium of Claim 34, wherein the document includes one of a computer-readable content and a non-computer-readable content.

36. The computer-readable medium of Claim 34, wherein the key type includes at least one of a title, a header, a footer, a document type, a document identifier, a subject identifier, a section identifier, a chapter identifier, and a page number.

37. The computer-readable medium of Claim 34, wherein the indicator is derived from knowledge including at least one of a general knowledge, a tribal knowledge, and a specific document formatting rule.

38. The computer-readable medium of Claim 34, wherein the placement indicator includes a page position.

39. The computer-readable medium of Claim 34, wherein the format indicator includes a character pattern including at least one of a pattern of digits and a pattern of separators.

40. The computer-readable medium of Claim 34, wherein the font indicator includes at least one of a typeface, a boldface, an underlined text portion, and a font size.

41. The computer-readable medium of Claim 34, wherein extracting the key includes copying the key.

42. The computer-readable medium of Claim 34, further comprising using optical character recognition to extract the key when the document includes a non-computer-readable content.

43. A system for extracting referential keys from a document, the system comprising:
   a parser configured to parse the document to identify at least one key, the key being identified from at least one indicator;
   an identifier configured to identify the key according to a key type, the key type being identified from the at least one indicator; and
   a classifier configured to extract the key and store the key in a location in a structured shell, the location corresponding to the key type such that the key can be found by
a search seeking one of the key and the key-type allowing a searcher to identify the document from which the key was extracted.

44. The system of Claim 43, wherein the document includes one of a computer-readable content and a non-computer-readable content.

45. The system of Claim 43, wherein the document includes a .pdf document.

46. The system of Claim 43, wherein the key type includes at least one of a title, a header, a footer, a document type, a document identifier, a subject identifier, a section identifier, a chapter identifier, and a page number.

47. The system of Claim 43, wherein the indicator includes at least one of a placement indicator, a format indicator, and a font indicator.

48. The system of Claim 47, wherein the indicator is derived from knowledge including at least one of a general knowledge, a tribal knowledge, and a specific document formatting rule.

49. The system of Claim 47, wherein the placement indicator includes a page position.

50. The system of Claim 47, wherein the format indicator includes a character pattern including at least one of a pattern of digits and a pattern of separators.

51. The system of Claim 47, wherein the font indicator includes at least one of a typeface, boldface, an underlined text portion, and a font size.

52. The system of Claim 43, wherein extracting the key includes copying the key.

53. The system of Claim 43, further comprising an optical character recognizer configured to use optical character recognition to extract the key when the document includes non-computer-readable content.

54. The system of Claim 43, wherein the structured shell includes an extensible markup language (XML) document.
55. A system for extracting referential keys from a document, the system comprising:
   a parser configured to parse a .pdf document to identify at least one key, the key
   being identified from at least one indicator, the indicator including at least one of
   a placement indicator, a format indicator, and a font indicator;
   an identifier configured to identify the key according to a key type, the key type being
   identified from the at least one indicator;
   a classifier configured to extract the key and store the key in a location in an
   extensible markup language (XML) document, the location corresponding to the
   key type such that the key can be found by a search seeking one of the key and the
   key-type allowing a searcher to identify the document from which the key was
   extracted.

56. The system of Claim 55, wherein the document includes one of a computer-readable
    content and a non-computer-readable content.

57. The system of Claim 55, wherein the key type includes at least one of a title, a header, a
    footer, a document type, a document identifier, a subject identifier, a section identifier, a chapter
    identifier, and a page number.

58. The system of Claim 55, wherein the indicator is derived from knowledge including at
    least one of a general knowledge, a tribal knowledge, and a specific document formatting rule.

59. The system of Claim 55, wherein the placement indicator includes a page position.

60. The system of Claim 55, wherein the format indicator includes a character pattern
    including at least one of a pattern of digits and a pattern of separators.

61. The system of Claim 55, wherein the font indicator includes at least one of a typeface, a
    boldface, an underlined text portion, and a font size.

62. The system of Claim 55, wherein extracting the key includes copying the key.

63. The system of Claim 55, further comprising using optical character recognition to
    extract the key when the document includes a non-computer-readable content.
64. A method of information searching in a document, comprising:

creating a reference key corresponding to at least one contextual indicator present in the
document;
parsing successive portions of the document;
identifying at least one portion of the document that includes the reference key; and
reviewing the information included in the at least one portion.

65. The method of claim 64, wherein creating a reference key further comprises creating
the key according to a predetermined key type.

66. The method of claim 65, wherein the key type comprises at least one of a title, a header,
footer, a document type, a document identifier, a subject identifier, a section identifier, a
chapter identifier and a page number.

67. The method of claim 64, wherein the indicator comprises at least one of a placement
indicator, a format indicator and a font indicator.

68. The method of claim 67, wherein the format indicator comprises a character pattern that
includes at least one of a pattern of digits and a pattern of separators.

69. The method of claim 67, wherein the font indicator comprises at least one of a typeface,
boldface, an underlined text portion and a font size.
Begin

Receive Document to be Mined

Parse Next Area of Document

Placement Indicator Identified?

Yes

Yes Placement Determinative?

Yes

Store Key as Placement Indicated Key

No

No

Format Indicator Identified?

Yes

Yes Format Determinative?

No

All Areas of Document Parsed?

Yes

End

Fig. 3
AMM Segment

INTEGRATED DRIVE GENERATOR (IDG) – REMOVAL/INSTALLATION

1. General
   A. This procedure has two tasks:
      (1) The IDG Removal
      (2) The IDG Installation.

- Extract Section Title
- First line in page body
- All uppercase
- Begins at left side of page

XML Segment

<TITLE>INTEGRATED DRIVE GENERATOR (IDG)...</TITLE>

Fig. 4
AMM Segment

TASK 24-11-11-000-801

2. Integrated Drive Generator (IDG) Removal (Fig. 401)
   A. General
   (1) The IDG is found on the accessory gearbox of

504

*Always of the form "TASK ###..."

XML Segment

<TASK><NUMBER>24-11-11-000-801</NUMBER></TASK>

Fig. 5
AMM Segment

24-11-11
Page 401
Jun 10/01

- Found on right side of page.
- Always of the form "##-##-##-##".

XML Segment

<ATA><CHAPTER>24</CHAPTER>...</ATA>

Fig. 6
AMM Segment

EFFECTIVITY
ALL

XML Segment

\[<\text{EFFECTIVITY}><\text{GENERAL}>\text{ALL}</\text{GENERAL}>>\text{/EFFECTIVITY}>\]

Fig. 7
<PAGE>
  <NUMBER>54</NUMBER>
  <ATA>
    <CHAPTER>24</CHAPTER>
    <SECTION>11</SECTION>
    <SUBJECT>11</SUBJECT>
    <PAGEBLOCK>401</PAGEBLOCK>
  </ATA>
  <TITLE>INTEGRATED DRIVE GENERATOR (IDG) - REMOVAL/INSTALLATION</TITLE>
  <EFFECTIVITY>
    <GENERAL>ALL</GENERAL>
  </EFFECTIVITY>
  <TASK>
    <NUMBER>24-11-11-000-801</NUMBER>
    <TITLE>Integrated Drive Generator (IDG) Removal</TITLE>
  </TASK>
...
</PAGE>
**INTERNATIONAL SEARCH REPORT**

**A. CLASSIFICATION OF SUBJECT MATTER**

IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic database consulted during the international search (name of database and, where practical, search terms used)

EPO-Internal, PAJ, WPI Data, IBM-TDB, INSPEC

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>US 6 424 982 B1 (VOGEL CLAUDE) 23 July 2002 (2002-07-23) column 2, lines 15-50 column 3, line 7 - column 6, line 46</td>
<td>1,22,43, 64</td>
</tr>
</tbody>
</table>

Further documents are listed in the continuation of box C. Patent family members are listed in annex.

* Special categories of cited documents:
  *A* document defining the general state of the art which is not considered to be of particular relevance
  *E* earlier document but published on or after the international filing date
  *L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
  *O* document referring to an oral disclosure, use, exhibition or other means
  *P* document published prior to the international filing date but later than the priority date claimed
  *T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
  *X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
  *Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
  *&* document member of the same patent family

Date of the actual completion of the international search 15 September 2005

Date of mailing of the international search report 22/09/2005

Name and mailing address of the ISA
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Ridderkerk Tel. (+31-70) 340-2240, Tx. 31 651 epo nl, Fac. (+31-70) 340-3016

Authorized officer Bowler, A

Form PCT/ISA/210 (second sheet) (January 2004)
<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>paragraphs '0029!' - '0031!' paragraph '0065!'; figure 2</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>the whole document</td>
<td>3,10-21, 24, 31-42, 45,52-63</td>
</tr>
<tr>
<td>Patent document cited in search report</td>
<td>Publication date</td>
<td>Patent family member(s)</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>-----------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CA 2366485 A1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP 2002541580 T</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WO 0062155 A1</td>
</tr>
<tr>
<td>US 2002065857</td>
<td>30-05-2002</td>
<td>NONE</td>
</tr>
</tbody>
</table>