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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 January 2004 (15.01.2004)

PCT

(10) International Publication Number
WO 2004/006133 A1

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

(21) International Application Number: PCT/US2002/021198

(22) International Filing Date: 3 July 2002 (03.07.2002)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **IOTAPI., COM, INC.** [US/US]; 514 Bryant, Suite 119, Palo Alto, CA 94301 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **DEHLINGER, Peter, J.** [US/US]; 58 Roosevelt Circle, Palo Alto, CA 94306 (US). **CHIN, Shao** [US/US]; 955 Mears Street, Stanford, CA 94305 (US).

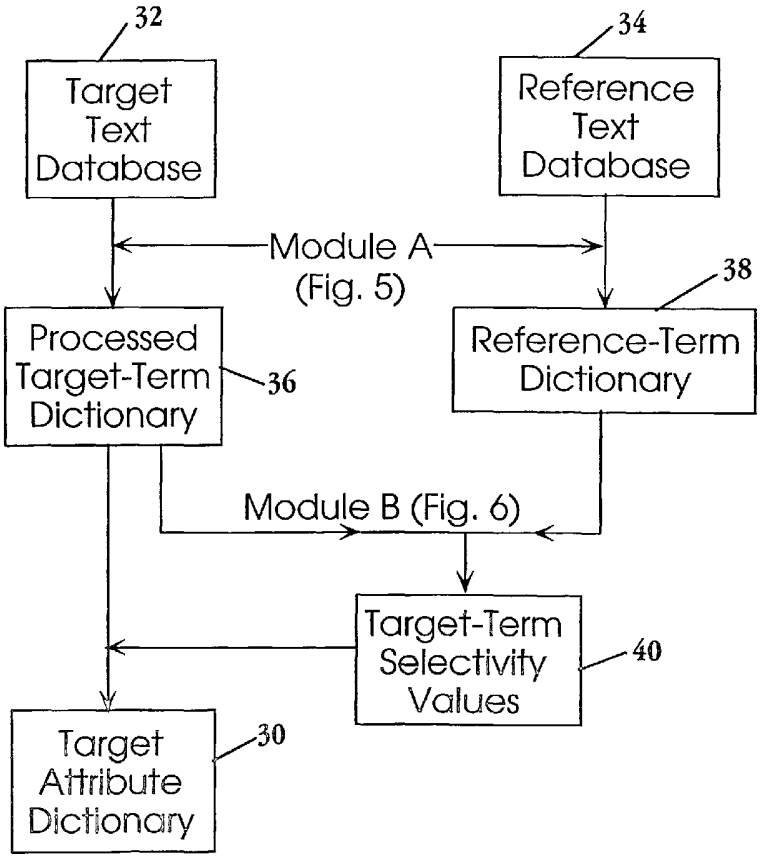
(74) Agents: **DEHLINGER, Peter,** et al.; Perkins Coie LLP, P.O. Box 2168, Menlo Park, CA 94026 (US).

(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, US, 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,

[Continued on next page]

(54) Title: TEXT-MACHINE CODE, SYSTEM AND METHOD



(57) Abstract: An automated method, system and machine readable code for comparing an input text in a selected field with each of a plurality of natural-language texts in the same field are disclosed. In the method, each of a plurality of terms characterizing the input text (32) is associated with a selectivity value (40) related to the frequency of occurrence of that term in a database (38) of digitally processed texts in the selected field (30), relative to the frequency of occurrence of the same term in a database of digitally processed texts in one or more unrelated fields. For each of the plurality of natural-language texts in the same field, a match score related to the number of terms derived from that text that match those of the input text, weighted by selectivity values of the matching terms, is determined. The highest-score matches are then determined.

WO 2004/006133 A1

ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search 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.

TEXT-MATCHING CODE, SYSTEM AND METHOD

Field of the Invention

This invention relates to the field of text matching, and in particular, to a method, machine-readable code, and system for matching one text with each of a plurality of texts in a related field.

Summary of the Invention

The invention includes, in one aspect, an automated method of comparing a target concept, invention, or event in a selected field with each of a plurality of natural-language texts in the same field. In practicing the method, each of a plurality of terms composed of non-generic words and, optionally, word groups characterizing the target concept, invention, or event, is associated with a selectivity value related to the frequency of occurrence of that term in a database of digitally processed texts in the selected field, relative to the frequency of occurrence of the same term in a database of digitally processed texts in one or more unrelated fields.

The method then determines for each of the plurality of natural-language texts in the same field, a match score related to the number of terms derived from that text that match those in the target concept, invention, or event, preferably weighted by selectivity values of the matching terms. Those natural-language texts in the same field having the highest match score or scores are then identified.

For use in comparing a target concept, invention, or event expressed in a natural-language input text, the method further includes constructing the plurality of terms by (a) identifying non-generic words in the input text, and (b) constructing from the identified words, a plurality of words groups, each containing two or more non-generic words that are proximately arranged in the input text.

To determine a match score, the method may assign to each of the terms in the target concept, invention, or event, a match value related to the corresponding selectivity value, and sum the match values of terms that match those of each of the plurality of digitally processed texts in the given field.

Information about the highest match-score texts may be displayed displaying as a two-dimensional matrix, one dimension representing terms

contained in the target concept, invention, or event, and the other dimension representing the highest matching encoded texts. In another embodiment, the information may be displayed as a list of highest matching encoded texts, and for each such text, a list of matching terms identified for that text.

The method is useful, for example, for comparing a concept, invention, or event in a selected technical field, or in a selected legal field.

In another aspect, the invention includes an automated system for comparing a target concept, invention, or event in a selected field with each of a plurality of natural-language texts in the same field. The system includes a database which provides, or from which can be determined, the selectivity value for each of a plurality of terms composed of non-generic words and, optionally, word groups representing proximately arranged non-generic words derived from a plurality of digitally-encoded natural-language texts (i) in the selected field and (ii) in one or more unrelated fields. The selectivity value of any term is determined from the frequency of occurrence of that term derived from the plurality of digitally-encoded natural-language texts in the selected field, relative to the frequency of occurrence of that word pair derived from the plurality of digitally-encoded natural-language texts in one or more unrelated fields.

An electronic computer in the system is operable to access the database, for retrieving or determining said selectivity value for any of the terms supplied to the database from the computer. Computer-readable code is operable, when read by the electronic computer, to perform the steps of (i) accessing the database, to retrieve or determine selectivity values for each of a plurality of terms composed of non-generic words and, optionally, word groups characterizing the target concept, invention, or event, (ii) determining for each of the plurality of natural-language texts in the same field, a match score related to the number of terms derived from that text that match those in the target concept, invention, or event, preferably weighted by selectivity values of the matching terms, and (iii) identifying from among the plurality of natural-language texts in the same field, one or more texts which have the highest match score or scores.

The electronic computer may be a central computer, where the code is operable to connect the central computer to each of a plurality of peripheral

computers on which a user can input information about the target concept, invention, or event, and can receive information about said one or more texts which have the highest match score or scores.

For use in comparing a target concept, invention, or event expressed in a natural-language input text, the code may further be operable to construct the plurality of terms by (a) identifying non-generic words in the input text, and (b) constructing from the identified words, a plurality of words groups, each containing two or more non-generic words that are proximately arranged in the input text. This code portion may be executed on one or more of the peripheral computers.

The code may be further operable to display on a peripheral computer, information about the texts having the highest match scores as a two-dimensional matrix, one dimension representing terms contained in the target concept, invention, or event, and the other dimension representing the highest matching encoded texts. Alternatively, the code may be operable to display information about the texts having the highest match scores as a list of highest matching encoded texts, and for each such text, a list of matching terms identified for that text. In another embodiment, the information may be displayed in the form of text abstracts, with the matched terms highlighted.

The above computer-readable code, and code portion forms yet another aspect of the invention. The code portion may be operable to identify terms in an input text by identifying those words and word groups in the input text having a selectivity value above a given threshold value. The code portion may be operable to generate words groups from pairs of adjacent descriptive words in the input text.

The code may be operable to associate a selectivity value with each of the terms by accessing a look-up table containing, for each of the descriptive words and, optionally, word groups in the plurality of digitally processed texts in the selected field, the selectivity value of the word, and optionally, word group.

The code may be operable to determine a match score by assigning to each of the terms in the target concept, invention, or event, a match value related to the corresponding selectivity value, and summing the match values of terms that match those of each of the plurality of digitally processed texts in the given field.

In another aspect, the invention includes a computer-accessible database

having a plurality of terms composed of non-generic words and, optionally, word groups representing proximately arranged non-generic words derived from a plurality of digitally-encoded natural-language texts in a given field. There is associated with the terms, a selectivity value determined from the frequency of occurrence of that term derived from a plurality of digitally-encoded natural-language texts in the selected field, relative to the frequency of occurrence of that term derived from a plurality of digitally-encoded natural-language texts in one or more unrelated fields. The selectivity value associated with terms below a given threshold value may be a constant or null value. Also associated with each term may be identifiers of the natural-language texts in the texts in the given field.

Where the selected field contributing to the database is a selected technical field, the one or more unrelated fields contributing to the database may be unrelated technical fields, and the plurality of texts contributing to the database may be patent abstracts or claims or technical-literature abstracts.

Where the selected field contributing to the database is a selected legal field, the one or more unrelated fields contributing to the database may be unrelated legal fields, and the plurality of texts contributing to the database may be legal-reporter case notes or head notes.

These and other objects and features of the invention will become more fully apparent when the following detailed description of the invention is read in conjunction with the accompanying drawings.

Brief Description of the Drawings

Fig. 1 illustrates components of the a system for searching texts in accordance with the invention;

Fig. 2 shows in an overview, flow diagram form, the processing of text libraries to form a target attribute dictionary;

Fig. 3 shows in an overview, flow diagram form, the steps in deconstructing a natural-language input text to generate search terms;

Fig. 4 in an overview, flow diagram, the steps in a text matching or searching operation performed by the system of the invention;

Fig. 5 is a flow diagram of Module A in the machine-readable code of the

invention, for converting target- or reference text libraries to corresponding processed-text libraries, and as indicated in Fig. 2;

Fig. 6 is a flow diagram of Module B in the machine-readable code of the invention, for determining the selectivity values of terms in the processed target-text libraries, also as indicated in Fig. 2;

Fig. 7 is a flow diagram of Module C for generating a target attribute library from the processed target-text library and selectivity value database, also as indicated in Fig. 2;

Fig. 8 is a flow diagram of Module D for matching an input text against a plurality of same-field texts, in accordance with the invention, and as indicated in Fig. 4;

Fig. 9 is a flow diagram of the algorithm in Module D for accumulating match values;

Fig. 10 is a flow diagram for construction groups of top-ranking texts having a high covering value;

Figs. 11-14 are histograms of search terms, showing the distribution of the terms among a selected number of top-ranked texts (light bars), and the distribution of the same terms among an equal number of cited US patent references (dark bars) for two different text inputs in the surgical field (Figs. 11 and 12), and in the diagnostics field (Figs. 13 and 14).

Detailed Description of the Invention

A. Definitions

Natural-language text" refers to text expressed in a syntactic form that is subject to natural-language rules, e.g., normal English-language rules of sentence construction. Examples include descriptive sentences, groups of descriptive sentences making up paragraphs, such as summaries and abstracts, and single-sentence texts, such as patent claims.

Sentence" is a structurally independent grammatical unit in a natural-language written text, typically beginning with a capital letter and ending with a period.

Target concept, invention, or event" refers to an idea, invention, or event

that is the subject matter to be searched in accordance with the invention. A target concept, invention, or concept may be expressed as a list of descriptive words and/or word groups, such as word pairs, as phrases or as natural-language text, e.g., composed of one or more sentences.

Target input text" refers to a target concept, invention, or event that is expressed in natural-language text, typically containing at least one, usually two or more complete sentences. Text summaries, abstracts and patent claims are all examples of target input texts.

Digitally-encoded text" refers to a natural-language text that is stored and accessible in digitized form, e.g., abstracts or patent claims or other text stored in a database of abstracts, full text or the like.

Abstract" refers to a summary form, typically composed of multiple sentences, of an idea, concept, invention, discovery or the like. Examples, include abstracts from patents and published patent applications, journal article abstracts, and meeting presentation abstracts, such as poster-presentation abstracts, and case notes form case-law reports.

Full text" refers to the full text of an article, patent, or case law report.

Field" refers to a field of text matching, as defined, for example, by a specified technical field, patent classification, group of classes or sub-classification, or a legal field or speciality, such "torts" or "negligence" or "property rights". An "unrelated field" is a field that is unrelated to or different from the field of the text matching, e.g., unrelated patent classes, or unrelated technical specialities, or unrelated legal fields.

Generic words" refers to words in a natural-language text that are not descriptive of, or only non-specifically descriptive of, the subject matter of the text. Examples include prepositions, conjunctions, pronouns, as well as certain nouns, verbs, adverbs, and adjectives that occur frequently in texts from many different fields. A dictionary of generic words, e.g., in a look-up table of generic words, is somewhat arbitrary, and can vary with the type of text analysis being performed, and the field of search as will be appreciated below. Typically generic words have a selectivity value (see below) less than 1to 1.25.

"Non-generic words" are those words in a text remaining after generic words

are removed. The following text, where generic words are enclosed by brackets, and non-generic words left unbracketed, will illustrate:

[A method and apparatus for] treating psoriasis [includes a] source [of] incoherent electromagnetic energy. [The] energy [is directed to a region of] tissue [to be] treated. [The] pulse duration [and the] number [of] pulses [may be] selected [to] control treatment parameters [such as the] heating [of] healthy tissue [and the] penetration depth [of the] energy [to] optimize [the] treatment. [Also, the] radiation [may be] filtered [to] control [the] radiation spectrum [and] penetration depth.

A "word string" is a sequence of non-generic words formed of non-generic words. In the example above, the first two sentence give rise to the two word strings:

treating psoriasis source incoherent electromagnetic energy.

energy directed region tissue treated.

A word group" is a group, typically a pair, of non-generic words that are proximately arranged in a natural-language text. Typically, words in a word group are non-generic words in the same sentence. More typically they are nearest or next-nearest non-generic word neighbours in a string of non-generic words. As an example, the above word string "treating psoriasis source incoherent electromagnetic energy" would generate the word pairs "treating psoriasis," treating source," "psoriasis source," "psoriasis incoherent," source incoherent," source electromagnetic," and so forth is all combination of nearest neighbors and next-nearest neighbors are considered.

Non-generic words and words groups generated therefrom are also referred to herein as "terms"

"Digitally processed text" refers to a digitally-encoded, natural-language text that has been processed to generate non-generic words and word groups.

"Database of digitally encoded texts" refers to large number, typically at least 100, and up to 1,000,000 or more such texts. The texts in the database have been preselected or are flagged or otherwise identified to relate to a specific field, e.g., the field of the desired search or an unrelated field.

"Dictionary" of terms refers to a collection of words and, optionally, word pairs, each associated with identifying information, such as the texts containing that word, or from which that word pair was derived, in the case of a "processed target-term" or "processed reference-term" dictionary, and additionally, selectivity value information for that word or word term in the case of a "target attribute dictionary." The words and word pairs in a dictionary may be arranged in some easily searchable form, such as alphabetically.

The "selectivity value" of a term (word or word group) is related to the frequency of occurrence of that term in a database of digitally processed texts in the selected field, relative to the frequency of occurrence of the same term in a database of digitally processed texts in one or more unrelated fields. The selectivity value of a given term may be calculated, for example, as the ratio of the percentage texts in a given field that contain that term, to the percentage texts in an unrelated field that contain the same term. A selectivity value so measured may be as low as 0.1 or less, or as high as 1,000 or greater.

A "descriptive term" or "descriptive search term" in a text is a term that has a selectivity value above a given threshold, e.g., 1.25 for a given non-generic word, and 1.5 for a given word pair.

A "match value" of a term is a value corresponding to some mathematical function of the selectivity value of that term, such as a fractional exponential function. For example, the match value of a given term having a selectivity value of X might be $X^{1/2}$ or $X^{1/3}$.

A "verbized" word refers a word that has a verb root. Typically, the verbized word has been converted to one form of the verb root, e.g., a truncated, present tense, active voice form of the verb. Thus, the verb root "light" could be the verbized form of light (the noun), light (the verb), lighted, lighting, lit, lights, has been lighted, etc.

Verb form" refers to the form of a verb, including present and past tense, singular and plural, present and past participle, gerund, and infinitive forms of a verb.

Verb phrase" refers to a combination of a verb with one or more auxiliary verbs including (i) to, for, (ii) shall, will, would, should, could, can, and may, might,

must, (iii) have has, had, and (iv) is are, was and were.

B. System and Method Overview

Fig 1 shows the basic elements of a text-matching system 20 in accordance with the present invention. A cental computer or processor 22 receives user input and user-processed information from a user computer 24. The user computer has a user-input-device, such as a keyboard or disc reader 28 by which the user can enter input text or text words describing an idea, concept, or event to be searched, and a display or monitor 26, for displaying search information to the user. A target attribute dictionary 30 in the system is accessible by the central computer in carrying out several of the operations of the system, as will be described. Typically, the system includes a separate target attribute dictionary for each different field of search.

In a typical system, the user computer is one of several remote access stations, each of which is operably connected to the central computer, e.g., as part of an internet system in which users communicate with the central computer through an internet connection. Alternatively, the system may be an intranet system in which one or multiple peripheral computers are linked internally to a central processor. In still another embodiment, the user computer serves as the central computer.

Where the system includes separate user computer(s) communicating with a central computer, certain operations relating to text processing are or can be carried out on a user computer, and operations related to text searching and matching are or can be carried out on the central computer, through its interaction with one or more target attribute dictionaries. This allows the user to input a target text, have the text deconstructed in a format suitable for text searching at the user terminal, and have the search itself conducted by the central computer. The central computer is, in this scheme, never exposed to the actual target text. Once a text search is completed, the results are reported to the user at the user-computer display.

Generating a target-attribute dictionary. Fig. 2 illustrates, in overview, the steps in processing target- and reference-text databases 32, 34, respectively, to form a target attribute dictionary. A target-text database is a database of digitally encoded texts, e.g., abstracts, summaries, and/or patent claims, along with pertinent identifying information, e.g., (i) pertinent patent information such as patent number, patent-office classification, inventor names, and patent filing and issues dates, (ii) pertinent journal-reference information, such as source, dates, and author, or (iii) pertinent law-reporter information, such as reporter name, dates, and appellate court. Such databases are available commercially from a variety of sources, such as the US Patent and Trademark Office, the European Patent Office PO, Dialog Search Service, legal reporter services, and other database sources whose database offerings are readily identifiable from their internet sites. The databases are typically supplied in tape or CD ROM form. In many of the examples described herein, the text databases are U.S. Patent Bibliographic databases which contain, for all patents issued between 1976 and 2000, various patent identifier information and corresponding patent abstracts. Thjs database is available in tape form from the USPTO.

The target-text database in the figure refers to a collection of digitally-encoded texts and associated identifiers in a given field of interest, e.g., a given technical or scientific or legal field. The target-text database contains the text sources and information that will be searched and identified in the search method of the invention.

The reference-text database in the figure refers to a collection of digitally encoded texts and associated information that are outside of the field of interest, i.e., unrelated to the field of the search, and serve as a "control" for determining the field-specificity of search terms in the target input, as will be described. , "Reference-text library" may refer to texts and associated information in one or more different fields that are unrelated to the field of the search. For example, the target-text database may include all post-1976 U.S. patent abstracts and identifying information in the field of surgery, and the reference-text database may include all post-1976 U.S. patent abstracts and identifying information in unrelated fields of machines, textiles, and electrical devices. As another example, the target-

text database may include all Medline (medical journal literature) abstracts and identifying information in the field of cancer chemotherapy, and the reference-field database, all Medline abstracts and identifying information in the fields or organ transplantation, plant vectors, and prosthetic devices. As still another example, the target-text database may include all head notes and case summaries and identifying information in federal legal reporters in the field of trademarks, and the reference-text database, all head notes and case summaries and identifying information in the same reporters in the fields of criminal law, property law, and water law.

With continuing reference to Fig. 2, the target-text database and reference-text database are processed by a Module A, as described below with reference to the flow diagram in Fig. 5. Briefly, Module AC in Module A operates to identify verb-root words, remove generic words, identify remaining words, and parse the text remaining after removal of generic words into word strings, typically 2-6 words long. The module uses a moving window algorithm to generate proximately arranged word pairs in each of the word strings. - Thus each text is deconstructed into a list of non-generic words and word groups, e.g., word pairs.

These words and word pairs are then arranged to form a target-term dictionary 36 or reference-term dictionary 38 containing all of the words and word pairs (terms) derivable from the texts in the associated text database, and for each term, text identifiers which identify, e.g., by patent number or other bibliographic information, each of the texts in the associated database that contains that term. The words and word pairs may be arranged in a conveniently searchable form, such as alphabetically within each class (words and word pairs) of terms. Thus, an entry in a processed target-term or reference-term dictionary generated from a patent-abstract database would include a non-generic word or word-pair term, and a list of all patent or publication numbers that contain that term.

The target-term and reference-term dictionaries just described, and identified at 36, 38, respectively in Fig. 2, are used to generate, for each term in the target-term dictionary, a target-term selectivity value that indicates the relative occurrence of that term in the processed target- and reference-term dictionaries. The determination of a selectivity value for each term in the target-term dictionary

is carried out by Module B, described below with reference to Fig. 6. In one general embodiment, the selectivity value is determined simply as the normalized ratio of the number of texts in the target-term dictionary containing that term to the number of texts in the reference-term dictionary that contain the same term.

Thus for example, assume that the term "electromagnetic" in the target-term dictionary contains 1,500 different text identifiers (meaning that this term occurs at least once in 1,500 different abstracts in the target-text database), and contains 500 different text identifiers in the reference-term dictionary. The ratio of occurrence of the word in the two dictionaries in thus 3:1, or 3. To determine the normalized ratio, this number is then multiplied by the ratio of total number of texts in the reference and target databases. In the example above, assuming that the target text database contains 50,000 texts, and the reference text database, 100,000, the selectivity ratio of 3 is multiplied by 100/50 or 2 to yield a normalized selectivity value for the term "electromagnetic" of 6.

In another embodiment, the processed target-term and reference-term dictionaries are generated by considering only some arbitrary number, e.g., 50,000, of the texts in the respective target-text and reference-text databases, for generating the target-term and processed-term dictionaries, so that the normalization factor for is always 1. It will be appreciated that by selecting a sufficiently large number of texts from each database, a statistically meaningful rate of occurrence of any term from the database texts is obtained.

To produce a target-attribute dictionary (shown at 30 in Figs. 1 and 2), the selectivity values just described, and stored at 40, are assigned to each of the associated terms in the target-term dictionary. For example, in the case above, the selectivity value of "6" is assigned to the term "electromagnetic" in the target-term dictionary. The target-attribute dictionary now contains, for each dictionary term, a list of text identifiers for all texts in the target-text database that contain that term and the selectivity value of the term calculated with respect to a reference text database of texts in an unrelated field.

This target-attribute dictionary forms one aspect of the invention. More generally, the invention provides a dictionary of terms (words and/or word groups, e.g., word pairs) contained in or generated from the texts in a database of natural-

language texts in a given field. Each entry in the dictionary is associated with a selectivity value which is related to the ratio of the number of texts in the given field that contain that term or from which that term can be generated (in the case of word groups) to the number of texts in a database of texts in one or more unrelated fields that contain the same term, normalized to the total number of texts in the two databases. The selectivity value associated with terms below a given threshold value may be a constant or null value.

As just described, the dictionary may additional include, for each term, text identifiers that identify all of the texts in the target text database that contain that term or from which the term can be derived.

Processing a target text into search terms. The concept, invention of event to be searched may be expressed as a group of words and, optionally, word pairs that are entered by the user at the user terminal. Such input, as will be seen, is already partly processed for searching in accordance with the invention.

In a more general embodiment, and with reference to Figure 3, the user inputs a natural-language text 42 that describes the concept, invention, or event as a summary, abstract or precis, typically containing multiple sentences, or as a patent claim, where the text may be in a single-sentence format. An exemplary input would be a text corresponding to the abstract or independent claim in a patent or patent application or the abstract in a journal article, or a description of events or conditions, as may appear in case notes in a legal reporter.

The input text, which is preferably entered by the user on user computer 24, is then processed on the user computer (or alternatively, on the central computer) to generate non-generic words contained in the text and word groups constructed from proximately arranged non-generic words in the text, as indicated at 44. The deconstruction of text into non-generic words and word groups, e.g., word pairs is carried out by Module C in the applicable computer, including Module AC which is also used in processing database texts, as noted above. The operation of Module C is described more fully below with reference to Fig. 7.

With continuing reference to Fig. 3, non-generic words and word pairs (collectively, terms) from the input text (or from terms that are input by the user in lieu of an input text, as above) are then supplied to the central computer which

performs the following functions: (i) For each term contained in or generated from the input text, the central computer "looks up" the corresponding selectivity value in target-attribute dictionary 30. (ii) Applying default or user-supplied selectivity-value threshold values, the central computer saves terms having above-threshold selectivity values. For example the default or suer-supplied selectivity values for the words and word pairs may be 1.25 and 1.5, respectively. The central computer would then save, as "descriptive" terms, only those input text words having a selectivity value of 1.25 or greater, and only those word pairs having a selectivity value of 1.5 or greater. The above operations are carried out by Module C in the system, detailed below with respect to Fig. 7. As indicated in Fig. 3, Module C may be executed partly on a user computer (processing of input text) and partly on a central computer (identifying descriptive terms).

The descriptive terms (words or word pairs), saved at 48, are then displayed to the user at the user display. The user may have several options at this point. The user may accept the terms as pertinent and appropriate for the search to be conducted, without further editing; the user may add synonyms to one or more of the words (including words in the word pairs) to expand the range of the search; the user may add or delete certain terms; and/or specify a lower or higher selectivity-value threshold for the terms, and ask the central computer to generate a new list of descriptive terms, based on the new threshold values.

These changes (if any) to the originally generated input-text descriptive terms are returned to the user and/or stored by the central computer for text matching, to be described below with reference to Figs. 4 and Figs. 8 and 9.

The invention thus provides, in another aspect, computer-readable code which is operable, when read by an electronic computer, to generate descriptive terms (words and/or multi-word groups) from a digitally encoded, natural-language input text that describes a concept, invention, or event in a selected field. The code operates to (a) identify non-generic words in the input text, (b) construct from the identified non-generic words, a plurality of word groups, each containing two or more words that are proximately arranged in the input text (where word groups are to be generated), (c) select terms from (b) as descriptive terms if, for each group considered, the group has a selectivity value above a selected threshold value,

and (d) storing or displaying the terms selected in (c). Also disclosed is a system using this code for generating descriptive terms from a digitally encoded, natural-language input text that describes a concept, invention, or event in a selected field, and an automated method carried out by the system.

In another aspect, the invention provides computer readable code, and a system and method for identifying descriptive words and/or word groups in a natural-language text. The method includes identifying from the text, those terms, including non-generic words contained in the text and/or word groups generated from the non-generic words, that have a selectivity value greater than a given or specified threshold.

Conducting a text-matching search. This section will provide an overview of the text-matching or text-searching operation in the invention. The purpose of the operation is to identify those texts originally present in a target-text database that most closely match the target input text (or terms) in content, that is, the most pertinent references in terms of content and meaning. The search is based on two strategies which have found to be useful, in accordance with the invention, in extracting content from natural-language texts: (i) by using selectivity values to identify those terms, i.e., words and optionally, word groups, having above-threshold selectivity values, and (ii) considering all of the high selectivity value terms collectively, the search compares target and search texts in a content-rich manner.

Preferably, the final match score will also reflect the relative "content" value of the different search terms, as measured by some function related to the selectivity value of that term. This function is referred to as a match value. For example, if a term has a selectivity value of 8, and the function is a cube root function ($SV^{1/3}$), the match value will be 2. A cube root function would compress the match values of terms having selectivity values to between 1 and 1,000 to 1 to 10; a square root function would compress the same range to between 1 and about 33.

Fig. 4 shows the overall flow of the components and operations in the text-matching method. The initial input in the method is the descriptive search terms generated from the input text as above. For each term (word and optionally, word

group), the code in the central computer (Module D described below with reference to Figs. 8 and 9) operates to look up that term in the target-attribute dictionary. If the selectivity value of the term is at or above a given threshold, the code operates to record the text identifiers of all of the texts that contain the term (word) or from which that term (word group) is generated. The text identifiers are placed form an accumulating or update list of text IDs, each associated with one or more terms, and therefore, with one or more match values those terms.

The steps are repeated until each term has been so processed. With each new term, the text IDs and match value associated with that term are added to update list of Ids, either as new ID's or as additional match values added to existing Ids, as described below with reference to Fig. 9. After all of the terms have been considered, the update list includes each text ID having at least one of the search terms, and the total match score for that text. The program then applies a standard ranking algorithm to rank the text entries in the updated in buffer 50, yielding some selected number, e.g., 25, 50, or 100 of the top ranked matching texts, stored at 52.

As will be described below with reference to Fig 9, the system may also find, from among the texts with the top match scores, e.g., top 100 texts, a group of texts, e.g., group of 2 or 3 texts, having the highest collective number of hits. This group will be said to be a spanning or covering group, in that the texts in the group maximally span or cover the terms from the target input.

Information relating to the top-ranked texts, and/or covering groups may be displayed to the user, e.g., in the form discussed below with respect to Fig. 14. This display is indicated at 26 in Fig. 4. In one embodiment of the invention, for use in a user-fee system, the information displayed at 26 may include information about text scores, and/or matching terms, and the text itself, but not include specific identifying information, such as patent numbers of bibliographic citations.

In this embodiment, the user selects, as at 54, those texts which are of greatest interest, based, for example, on the match score, the matching terms, and/or the displayed text of a given reference. This input is fed to the central computer, which then retrieves the identifying information for the texts selected by the user, as at 56, and supplies this to the user at display 26. This allows for a

variety of user-fee arrangements, e.g., one in which the user pays a relatively low rate to examine the quality of a search, but a higher rate to retrieve specific texts of interest.

C. Text processing: Module AC

There are two related text-processing operations in the invention. The first is used to deconstruct each text in the target-text or the reference-text database into a list of words and word groups, e.g., word pairs, that are contained in or derivable from that text. The second is the deconstruction of a target text into meaningful search terms, e.g., words and word pairs. Both text-processing operations involve a Module AC which processes a text into terms, that is, non-generic words and optionally word groups formed proximately arranged non-generic words. Module AC is described in this section with reference to Fig. 5, and illustrated with respect to a model patent claim.

The first step in text processing module of the program is to "read" the text for punctuation and other syntactic clues that can be used to parse the text into smaller units, e.g., single sentences, phrases, and/or subphrases, and more generally, word strings. These steps are represented by parsing function 60 in Module AC. The design of and steps for the parsing function will be appreciated form the following description of its operation.

For example, if the text is a multi-sentence paragraph, the parsing function will first look for sentence periods. A sentence period should be followed by at least one space, followed by a word that begins with a capital letter, indicating the beginning of a the next sentence, or should end the text, if the final sentence in the text. Periods used in abbreviations can be distinguished either from an internal dictionary of common abbreviations and/or by a lack of a capital letter in the word following the abbreviation.

Where the text is a patent claim, the preamble of the claim can be separated from the claim elements by a transition word "comprising" or "consisting" or variants thereof. Individual elements may be distinguished by semi-colons and/or new paragraph markers, and/or element numbers of letters, e.g., 1, 2, 3, or i, ii, iii, or a, b, c.

As will be appreciated below, the parsing algorithm need not be too strict, or particularly complicated, since the purpose is simply to parse a long string of words (the original text) into a series of shorter ones that encompass logical word groups. In addition to punctuation clues, the parsing algorithm may also use word clues. For example, by parsing at prepositions other than "of", or at transition words, useful word-strings can be generated.

After the initial parsing, the program carries out word classification functions, indicated at 62, which act to classify the words in the text into one of three basic groups: (i) a group of verbs and verb-root words, (ii) generic words, and (iii) remaining groups, i.e., words other than those in groups (i) or (ii), which tend to be nouns and adjectives. The verb word group includes all words that act as verbs or have a verb root, including words that may (and often are) nouns or adjectives, such as "light," "monitor," "discussion," "interference," and so on. These words are identified readily by comparing each word with one of the forms of the verb root compiled in verb dictionary, such as given in Attachment A. The purpose of verbizing all verb-root words is twofold. First, the resulting search term containing the verb root will be more robust, since it will encompass a variety of ways that the verb-root word may be used in natural-language text, i.e., either as a verb, adverb, noun, or adjective. Second, to the extent meaningful synonyms can be attached to verbs, the program will be able to automatically generate word synonyms.

Once verb and verb-root words have been identified, it may be useful, although not necessary, to identify actual verbs. In a patent claim, for example, the verb words "including" or "include(s)" and "comprising" or "comprise(s)", and "consisting of" and "consists of" almost invariably indicate a true verb. In technical texts, which tend to be rich in passive voice verbs, the presence of compound verbs containing "to be" or "to have" forms indicates that associated verb word is a true verb. The purpose of identifying true verbs is to provide another "marker" for sentence parsing.

The group of generic words include all articles, prepositions, conjunctions, and pronouns as well as many noun or verb words that are so generic within the texts of a database as to have little or no meaning in terms of describing a

particular invention, idea, or event. For example, in the patent or engineering field, the words "device," "method," "apparatus," "member," "system," "means," "identify," "correspond," or "produce" would be considered generic. As will be appreciated below, such words could also be retained at this stage, and eliminated at a later stage, on the basis of a low selectivity factor value. However, for convenience, the program includes a dictionary of generic words that are removed from the text at this initial stage in text processing.

The words remaining after tagging all verb words and generic words are object or noun words that typically function as nouns or adjectives in the text.

As an example of these text processing operations, consider the processing of the following claim, which has been parsed by paragraph markers, with verb-root words indicating by italics, true verbs by bold italics, generic words, by normal font, and remainder "noun" words, by bold type.

A device for *monitoring* **heart rhythms**, ***comprising***:

means for *storing digitized* **electrogram segments including** *signals* indicative of *depolarizations* of a **chamber** or **chamber** of a **patient's heart**;

means for *transforming* the *digitized signals* into *signal* **wavelet coefficients**;

means for identifying higher **amplitude** ones of the *signal* **wavelet coefficients**; and

means for generating a *match* **metric** corresponding to the higher **amplitude** ones of the *signal* **wavelet coefficients** and a corresponding set of **template wavelet coefficients** derived from **signals** indicative of a **heart** *depolarization* of known type, and

identifying the **heart rhythms** in response to the *match* **metric**.

The text may be further parsed at all prepositions other than "of". When this is done, and generic words are removed, the program generates the following strings of non-generic verb and noun words.

monitoring **heart rhythms**//*storing digitized* **electrogram segments**//*signals depolarizations* **chamber patient's heart**// *transforming*

digitized signals//signal **wavelet coefficients// amplitude** *signal* **wavelet coefficients//** *match* **metric// amplitude** *signal* **wavelet coefficients// template wavelet coefficients//**

signals heart *depolarization//* **heart rhythms//***match* **metric.**

The operation for generating words strings of non-generic words is indicated at 64 in Fig. 5, and generally includes the above steps of removing generic words, and parsing the remaining text at natural punctuation or other syntactic cues, and/or at certain transition words, such as prepositions other than "of."

The desired word groups, e.g., word pairs, are now generated from the words strings. This may be done, for example, by constructing every permutation of two words contained in each string. One suitable approach that limits the total number of pairs generated is a moving window algorithm, applied separately to each word string, and indicated at 66 in the figure. The overall rules governing the algorithm, for a moving "three-word' window, are as follows:

1. consider the first word(s) in a string. If the string contains only one word, no pair is generated;

2. if the string contains only two words, a single two-word pair is formed;

3. If the string contains only three words, form the three permutations of word pairs, i.e., first and second word, first and third word, and second and third word;

4. If the string contains more than three words, treat the first three words as a three-word string to generate three two-words pairs; then move the window to the right by one word, and treat the three words now in the window (words 2-4 in the string) as the next three-word string, generating two additional word pairs (the word pair formed by the second and third words in preceding group will be the same as the first two words in the present group) string;

5. continue to move the window along the string, one word at a time, until the end of the word string is reached.

For example, when this algorithm is applied to the word string : *store digitize* **electrogram segment**, it generates the word pairs: store-digitize, store-electrogram, digitize-electrogram, digitize-segment, electrogram-segment, where

the verb-root words are expressed in their singular, present-tense form and all nouns are in the singular.

D. Processing text databases to form a target-attribute dictionary

This section will describe the processing of the texts in a text database—either target text database 32 or reference-text database 34-- to form the corresponding target-term dictionary 36 or reference-term dictionary 38, (Module A), the use of the two dictionaries to calculate target-term selectivity values (Module B), and finally, the construction of target-attribute dictionary 30.

Processing the text databases. The operation of Module A will be described with respect to the construction of target-term dictionary 36 from target-text database 32, it being recognized that the same operations are carried out for constructing reference-term dictionary 40 from reference-text database 34.

As noted above, each database contains a large number of natural-language texts, such as abstracts, summaries, full text, claims, or head notes, along with reference identifying information for each text, e.g., patent number, literature reference, or legal reporter volume. For purposes of illustration, the two database are US patent bibliographic databases containing information about all US patents, including an abstract patent, issued between 1976 and 2000, in a particular field, in this case, several classes in covering surgical devices and procedures. That each,, each entry in the database includes bibliographic information for one patent and the associated abstract.

With reference to Fig. 5, program represented by Module A initially retrieves a first entry, as at 68, where the entry number is indicated at t. The abstract from this entry is then processed by Module AC according to the steps detailed above to generate for that abstract, a list of non-generic words and word pairs constructed from proximately arranged non-generic words in the parsed word strings of the text.

Each term. i.e., word and word pair, is then added to an updated dictionary 70, typically in alphabetical order, and with the words and word-groups either mixed together or treated as separate sections of the dictionary. For each

word and word pair added to the dictionary, the program operates to add the text ID, i.e., patent number, of the text containing that word or from which the word pair is derived. This operation is indicated at 72. If the words being entered are from the first text, the program merely enters each new term and adds the associated text ID. For all subsequent text, t>1, the program carries out the following operations.

1. for each term in the new text, scan the updated dictionary for that term;

2. If the term is found, add the associated text Id to the already existing text Id(s) for that term;

3. if the term is not found in dictionary 72, add it, along with the associated text Id.

4. repeat steps 1-3 until all of the terms in new text t have been added to the dictionary, either as new terms or additional Id's to existing terms.

This completes the processing of text t. The program then increments t, pulls up the next text from the database and carries out the same operations to add the terms from the new text to the updated dictionary. The text processing is repeated until all of the texts t in the database (either the target- or reference-text database) have been processed, as indicated by the logic at 74. At this point, update dictionary 74 becomes dictionary 32 (or 34), and includes (i) all of the non-generic words and word groups contained in and derived from the texts in the corresponding database, and (ii) for each term, all of the text Ids containing that term.

Calculating selectivity values. Module B shown in Fig. 6 shows how the two dictionaries whose construction was just described are used to generate selectivity values for each of the terms in the target-term dictionary. In this flowchart, "w" indicates a term, either a word or word pair in the corresponding dictionary.

The program is initialized at w=1, meaning the first term in the target-term dictionary. The program calls up this term in both dictionaries, as at 76. The program then determines at 78 the occurrence O_t of that term in the target-term dictionary as the total number of Ids (different texts) associated with that term/divided by the total number of texts processed from the target-database to

produce the dictionary. For example, if the word "heart" has 500 associated text Ids, and the total number of texts used in generating the target-term dictionary is 50,000, the occurrence of the term is 500/50,000, or 1/100. The program similarly calculates at 80 the occurrence of O_r of the same term in the reference-term dictionary. For example, if the word "heart" appears in 200 texts out of a total of 100,000 used to construct the reference-term dictionary, its occurrence O_r is 200/100,000 or 1/500. The selectivity value of the term is then calculated at 82 as O_t/O_r, or, in the example given, 1/100 divided by 1/500 or 5. As noted above, the Ot and Or values may be equivalently calculated using total text numbers, then normalizing the ratio by the the total number of target and reference texts used in constructing the respective dictionaries.

The selectivity value thus calculated is then added to that term in the target-term dictionary, as at 88. Although not indicated, the selectivity value may be assigned an arbitrary or zero value in either of the two following cases:

1. the occurrence of the term in either the target- or reference-term dictionary is below a selected threshold, e.g., 1, 2 or 3;

2. the calculated selectivity value is below a given threshold, e.g., 0.5 or 0.75.

In the first case, if O_t is significant and O_r is zero or very small, the selectivity value may be assigned a large number such as 100. If O_t is zero or very small, the term may be assigned a zero or very small value. In addition, a term with a low selectivity value, e.g., below 0.5, may be dropped from the dictionary, since it may have not real search value.

These operations are repeated until all of the terms in the target-term dictionary have been considered, as indicated by the logic at 90. At this point, all the target-term dictionary contains all of the pertinent selectivity values, and is thus referred to as the target-attribute dictionary.

Entries in the dictionary might have the following form, where a term is followed by its selectivity value, and a list of all associated text IDs.

heart (622.6906);
US 20020077687 20020620 US 2001814533 20010321;
US 20020077571 20020620 US 200120771 20011212;
US 20020077566 20020620 US 1998170793 19981013;
US 20020077563 20020620 US 2000740258 20001218;

US 20020077562 20020620 US 2000738869 20001215;
US 20020077555 20020620 US 2001877615 20010608;
US 20020077554 20020620 US 2000739062 20001218;
US 20020077538 20020620 US 2000681068 20001219;
US 20020077534 20020620 US 200128902 20011218;
US 20020077532 20020620 US 2001993175 20011106;

US 3095872 19630702

heart-monitor (238.5)
US 20020077538 20020620 US 2000681068 20001219;
US 20020068874 20020606 US 2001784413 20010213;
US 20020065449 20020530 US 200258475 20020128;
US 20020052558 20020502 US 2001907895 20010712;
US 20020052557 20020502 US 2001793653 20010227;
US 20020045837 20020418 US 2001910837 20010724;
US 20020038094 20020328 US 2001988605 20011120;
US 20020019585 20020214 US 2001875162 20010607;
US 20020013535 20020131 US 2001861904 20010521;
US 20020007117 20020117 US 2001835166 20010416;

US 3135264 19640602

Each text identifier may include additional identifying information, such as the page and lines numbers in the text containing that term, or passages from the text containing that term, so that the program can output, in addition to text identification for highest matching texts, pertinent portions of that text containing the descriptive search term at issue.

E. Automated Search Method and System

As noted above, the automated search method of the invention involves first, extracting meaningful search terms or descriptors from an input text (when the input is a natural-language text), and second, using the extracted search terms to identify natural-language texts describing ideas, concepts, or events that are pertinent to the input text. This section considers the two operations separately.

Extracting meaning search terms. Fig. 7 is a flow diagram of the operations carried out by Module C of the program for extracting meaning search terms or descriptors. The input text, indicated at 92, is a natural-language text, e.g., e.g., an abstract, summary, patent claim head note or short expository text

describing an invention, idea, or event. This text is processed by Module AC as described with reference to Fig. 5 above, to produce a list of non-generic words and word pairs formed by proximately arranged word pairs. These terms are referred to here as "unfiltered terms" and are stored at 94, e.g., in a buffer.

An unfiltered term will be deemed a meaningful search term is its selectivity value is above a given, and preferably preselected selectivity value, e.g., a value of at least 1.25 for individual word terms, and a value of at least 1.5 for word-pair terms. This is done, in accordance with the flow diagram shown in Fig. 7, by calling up each successive selectivity value from 94, finding its selectivity value from dictionary 30, and saving the term as a meaningful search term if its selectivity value is above a given threshold, as indicated by the logic at 94. When all of the terms have been so processed, the program generates a list 96 of filtered search terms (descriptive terms). The program may also calculate corresponding match values for each of the terms in 96. As indicated above, the match value is determined from some function of the selectivity value, e.g., the cube root or square root of the selectivity value, and serves as a weighting factor for the text-match scoring, as will be seen in the next section.

Text matching and scoring. The next step in program operation is to employ the descriptive search terms generated as above to identify texts from the target-text database that contain terms that most closely match the descriptive search terms. The text-matching search operation may be carried out in various ways. In one method, each the program actually processes the individual texts in the target database (or other searchable database), to deconstruct the texts indo word and word-pair terms, as above, and then carries out a term-by-term matching operation with each text, looking for and recording a match between descriptive term and terms in each text. The total number of term matches for a text, and the total match score, is then recorded for each text. After processing all of the texts in the database in this way, the program identifies those texts having the highest match scores. This search procedure is relatively slow, since each text must be individually processed before a match score can be calculated.

A preferred and much faster search method uses the target-attribute

dictionary for the search operation, as will now be discussed with respect to Figs. 8 and 9. In this method, the program is initialized at w=1, and gets term w from the list of descriptive terms at 96. The program then accesses the target-attribute dictionary to record the text Ids and the corresponding match value for that term, as indicated at 98. This operation, it will be appreciated, is effective to record all text Ids in the target database containing the term being considered.

The next step is to assign and accumulate match values for all of Ids being identified, on an ongoing basis, that is, with each new group of text Ids associated with each new descriptive term. This operation is indicated at 100 in Fig. 8 and given in more detail in Fig. 9. The updated list of all Ids, indicated at 102, contains a list of all text Ids which have been identified at any particular point during the search. Initially the list includes all of those text Ids associated with the first descriptive term. When a new term, e.g., the second term is considered. as at 104, the program operates to compare each new text Id with each existing text Id in the updated list, as indicated at 106. If the text Id being examined already exists in the updated list, the match value for the new term is assigned to that text Id, which now includes at least two match values. This operation, and the decision logic, is indicated at 110, 112 in the figures. If the text Id being examined is not already in the updated list, it is added to the list, as at 108, along with the corresponding match score. This process is repeated until all of the text Ids for a new term have been so processed, either by being added to the updated list or being added to an already existing text Id. The program is now ready to proceed to the next descriptive term, according to the program flow in Fig. 8.

Once all of the descriptive terms have been processed, as at 116, the updated Id list includes, for each text Id found, one or more match values which indicate the number of matched terms associated with that text ID, and the match value for each of the terms. The total match score for each text Id is then readily determined, e.g., by adding the match values associated with each text Id. The scores for the text Ids are then ranked, as at 118 and a selected number of highest-matching scores, e.g., top 50 or 100 scores, are output at 120, e.g., at the user computer display.

Fig. 8 also shows at 122 a logic decision for another feature another of the invention—the ability to group top-ranked hits into groups of "covering" or "spanning" references containing an optimal number of terms, as will now be described with reference to Fig. 10.

Finding a spanning group of references. The purpose of this function is to identify, out of a group of X, e.g., X=50-100, top-matching texts, some relatively small number Y, e.g., Y=2-4, of these top ranked texts that can the highest total number of matched terms.

With reference to Fig. 10, the program operation here starts with some number X of top-ranked texts for a specific target input, as at 124. For each of these texts, , the program constructs an N-dimensional vector, where N is the number of target descriptive terms, as indicated at 126. Each of the N vector coefficients in the N-dimensional vector is either a 1, indicating a given term is present in the text, or a 0, indicating the term is absent from that text.

The program generates all permutations of X texts, without regard to order, as indicated at 130. Thus, for example, if there X=100 top ranked texts, and Y = 3, the program would generate 100!/97!3! different groups of three texts, or triplets.

For each group, e.g., triplet, the program adds the vectors in the group, using a logical OR operation, as at 132. This operation creates a new vector whose coefficients are 1 if the corresponding term is present in any of the Y vectors being added, and is 0 only if the corresponding term is absent from all Y vectors. The 1-value coefficients may now be replaced by the match values of the corresponding terms, as at 134.

The "covering" score for each permutation groups is now calculated as the sum of the vector coefficients, as at 136. The scores are then sorted, and the highest ranking group of text identified by the highest vector-sum value, as at 136 and 138. As indicated at 140, the combined vector with the highest scores give the best combinations of terms matching the target terms.

The following examples illustrate four searches carried out in accordance with the invention. Each example gives the target text (abstract and claim that served as the input text for the search), the words and word pairs generated from

the input text, and the corresponding selectivity values, and the patents numbers of the top-ranked abstracts found in the search. In each case, the target input is from an issued US patents having predominantly issued US patents as cited references. For each of these patents, the occurrences of descriptive terms in the cited references are compared with those in the same number of top-ranked patents found in the search. The histograms for Examples 1-4 are given in Examples 11-14.

Example 1:

Method and apparatus for detection and treatment of cardiac arrhythmias

Abstract:

A device for monitoring heart rhythms. The device is provided with an amplifier for receiving electrogram signals, a memory for storing digitized electrogram segments including signals indicative of depolarizations of a chamber or chamber of a patient's heart and a microprocessor and associated software for transforming analyzing the digitized signals. The digitized signals are analyzed by first transforming the signals into signal wavelet coefficients using a wavelet transform. The higher amplitude ones of the signal wavelet coefficients are identified and the higher amplitude ones of the signal wavelet coefficients are compared with a corresponding set of template wavelet coefficients derived from signals indicative of a heart depolarization of known type. The digitized signals may be transformed using a Haar wavelet transform to obtain the signal wavelet coefficients, and the transformed signals may be filtered by deleting lower amplitude ones of the signal wavelet coefficients. The transformed signals may be compared by ordering the signal and template wavelet coefficients by absolute amplitude and comparing the orders of the signal and template wavelet coefficients. Alternatively, the transformed signals may be compared by calculating distances between the signal and wavelet coefficients. In preferred embodiments the Haar transform may be a simplified transform which also emphasizes the signal contribution of the wider wavelet coefficients.

Claim:

1. A device for monitoring heart rhythms, comprising:

means for storing digitized electrogram segments including signals indicative of depolarizations of a chamber or chamber of a patient's heart;

means for transforming the digitized signals into signal wavelet coefficients;

means for identifying higher amplitude ones of the signal wavelet coefficients; and

means for generating a match metric corresponding to the higher

amplitude ones of the signal wavelet coefficients and a corresponding set of template wavelet coefficients derived from signals indicative of a heart depolarization of known type, and identifying the heart rhythms in response to the match metric. (Main Claim)

Target words/pairs with SF:

monitor (2978,3393,2.63307)
hear (2322,307,22.6906)
rhythms (94,10,28.2)
amplify (665,2257,0.883917)
electrogram (50,1,150)
memory (708,6367,1.5)
store (1545,7680,1.5)
digitized (156,174,2.68966)
segment (970,2202,1.5)
signal (5733,17424,1.5)
depolarize (79,7,33.8571)
chamber (3091,8860,1.5)
patient's (4133,59,210.153)
microprocessor (391,752,1.55984)
software (109,906,1.5)
transform (370,1371,1.5)
wavelet (16,17,2.82353)
coefficients (276,1016,1.5)
higher (469,2566,1.5)
amplitude (778,1089,2.14325)
template (104,419,1.5)
derive (760,1563,1.45873)
Haar (1,1,3)
filter (1491,2726,1.64087)
delete (32,264,0.363636)
order (27,315,0.257143)
compare (1324,4042,1.5)
calculate (1016,2312,1.5)
distances (1241,5070,0.73432)
contribution (42,51,2.47059)
wider (93,333,0.837838)
match (499,1718,0.871362)
metric (7,125,0.168)

hear---monitor(159,2,238.5)
monitor---rhythms(2,1,6)
hear---rhythms(17,1,51)
amplify---electrogram(0,1,0)
amplify---memory(2,33,0.181818)
electrogram---memory(1,1,3)
digitized---store(15,25,1.8)

electrogram---store(4,1,12)
digitized---electrogram(1,1,3)
digitized---segment(0,3,0)
electrogram---segment(1,1,3)
electrogram---signal(17,1,51)
segment---signal(17,52,0.980769)
depolarize---segment(0,1,0)
depolarize---signal(11,1,33)
chamber---signal(5,20,0.75)
chamber---depolarize(0,1,0)
chamber---patient's(14,1,42)
chamber---hear(95,6,47.5)
hear---patient's(230,2,345)
microprocessor---patient's(0,1,0)
hear---microprocessor(4,1,12)
hear---software(0,1,0)
microprocessor---software(4,7,1.71429)
digitized---transform(1,1,3)
signal---transform(49,73,2.0137)
digitized---signal(43,57,2.26316)
signal---wavelet(4,4,3)
coefficients---signal(8,19,1.26316)
coefficients---wavelet(3,4,2.25)
coefficients---transform(4,23,0.521739)
transform---wavelet(5,8,1.875)
amplitude---higher(5,8,1.875)
higher---signal(14,65,0.646154)
amplitude---signal(155,264,1.76136)
amplitude---wavelet(0,1,0)
higher---wavelet(0,1,0)
coefficients---higher(3,9,1)
amplitude---coefficients(0,3,0)
template---wavelet(0,1,0)
coefficients---template(0,1,0)
derive---wavelet(0,1,0)
coefficients---derive(3,7,1.28571)
hear---signal(172,16,32.25)
depolarize---hear(16,1,48)
Haar---signal(0,1,0)
Haar---transform(0,1,0)
Haar---wavelet(0,1,0)
filter---transform(2,14,0.428571)
filter---signal(153,265,1.73208)
amplitude---delete(0,1,0)
delete---signal(0,1,0)
order---signal(0,1,0)
order---template(0,2,0)
signal---template(5,1,15)

amplitude---compare(16,25,1.92)
amplitude---order(0,1,0)
compare---order(0,1,0)
compare---signal(87,450,0.58)
calculate---distances(7,77,0.272727)
calculate---signal(44,82,1.60976)
distances---signal(17,54,0.944444)
distances---wavelet(0,1,0)
contribution---signal(6,3,6)
signal---wider(0,3,0)
contribution---wider(0,1,0)
contribution---wavelet(0,1,0)
wavelet---wider(0,1,0)
coefficients---wider(0,1,0)
match---metric(0,1,0)
depolarize---rhythms(0,1,0)

Top 50 hits with scores

abs 060583274(67.3522) abs 054921287(67.0247) abs 053953932(67.0247) abs 043643973(58.3504) abs 043677533(56.4689) abs 051333503(56.2582) abs 052923487(55.8248) abs 059719338(55.5909) abs 058823123(54.3016) abs 059546611(53.3994) abs 059351608(53.3994) abs 048650366(52.8353) abs 056626894(52.2547) abs 058171312(52.2547) abs 060630779(52.2547) abs 054337305(52.2547) abs 058938818(52.2547) abs 058632913(52.2547) abs 055424309(51.6237) abs 041778006(50.9472) abs 054151716(50.9472) abs 050835653(50.6756) abs 039616231(49.8025) abs 052679420(49.8025) abs 049586416(49.8025) abs 049778994(49.8025) abs 058171320(49.8025) abs 055541771(49.8025) abs 057525218(49.8025) abs 050147013(48.9413) abs 057557381(47.2613) abs 053185935(45.3411) abs 049586327(45.3411) abs 050923307(45.3411) abs 061578592(44.6995) abs 055077846(43.9862) abs 054115310(43.9862) abs 046257306(43.9602) abs 056073852(43.9549) abs 053342208(43.9549) abs 059958715(43.5548) abs 056200021(43.1821) abs 039788482(43.1022) abs 057662274(42.945) abs 055031587(42.8516) abs 042365244(42.8004) abs 049286900(42.7258) abs 050781340(42.7258) abs 039601412(41.8919) abs 049827383(41.6557)

Example 2

Method, apparatus and system for removing motion artifacts from measurements of bodily parameters

Abstract:

A method for removing motion artifacts from devices for sensing bodilyparameters and apparatus and system for effecting same. The method includes analyzing segments of measured data representing bodily parameters and possibly noise from motion artifacts. Each segment of measured data may correspond to a single light signal transmitted and detected after transmission or

reflection through bodily tissue. Each data segment is frequency analyzed to determine dominant frequency components. The frequency component which represents at least one bodily parameter of interest is selected for further processing. The segment of data is subdivided into subsegments, each subsegment representing one heartbeat. The subsegments are used to calculate a modified average pulse as a candidate output pulse. The candidate output pulse is analyzed to determine whether it is a valid bodily parameter and, if yes, it is output for use in calculating the at least one bodily parameter of interest without any substantial noise degradation. The above method may be applied to red and infrared pulse oximetry signals prior to calculating pulsatile blood oxygen concentration. Apparatus and systems disclosed incorporate methods disclosed according to the invention.

Claim:

1. A method of removing motion-induced noise artifacts from a single electrical signal representative of a pulse oximetry light signal, comprising:

receiving a segment of raw data spanning a plurality of heartbeats from said single electrical signal;

analyzing said segment of raw data for candidate frequencies, one of which said candidate frequencies may be representative of a valid plethysmographic pulse;

analyzing each of said candidate frequencies to determine a best frequency including narrow bandpass filtering said segment of raw data at each of said candidate frequencies;

outputting an average pulse signal computed from said segment of raw data and said best frequency; and

repeating the above steps with a new segment of raw data. (Main Claim)

Target words/pairs with SF:

remove (5711,12291,1.5)
motion (1181,2651,1.5)
artifacts (263,57,13.8421)
sense (4497,8708,1.54927)
bodily (10148,8611,3.53548)
parameters (997,1659,1.80289)
segment (970,2202,1.5)
measure (4930,8592,1.72137)
noise (477,1280,1.5)
light (2161,5201,1.5)
signal (5733,17424,1.5)
transmit (3147,9513,1.5)
detect (3975,9067,1.31521)
reflect (1146,3388,1.5)
tissue (4351,131,99.6412)
frequency (2126,4687,1.5)
dominant (21,46,1.5)
subdivided (27,155,1.5)

subsegment (2,1,6)
hear (2322,307,22.6906)
calculate (1016,2312,1.5)
modify (588,2379,1.5)
average (559,1100,1.52455)
pulse (2661,2370,3.36835)
candidate (21,137,1.5)
output (2491,9859,1.5)
valid (40,302,1.5)
degradation (42,307,0.410423)
red (134,250,1.608)
infrared (280,567,1.5)
oximetry (69,1,207)
pulsatile (116,1,348)
blood (4415,162,81.7593)
oxygen (1031,732,4.22541)
concentrate (883,2016,1.5)
motion-induced (3,1,9)
electrical (3404,10634,1.5)
raw (50,470,1.5)
spanning (52,126,1.5)
plethysmographic (15,1,45)
best (92,236,1.5)
narrow (428,1051,1.5)
bandpass (50,92,1.63043)
filter (1491,2726,1.64087)
compute (1140,5673,1.5)
repeat (639,1648,1.16323)

motion---remove(4,4,3)
artifacts---remove(19,1,57)
artifacts---motion(27,1,81)
bodily---sense(124,82,4.53659)
parameters---sense(63,42,4.5)
bodily---parameters(16,1,48)
measure---segment(6,19,0.947368)
bodily---segment(43,41,3.14634)
bodily---measure(181,76,7.14474)
measure---parameters(154,122,3.78689)
bodily---noise(8,1,24)
noise---parameters(1,5,0.6)
light---signal(92,247,1.11741)
light---transmit(323,489,1.9816)
signal---transmit(498,1747,0.85518)
detect---signal(524,1138,1.38137)
detect---transmit(71,175,1.21714)
reflect---transmit(87,142,1.83803)
bodily---transmit(101,41,7.39024)

bodily---reflect(31,36,2.58333)
reflect---tissue(19,1,57)
bodily---tissue(487,10,146.1)
frequency---segment(3,7,1.28571)
dominant---frequency(2,4,1.5)
bodily---frequency(14,5,8.4)
frequency---parameter(7,14,1.5)
segment---subdivided(2,4,1.5)
segment---subsegment(1,1,3)
subdivided---subsegment(0,1,0)
hear---subdivided(0,1,0)
hear---subsegment(0,1,0)
calculate---subsegments(0,1,0)
modify---subsegments(0,1,0)
calculate---modify(2,2,3)
average---calculate(26,29,2.68966)
average---modify(2,1,6)
modify---pulse(9,4,6.75)
average---pulse(31,8,11.625)
candidate---output(0,3,0)
candidate---pulse(0,1,0)
output---pulse(141,194,2.18041)
bodily---valid(0,1,0)
parameter---valid(1,2,1.5)
calculate---output(41,47,2.61702)
degradation---parameter(0,2,0)
degradation---noise(0,5,0)
infrared---red(25,2,37.5)
pulse---red(1,1,3)
infrared---pulse(10,6,5)
infrared---oximetry(1,1,3)
oximetry---pulse(44,1,132)
pulse---signal(238,366,1.95082)
oximetry---signal(4,1,12)
calculate---pulsatile(1,1,3)
blood---calculate(53,1,159)
blood---pulsatile(23,1,69)
oxygen---pulsatile(1,1,3)
blood---oxygen(109,2,163.5)
blood---concentrate(77,1,231)
concentrate---oxygen(76,42,5.42857)
motion-induced---remove(0,1,0)
noise---remove(22,27,2.44444)
motion-induced---noise(0,1,0)
artifacts---motion-induced(0,1,0)
artifacts---noise(15,1,45)
electrical---signal(716,870,2.46897)
electrical---pulse(149,88,5.07955)

light---pulse(58,47,3.70213)
light---oximetry(2,1,6)
raw---segment(1,1,3)
segment---spanning(0,1,0)
raw---spanning(0,1,0)
hear---raw(1,1,3)
hear---spanning(1,1,3)
candidate---frequencies(0,3,0)
candidate---valid(2,1,6)
frequencies---valid(0,2,0)
frequencies---plethysmographic(0,1,0)
plethysmographic---valid(0,1,0)
pulse---valid(2,1,6)
plethysmographic---pulse(1,1,3)
best---frequency(1,1,3)
best---narrow(0,1,0)
frequency---narrow(2,10,0.6)
bandpass---frequency(2,12,0.5)
bandpass---narrow(4,4,3)
filter---narrow(8,10,2.4)
bandpass---filter(34,67,1.52239)
bandpass---segment(0,1,0)
filter---segment(2,4,1.5)
filter---raw(0,1,0)
average---output(6,19,0.947368)
average---signal(52,70,2.22857)
compute---pulse(8,5,4.8)
compute---signal(60,146,1.23288)
best---segment(1,1,3)
best---raw(0,1,0)
frequency---raw(0,1,0)
repeat---segment(3,10,0.9)
raw---repeat(0,1,0)

Top 50 hits with scores

abs 054311705(92.6054) abs 053237765(75.6461) abs 048921017(72.4426) abs 048196460(72.4426) abs 056761414(72.1257) abs 051935430(65.4252) abs 060263148(65.2098) abs 048906190(64.1688) abs 048197521(62.9355) abs 045923655(61.5257) abs 056877226(59.7926) abs 055884270(58.6479) abs 043134459(57.7771) abs 055752853(56.8453) abs 054311594(56.8188) abs 058239669(56.789) abs 051313910(56.7492) abs 056669569(55.6319) abs 060979755(54.3863) abs 054996279(54.0249) abs 053721365(54.0249) abs 058039082(54.0249) abs 049459090(53.7931) abs 053139410(53.5442) abs 056178522(53.2941) abs 060919736(52.3188) abs 048076317(52.2418) abs 060615834(51.5366) abs 058003495(50.3704) abs 061528811(50.3483) abs 056621051(50.1209) abs 061415723(49.6611) abs 058170081(49.2808) abs 053701143(49.2142) abs 056010796(49.2142) abs 051118173(48.8561) abs 053516869(48.5869) abs 054562538(48.5869) abs

059958561(48.5082) abs 051881080(48.4503) abs 052857832(48.4503) abs 052857840(48.4503) abs 055558828(48.4402) abs 053682246(48.4402) abs 058852131(48.4402) abs 057133557(48.4402) abs 059719303(48.3887) abs 044936923(48.3844) abs 059220180(48.3047) abs 058239510(48.2078)

Example 3

Alkylation process using refractive index analyzer

Abstract:

An alkylation process that employs a refractive index analyzer to monitor, control, and/or determine acid catalyst strength before, during, or after the alkylation reaction. In a preferred embodiment, the invention relates to the alkylation of an olefinic feedstock with a sulfuric acid catalyst. The acid typically enters the alkylation reactor train at between from about 92 to about 98 weight percent strength. The concentration of acid is controlled and maintained by monitoring the refractive index of the acid in the product mixture comprising alkylate, mineral acid, water, and red oil. At least one online analyzer using a refractometer prism sensor producing real-time measurements of the refractive index of the solution may be compared to the results of manual laboratory tests on the acid strength of the catalyst using manual sample analyses or titration methods. Periodically, after calibration of the system, samples may be taken to verify the precision of the online analyzer, if desired. In a preferred embodiment, at least one sensor is connected to at least one transmitter and is capable of providing information related to the concentration of alkylation catalyst in the mixture such that the concentration level of acid in the mixture may be monitored and maintained.

What is claimed is:

1. In a method for determining the concentration of acid in a solution containing unkown quantities of said acid within an alkylation reactor, the method comprising forming a solution containing said acid in said alkylation reactor and measuring the concentration of said acid, the improvement comprising measuring the concentration of said acid within the said alkylation reactor with a refractive index sensor having: (a) a refracting prism with a measuring surface in contact with the solution; and (b) an image detector capable of producing a digital signal related to the refractive index of the solution by determining a bright-dark boundary between reflected and refracted light from the measuring surface and correlating the refractive index of the solution to the concentration of said acid in said solution. (Main Claim)

Target words/pairs with SF:

alkylation (50,4,37.5)
refract (136,1313,1.5)
index (160,1820,1.5)
acid (6183,1865,9.94584)
catalyst (1217,277,13.1805)

strength (277,1738,1.5)
react (5260,4195,3.76162)
olefinic (40,8,15)
feed (2784,8745,1.5)
sulfuric (133,80,4.9875)
enter (942,2284,1.5)
train (59,932,1.5)
weight (2006,4237,1.42034)
concentrate (2757,2016,4.10268)
control (4892,24121,1.5)
monitor (668,3393,1.5)
mix (5629,5522,3.05813)
mineral (405,454,2.67621)
red (193,250,2.316)
oil (2005,1475,4.07797)
online (1,36,0.0833333)
prism (4,186,0.0645161)
sense (1411,8708,1.5)
real-time (20,263,1.5)
measure (1428,8592,1.5)
manual (235,1885,1.5)
laboratory (131,60,6.55)
sample (2256,2114,3.20151)
titration (23,2,34.5)
calibrate (121,1083,1.5)
precision (48,693,0.207792)
connect (4157,26955,0.46266)
transmit (366,9513,0.115421)
level (2234,6434,1.5)
unkown (1,1,3)
form (12123,43518,1.5)
surface (5962,29798,1.5)
contact (3736,13940,0.804017)
image (191,6524,0.0878296)
detect (1464,9067,0.484394)
digital (26,3191,0.0244437)
signal (586,17424,0.100895)
bright-dark (1,1,3)
bound (237,1031,0.689622)
reflect (173,3388,0.153188)
light (816,5201,0.470679)
correlate (70,745,1.5)

alkylation---refract(0,1,0)
alkylation---index(0,1,0)
index---refract(36,403,0.26799)
acid---catalyst(15,10,4.5)
acid---strength(3,2,4.5)

catalyst---strength(0,1,0)
alkylation---react(12,1,36)
alkylation---olefinic(0,1,0)
alkylation---feed(1,1,3)
feed---olefinic(13,1,39)
acid---sulfuric(112,58,5.7931)
catalyst---sulfuric(1,1,3)
acid---enter(4,1,12)
acid---alkylation(3,1,9)
alkylation---enter(0,1,0)
enter---react(30,17,5.29412)
alkylation---train(0,1,0)
react---train(1,1,3)
strength---weight(3,11,0.818182)
acid---concentrate(56,23,7.30435)
concentrate---control(61,63,2.90476)
acid---control(16,5,9.6)
monitor---refract(2,3,2)
index---monitor(2,1,6)
acid---refract(0,1,0)
acid---index(0,1,0)
mineral---mix(8,35,0.685714)
mix---red(3,3,3)
mineral---red(1,1,3)
mineral---oil(21,42,1.5)
oil---red(0,1,0)
online---refract(0,1,0)
online---prism(0,1,0)
prism---refract(0,8,0)
refract---sense(0,1,0)
prism---sense(0,1,0)
prism---real-time(0,1,0)
real-time---sense(1,1,3)
measure---sense(58,294,0.591837)
measure---real-time(4,7,1.71429)
real-time---refract(0,1,0)
measure---refract(0,11,0)
index---measure(1,16,0.1875)
laboratory---manual(1,1,3)
manual---strength(0,1,0)
catalyst---manual(0,1,0)
catalyst---sample(3,2,4.5)
manual---sample(3,5,1.8)
manual---titration(0,1,0)
sample---titration(7,1,21)
calibrate---sample(9,13,2.07692)
online---precision(0,1,0)
connect---sense(37,289,0.384083)

concentrate---transmit(1,5,0.6)
alkylation---transmit(0,1,0)
alkylation---concentrate(1,1,3)
catalyst---concentrate(6,3,6)
alkylation---catalyst(7,1,21)
concentrate---mix(82,36,6.83333)
level---mix(26,13,6)
concentrate---level(35,18,5.83333)
acid---level(16,2,24)
mix---monitor(5,7,2.14286)
acid---unkown(0,1,0)
alkylation---unkown(0,1,0)
acid---form(674,113,17.8938)
alkylation---form(0,1,0)
alkylation---measure(0,1,0)
measure---react(52,20,7.8)
concentrate---react(38,13,8.76923)
concentrate---measure(124,93,4)
acid---measure(6,1,18)
acid---react(175,58,9.05172)
index---sense(3,6,1.5)
measure---surface(17,225,0.226667)
detect---image(3,95,0.0947368)
digital---image(0,170,0)
detect---digital(1,56,0.0535714)
detect---signal(51,1138,0.134446)
digital---signal(3,861,0.010453)
bound---bright-dark(0,1,0)
bright-dark---reflect(0,1,0)
bound---reflect(0,3,0)
bound---refract(0,1,0)
reflect---refract(1,28,0.107143)
light---reflect(40,613,0.195759)
light---refract(3,27,0.333333)
correlate---measure(4,31,0.387097)
correlate---surface(4,6,2)
refract---surface(6,46,0.391304)
correlate---refract(0,1,0)
correlate---index(0,10,0)

Top 50 hits scores

abs 045432376(55.0573) abs 056311389(45.4946) abs 042762570(41.5402) abs 040160742(37.0449) abs 06106789&(33.7383) abs 050930885(33.3309) abs 054078300(32.5542) abs 055830498(32.5542) abs 056492812(31.2627) abs 056521472(31.0018) abs 048636975(29.8641) abs 047676043(29.5226) abs 051146754(29.2883) abs 046832114(29.2177) abs 057668892(28.6501) abs 047286040(27.9563) abs 058825908(27.8595) abs 039536041(27.7305) abs 059225343(27.6995) abs 042542227(27.2632) abs

046832106(27.1808) abs 044329393(27.0151) abs 060902973(27.0066) abs 054260531(26.5944) abs 050193570(26.5131) abs 048792454(26.3914) abs 051941627(26.326) abs 051065118(26.1909) abs 060276924(25.6209) abs 056959494(25.6209) abs 041073150(25.3683) abs 040727467(25.3683) abs 041995864(25.3683) abs 040040127(25.3683) abs 043851134(25.3422) abs 053466764(25.0261) abs 048574546(24.94) abs 054340829(24.8966) abs 057982686(24.8152) abs 047537795(24.7818) abs 049884867(24.7818) abs 053166795(24.743) abs 050472135(24.5316) abs 043270735(23.8176) abs 048491861(23.7425) abs 050733516(23.7425) abs 045004900(23.6156) abs 047661160(23.3594) abs 041917421(23.2999) abs 054037484(23.1291)

Example 4

Multiple fluid sample processor with single well addressability

Abstract:

A method for single well addressability in a sample processor with row and column feeds. A sample processor or chip has a matrix of reservoirs or wells arranged in columns and rows. Pressure or electrical pumping is utilized to fill the wells with materials. In a preferred embodiment, single well addressability is achieved by deprotecting a single column (row) and coupling each transverse row (column) independently. After the coupling step, the next column (row) is deprotected and then coupling via rows (columns) is performed. Each well can have a unique coupling event.

In other embodiments, the chemical events could include, for example, oxidation, reduction or cell lysis.

What is claimed is:

1. A method for addressing wells individually in a fluid processing device having a plurality of N column channels and a plurality of M row channels and a plurality of wells arranged in a matrix of N columns and M rows, each of said row channels and column channels having a respective opening corresponding to said plurality of wells, said method comprising:

deprotecting a first of N columns of wells with a first deprotecting agent through a first of said plurality of column channels while protecting each of said N columns except the first of N columns;

coupling M rows of wells with a respective first plurality of coupling agents through said plurality of row channels;

deprotecting a second of N columns of wells with a second

deprotecting agent through a second of said plurality of column channels while protecting each of said N columns except the second of

N columns of wells;

coupling M rows of wells with a respective second plurality of coupling agents through said plurality of row channels; and

continuing the deprotecting and coupling steps until all N columns of wells and M rows have been addressed. (Main Claim)

Target words/pairs with SF:

addressability (1,1,3)
sample (2256,2114,3.20151)
row (175,1873,1.5)
column (1109,1398,2.37983)
feed (2784,8745,1.5)
chip (75,2051,1.5)
matrix (563,1282,1.5)
reserve (810,1695,1.5)
wells (197,174,3.39655)
pressure (4452,11218,1.5)
electrical (663,10634,1.5)
pump (2187,2305,2.84642)
fill (1228,4305,1.5)
deprotect (8,1,24)
couple (1003,9521,0.316038)
transverse (452,3165,0.428436)
chemical (2299,2243,3.0749)
reduce (3746,9984,1.5)
cell (1643,3693,1.5)
lysis (17,1,51)
address (21,2374,0.0265375)
fluid (5391,5884,2.74864)
N (1534,1217,3.78143)
channel (1227,5364,1.5)
M (673,554,3.6444)
open (3523,12717,1.5)
agent (4833,2572,5.63725)
protect (1039,4004,1.5)

addressability---sample(0,1,0)
column---row(8,275,0.0872727)
feed---row(0,22,0)
column---feed(24,14,5.14286)
chip---sample(2,3,2)
matrix---sample(12,2,18)
chip---matrix(0,1,0)
chip---reserve(2,1,6)
matrix---reserve(3,1,9)
matrix---wells(3,1,9)
reserve---wells(8,4,6)
electrical---pressure(11,76,0.434211)
pressure---pump(153,107,4.28972)
electrical---pump(7,9,2.33333)
fill---wells(6,5,3.6)
addressability---deprotect(0,1,0)
addressability---column(0,1,0)

column---deprotect(0,1,0)
couple---deprotect(0,1,0)
column---couple(9,25,1.08)
column---transverse(2,5,1.2)
couple---transverse(0,4,0)
couple---row(0,36,0)
row---transverse(4,41,0.292683)
chemical---reduce(43,18,7.16667)
cell---chemical(4,4,3)
cell---reduce(13,69,0.565217)
lysis---reduce(0,1,0)
cell---lysis(5,1,15)
address---wells(0,1,0)
N---fluid(1,2,1.5)
column---fluid(47,14,10.0714)
N---column(0,5,0)
N---channel(0,43,0)
channel---column(4,6,2)
M---column(0,1,0)
M---channel(0,4,0)
channel---row(4,11,1.09091)
M---row(0,3,0)
row---wells(6,1,18)
channel---wells(0,2,0)
N---matrix(0,4,0)
columns---matrix(3,38,0.236842)
M---N(9,24,1.125)
column---open(15,3,15)
channel---open(69,146,1.41781)
N---deprotect(0,1,0)
N---wells(0,4,0)
columns---wells(3,1,9)
agent---deprotect(1,1,3)
agent---column(3,1,9)
agent---channel(2,4,1.5)
N---protect(0,1,0)
columns---protect(4,2,6)
M---couple(0,1,0)
M---wells(0,1,0)
agents---couple(28,61,1.37705)
agents---row(0,1,0)
M---address(0,3,0)
address---rows(1,127,0.023622)

Top 50 hits scores

abs 061465917(25.7438) abs 053306520(20.1164) abs RE0366609(20.1164) abs 057388253(20.0501) abs 051588870(18.0327) abs 057664706(17.7579) abs 049561502(17.3974) abs 053044878(16.2711) abs

056353588(16.2711) abs 051475383(15.5882) abs 053825110(15.5882) abs 050873601(15.5882) abs 050841840(15.5376) abs 041607252(15.5376) abs 061067792(15.1264) abs 061000841(14.8703) abs 061209856(14.8703) abs 047450746(14.7973) abs 043626998(14.7796) abs 061562084(14.395) abs 043138284(14.2498) abs 052522105(14.2358) abs 046576762(14.139) abs 051358506(14.0315) abs 048226050(14.0315) abs 053955889(13.9945) abs 047449080(13.9159) abs 057859558(13.8528) abs 056351623(13.8528) abs 053957512(13.7256) abs 058887253(13.7256) abs 054746728(13.7049) abs 061361976(13.6892) abs 059894318(13.6526) abs 061565768(13.5869) abs 052061515(13.4768) abs 058007849(13.4495) abs 057053820(13.3965) abs 053407253(13.3965) abs 050966707(13.3848) abs 056501226(13.3848) abs 041559819(13.3475) abs 047802880(13.3178) abs 048714534(13.2503) abs 048140899(13.2503) abs 042723833(13.2306) abs 056482663(13.1287) abs 055189237(13.109) abs 058884556(13.0979) abs 051417189(13.0556)

Although the invention has been described with respect to particular features and embodiments, it will be appreciated that various modifications and changes may be made without departing from the spirit of the invention.

IT IS CLAIMED:

1. Computer-readable code that is operable, when read by an electronic computer, to compare a target concept, invention, or event in a selected field with each of a plurality of natural-language texts in the same field, by the steps of:

(a) associating with each of a plurality of terms composed of non-generic words and, optionally, word groups characterizing the target concept, invention, or event, a selectivity value related to the frequency of occurrence of that term in a database of digitally processed texts in the selected field, relative to the frequency of occurrence of the same term in a database of digitally processed texts in one or more unrelated fields,

(b) determining for each of the plurality of natural-language texts in the same field, a match score related to the number of terms derived from that text that match those in the target concept, invention, or event, and

(c) identifying from among the plurality of natural-language texts in the same field, one or more texts which have the highest match score or scores.

2. The code of claim 1, which operates to determine a match score related to (i) the number of terms derived from that text that match those in the target concept, invention, or event, and (ii) for each term in the target concept, invention or event, a match value related to the term's selectivity value, and

3. The code of claim 2, which is operable to assign to each of the terms in the target concept, invention, or event, a null match value if the selectivity value for that term is below a given threshold.

4. The code of claim 3, which is operable to assign to each of the terms in the target concept, invention, or event, a match value related to a fractional exponential of the corresponding selectivity value.

5. The code of claim 1, for use in comparing a target concept, invention, or event expressed in a natural-language input text, which is further operable to

construct said plurality of terms by (a) identifying non-generic words in the input text, and (b) constructing from the identified words, a plurality of words groups, each containing two or more non-generic words that are proximately arranged in the input text.

6. The code of claim 1, which is operable to carry out step (a) by accessing a database composed words and word-group terms contained in and derived from the texts in the selected field, respectively, and associated with each of said terms, a selectivity value determined from the frequency of occurrence of that term contained in or derived from a plurality of digitally-encoded natural-language texts in the selected field, relative to the frequency of occurrence of that term contained in or derived from a plurality of digitally-encoded natural-language texts in one or more unrelated fields.

7. The code of claim 1, wherein said database includes, associated with each of said terms, a list of indicators that identify the digitally encoded texts in the selected field containing that word, or from which that word group is derived, respectively.

8. The code of claim 7, which is operable, in carrying out step (c), to (i) access the database to determine, for each word and word group from the input text, a match score related to the selectivity value of that word or word group, respectively, and the identity to texts in the selected field containing that word or word group, (ii) add the match values for each of the texts so identified, and (iii) determine, from the rank the identified texts in the selected field with the highest total match scores.

9. The system of claim 8, wherein the match score associated with each word is a function of the associated selectivity value.

10. The system if claim 9, wherein the match score is a fractional exponential of the selectivity value.

11. The code of claim 1, which is further operable, when read by an electronic computer, to identify a selected number X references, where $X \geq 2$, which collectively, provide the greatest number of matches with the terms characterizing the target concept, invention, or event, by the steps of:

constructing groups of all permutations of X references selected from a highest-matching group of Y references, where Y is larger than X,

determining for each group, the number of terms that are represented in at least one of the references of the permutation,

selecting the group containing the highest number of terms.

12. An automated system for comparing a target concept, invention, or event in a selected field with each of a plurality of natural-language texts in the same field, comprising

(a) a database which provides, or from which can be determined, the selectivity value for each of a plurality of terms composed of non-generic words and, optionally, word groups representing proximately arranged non-generic words derived from a plurality of digitally-encoded natural-language texts (i) in the selected field and (ii) in one or more unrelated fields, where the selectivity value of any term is determined from the frequency of occurrence of that term derived from the plurality of digitally-encoded natural-language texts in the selected field, relative to the frequency of occurrence of that term derived from the plurality of digitally-encoded natural-language texts in one or more unrelated fields,

(b) an electronic computer which can access said database, and

(c) computer-readable code which is operable, when read by the electronic computer, to perform the steps of (i) accessing said database, to retrieve or determine selectivity values for each of a plurality of terms composed of non-generic words and, optionally, word groups characterizing the target concept, invention, or event, (ii) determining for each of the plurality of natural-language texts in the same field, a match score related to the number of terms derived from that text that match those in the target concept, invention, or event, and (iii) identifying from among the plurality of natural-language texts in the same field, one or more texts that have the highest match score or scores.

13. The system of claim 12, wherein said match score is related to the number of terms derived from that text that match those in the target concept, invention, or event, weighted by selectivity values of the matching terms.

14. An automated method of comparing a target concept, invention, or event in a selected field with each of a plurality of natural-language texts in the same field, by the steps of:

(a) associating with each of a plurality of terms composed of descriptive words and, optionally, word groups characterizing the target concept, invention, or event, a selectivity value related to the frequency of occurrence of that term in a database of digitally processed texts in the selected field, relative to the frequency of occurrence of the same term in a database of digitally processed texts in one or more unrelated fields,

(b) determining for each of the plurality of natural-language texts in the same field, a match score related to the number of terms derived from that text that match those in the target concept, invention, or event, and

(c) identifying from among the plurality of natural-language texts in the same field, one or more texts which have the highest match score or scores.

15. The method of claim 14, wherein the match score related to the number of terms derived from that text that match those in the target concept, invention, or event, weighted by selectivity values of the matching terms.

16. The method of claim 14, for use in comparing a target concept, invention, or event expressed in a natural-language input text, which further includes constructing said plurality of terms by (a) identifying non-generic words in the input text, and (b) constructing from the identified words, a plurality of words groups, each containing two or more descriptive words that are proximately arranged in the input text.

17 The method of claim 14, wherein said determining includes assigning to each of said terms in the target concept, invention, or event, a match value related

to the corresponding selectivity value, and summing the match values of terms that match those of each of the plurality of digitally processed texts in the given field.

18. The method of claim 17, wherein the match value assigned to each of the terms in the target concept, invention, or event, is a null value if the selectivity value for that term is below a given threshold.

19. The method of claim 17, wherein the match value assigned to each of the terms in the target concept, invention, or event is a fractional exponential of the corresponding selectivity value.

20. The method of claim 14, wherein the selected field is a selected technical field, and the one or more unrelated fields contributing to the database are unrelated technical fields.

21. The method of claim 20, wherein the target concept, invention, or event expressed is a natural-language input text in the form of a patent claim.

22. The method of claim 14, wherein the selected field is a selected legal field, and the one or more unrelated fields contributing to the database are unrelated legal fields.

23. The method of claim 22, wherein the plurality of texts contributing to the word-group database are legal-reporter case notes or head notes.

24. A computer-accessible database comprising

a plurality of terms composed of non-generic words and, optionally, word groups representing proximately arranged non-generic words derived from a plurality of digitally-encoded natural-language texts in a given field, and

associated with each such term, a selectivity value determined from the frequency of occurrence of that term derived from a plurality of digitally-encoded natural-language texts in the selected field, relative to the frequency of occurrence

of that term derived from a plurality of digitally-encoded natural-language texts in one or more unrelated fields.

25. The database of claim 24, wherein each term also has associated with it, identifiers of the natural-language texts in the give field.

26. The database of claim 24, wherein the texts contributing to the database is a selected technical field, the one or more unrelated fields contributing to the database are unrelated technical fields, and the plurality of texts contributing to the database are patent abstracts or claims or technical-literature abstracts.

27. The database of claim 24, wherein the selected field contributing to the database is a selected legal field, the one or more unrelated fields contributing to the database are unrelated legal fields, and the plurality of texts contributing to the database are legal-reporter case notes or head notes.

28. The database of claim 24, wherein the selectivity value associated with terms below a given threshold value is a constant or null value.

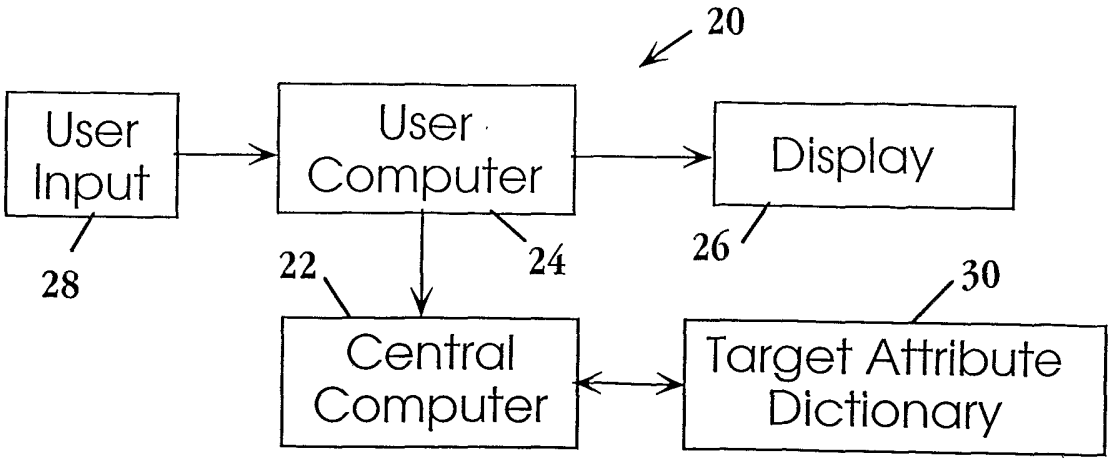

20
User Input
User Computer
Display
28
22
24
26
30
Central Computer
Target Attribute Dictionary

Fig. 1

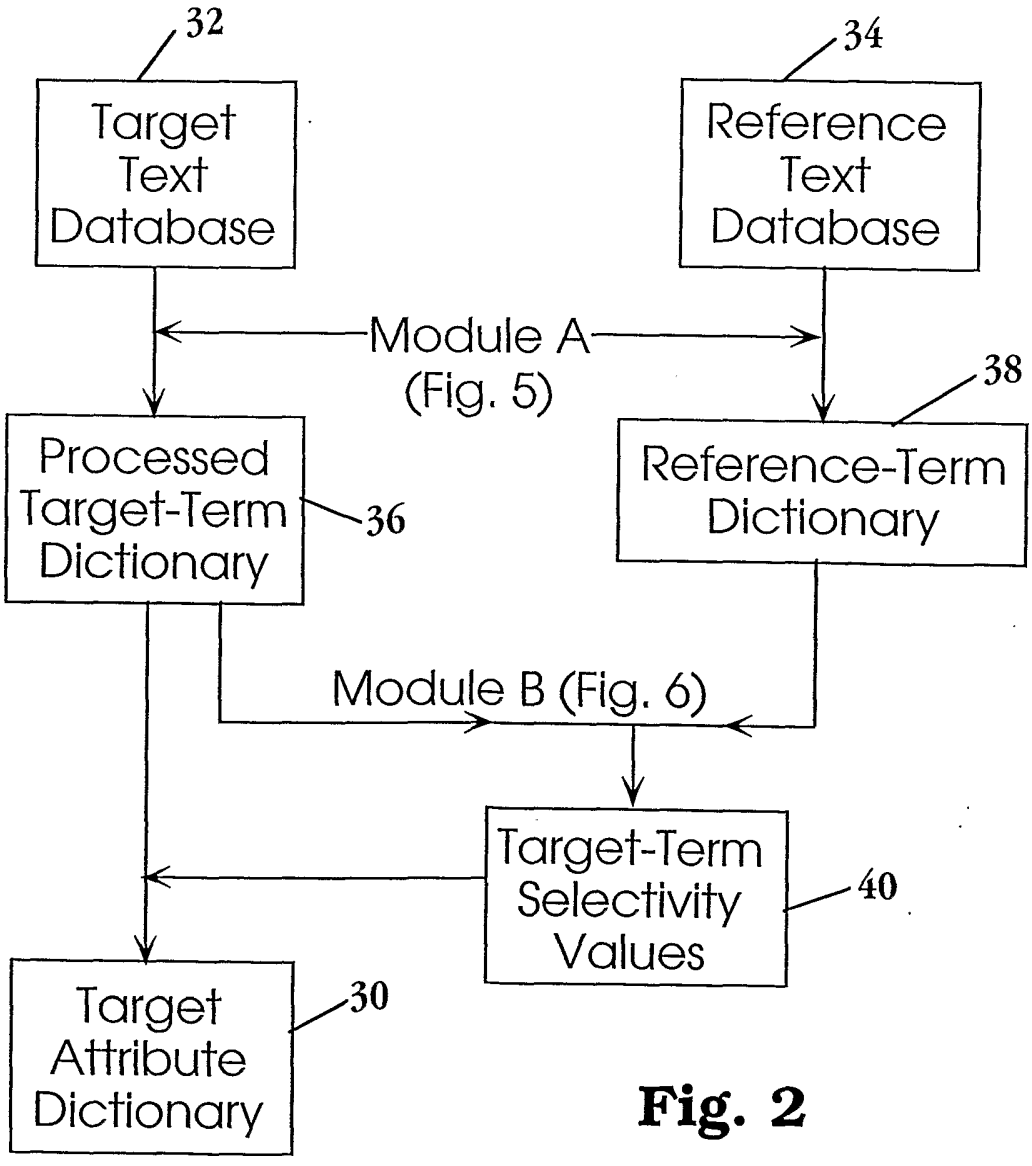

32
34
Target Text Database
Reference Text Database
Module A (Fig. 5)
38
Processed Target-Term Dictionary
36
Reference-Term Dictionary
Module B (Fig. 6)
Target-Term Selectivity Values
40
Target Attribute Dictionary
30

Fig. 2

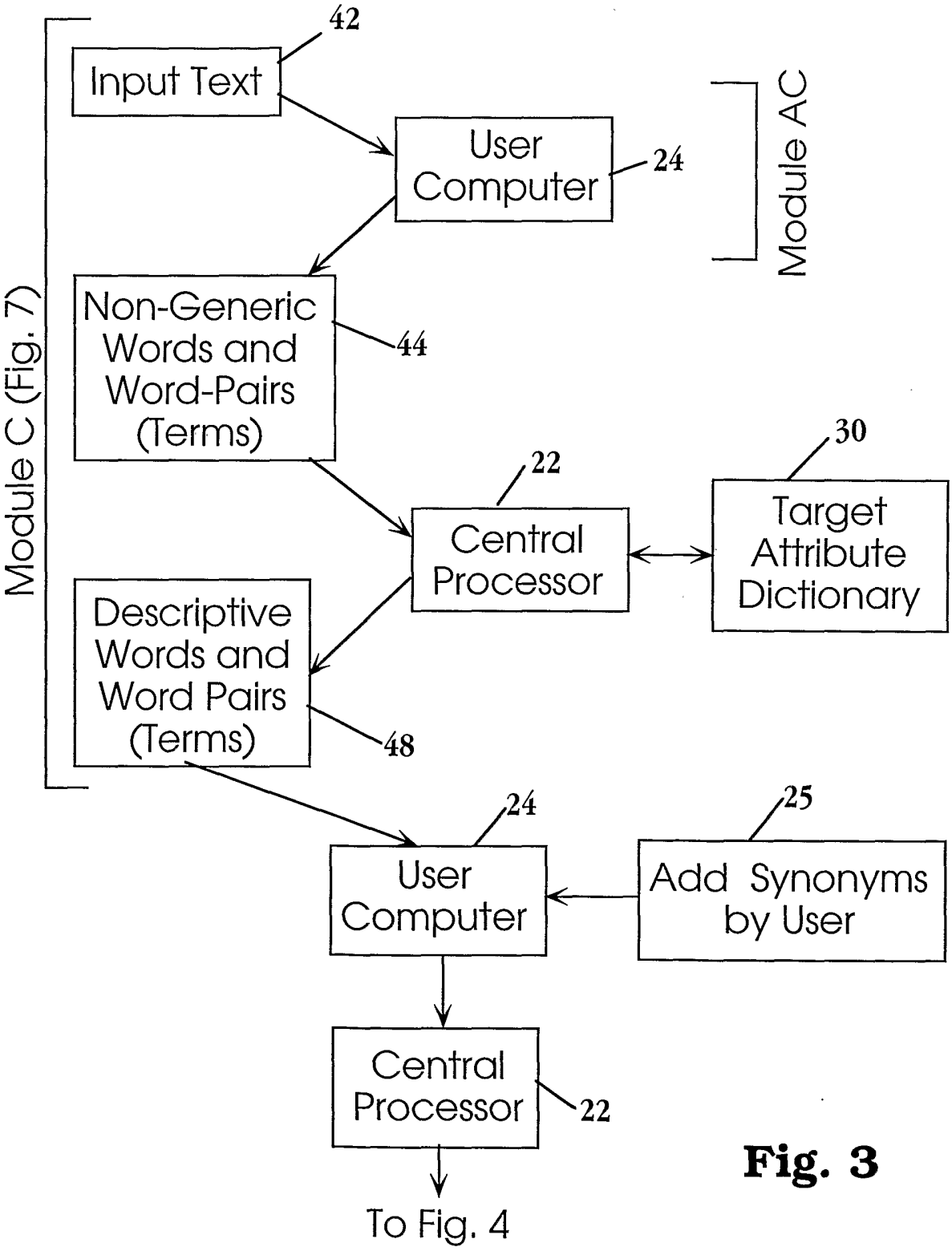

42
Input Text
User Computer
24
Module AC
Module C (Fig. 7)
Non-Generic Words and Word-Pairs (Terms)
44
22
30
Central Processor
Target Attribute Dictionary
Descriptive Words and Word Pairs (Terms)
48
24
25
User Computer
Add Synonyms by User
Central Processor
22
To Fig. 4

Fig. 3

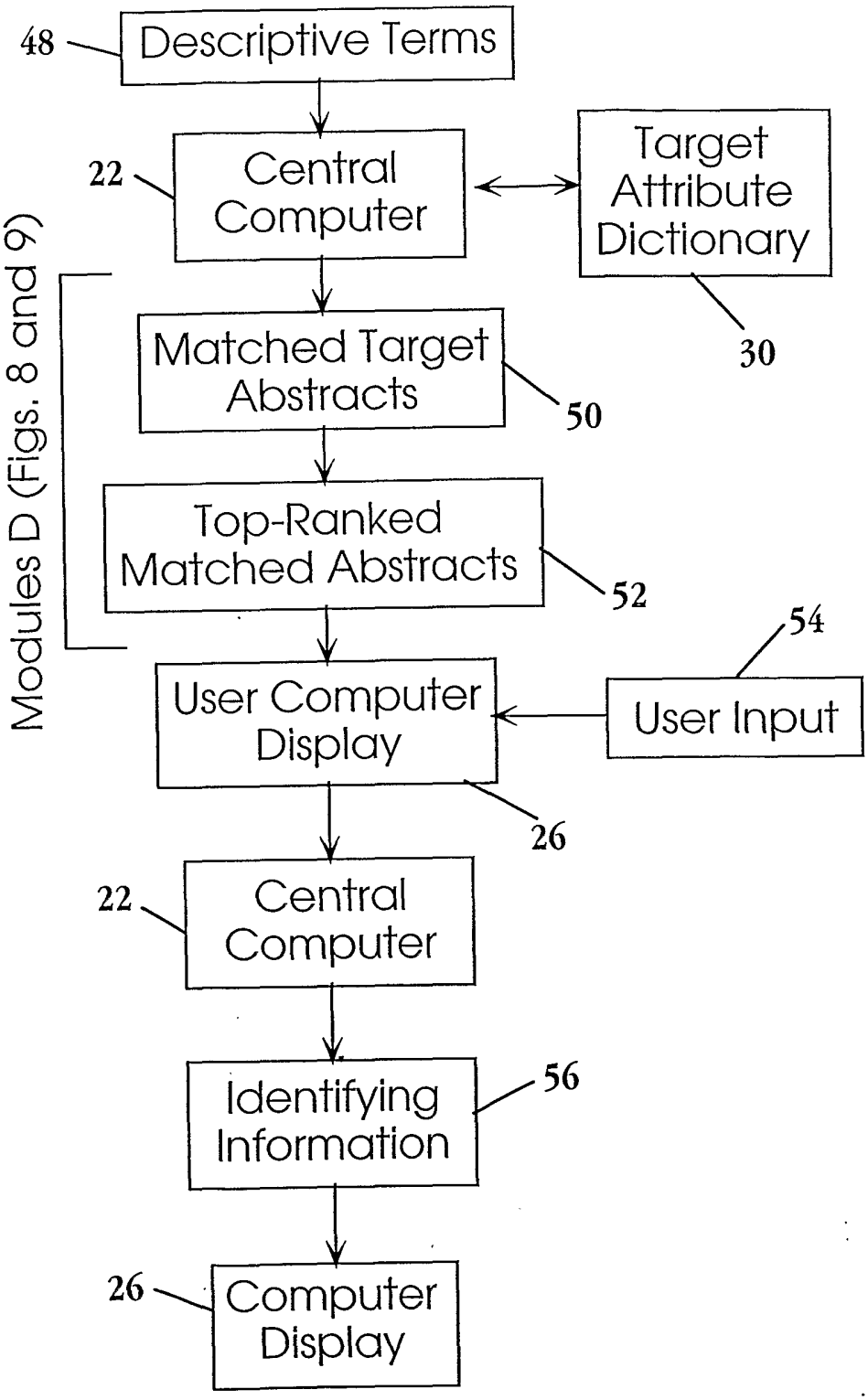

48
Descriptive Terms
22
Central Computer
Target Attribute Dictionary
30
Matched Target Abstracts
50
Top-Ranked Matched Abstracts
52
Modules D (Figs. 8 and 9)
54
User Input
User Computer Display
26
22
Central Computer
Identifying Information
56
26
Computer Display

Fig. 4

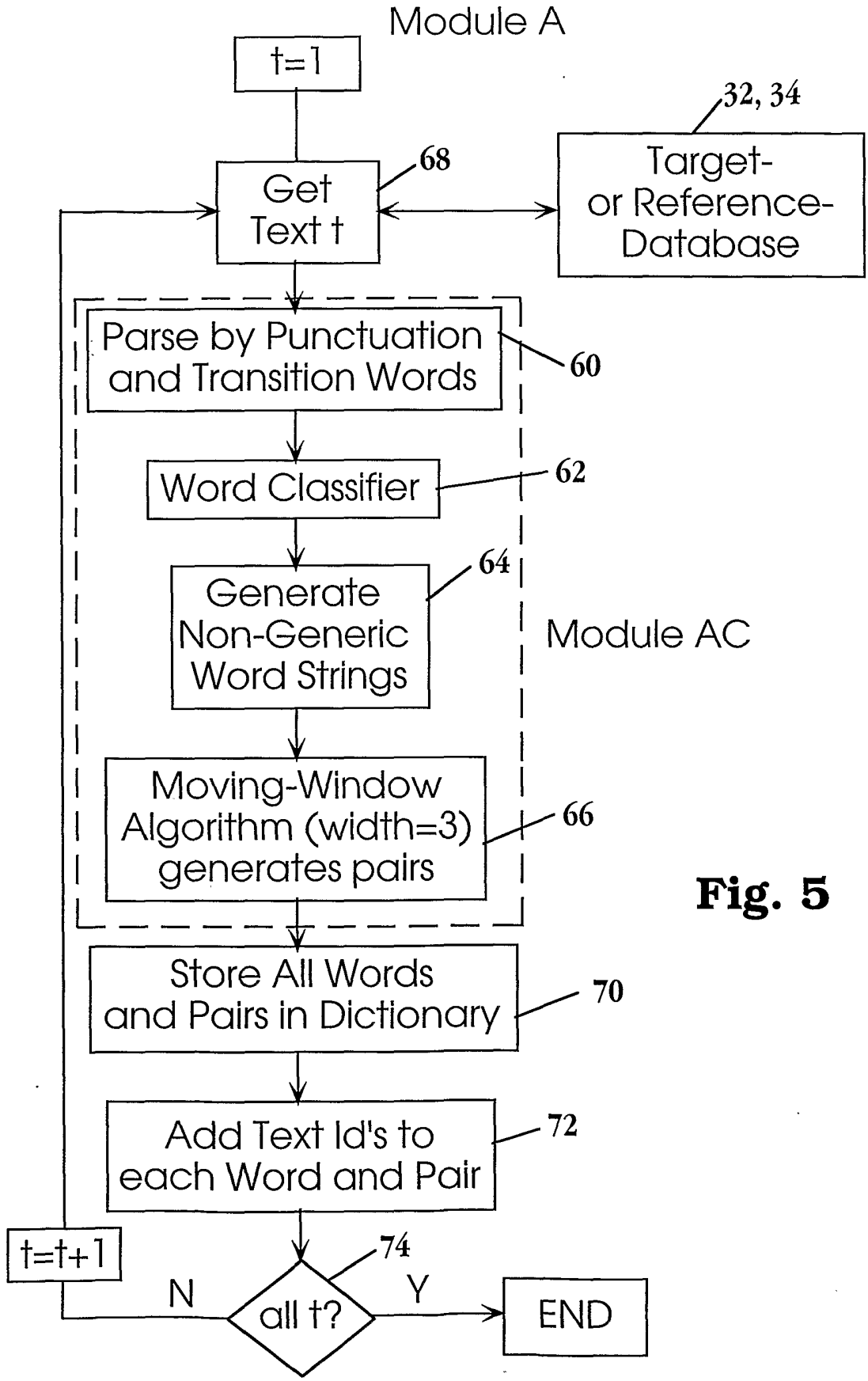

Module A
t=1
32, 34
Target-
or Reference-
Database
68
Get
Text t
Parse by Punctuation
and Transition Words
60
Word Classifier
62
64
Generate
Non-Generic
Word Strings
Module AC
Moving-Window
Algorithm (width=3)
generates pairs
66
Store All Words
and Pairs in Dictionary
70
Add Text Id's to
each Word and Pair
72
t=t+1
74
N
all t?
Y
END

Fig. 5

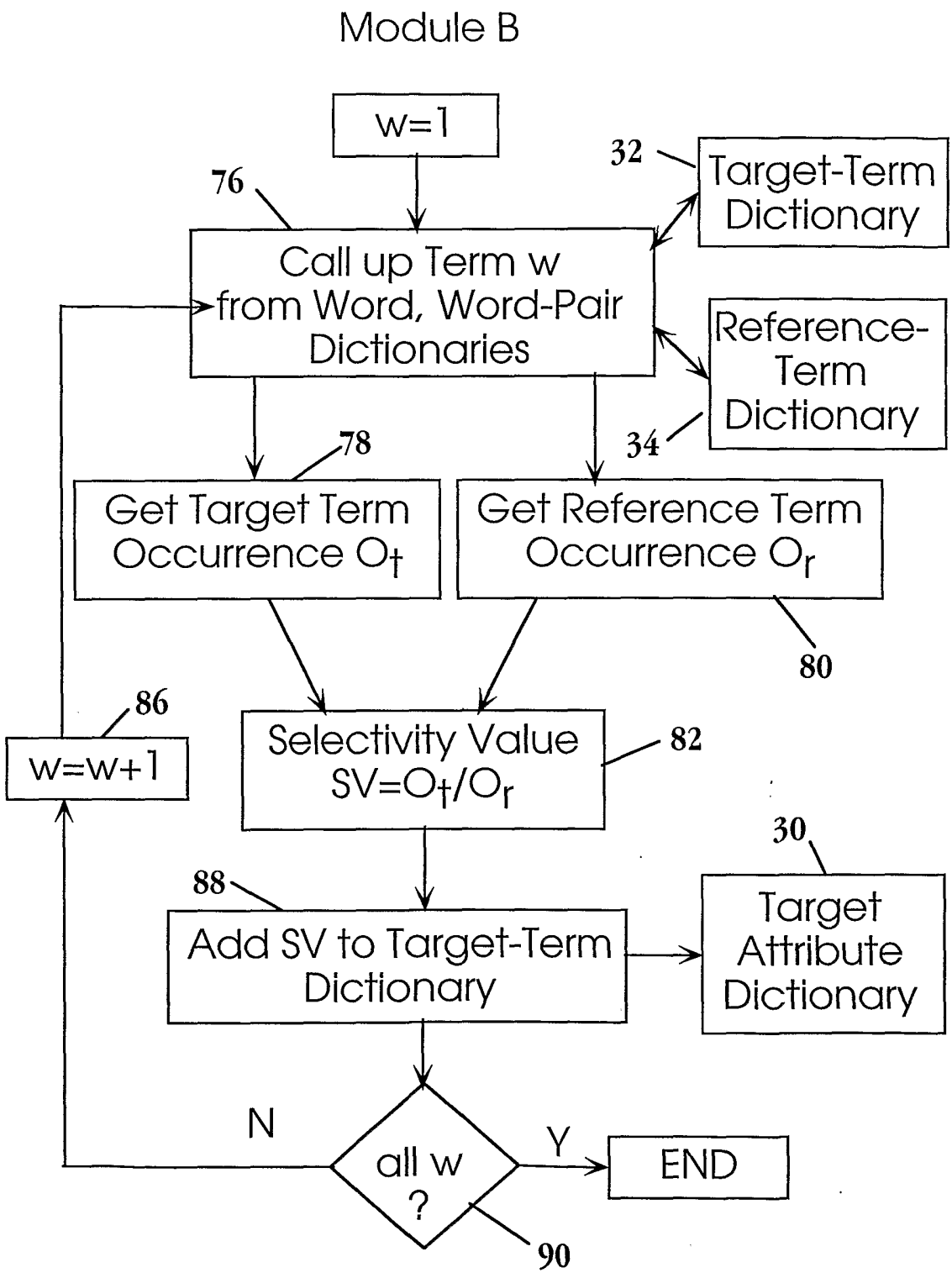

Module B
w=1
76
32
Target-Term Dictionary
Call up Term w from Word, Word-Pair Dictionaries
Reference-Term Dictionary
34
78
Get Target Term Occurrence O_t
Get Reference Term Occurrence O_r
80
86
w=w+1
Selectivity Value $SV=O_t/O_r$
82
30
88
Add SV to Target-Term Dictionary
Target Attribute Dictionary
N
all w ?
Y
END
90

Fig. 6

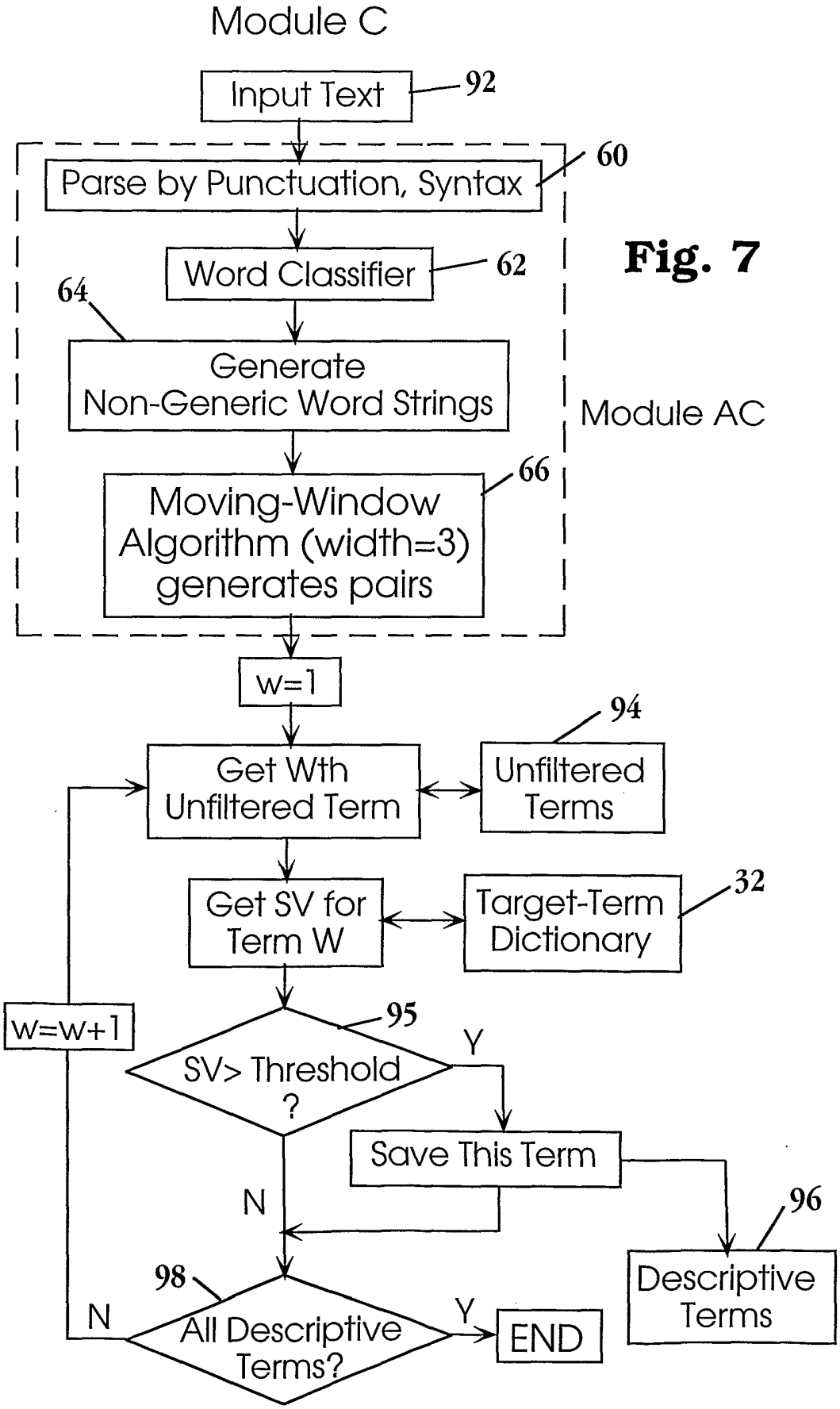

Module C
Input Text
92
Parse by Punctuation, Syntax
60
Word Classifier
62
64
Generate
Non-Generic Word Strings
Module AC
66
Moving-Window
Algorithm (width=3)
generates pairs
w=1
94
Get Wth
Unfiltered Term
Unfiltered
Terms
Get SV for
Term W
Target-Term
Dictionary
32
w=w+1
95
SV> Threshold
?
Y
Save This Term
N
96
Descriptive
Terms
98
N
All Descriptive
Terms?
Y
END

Fig. 7

Module D

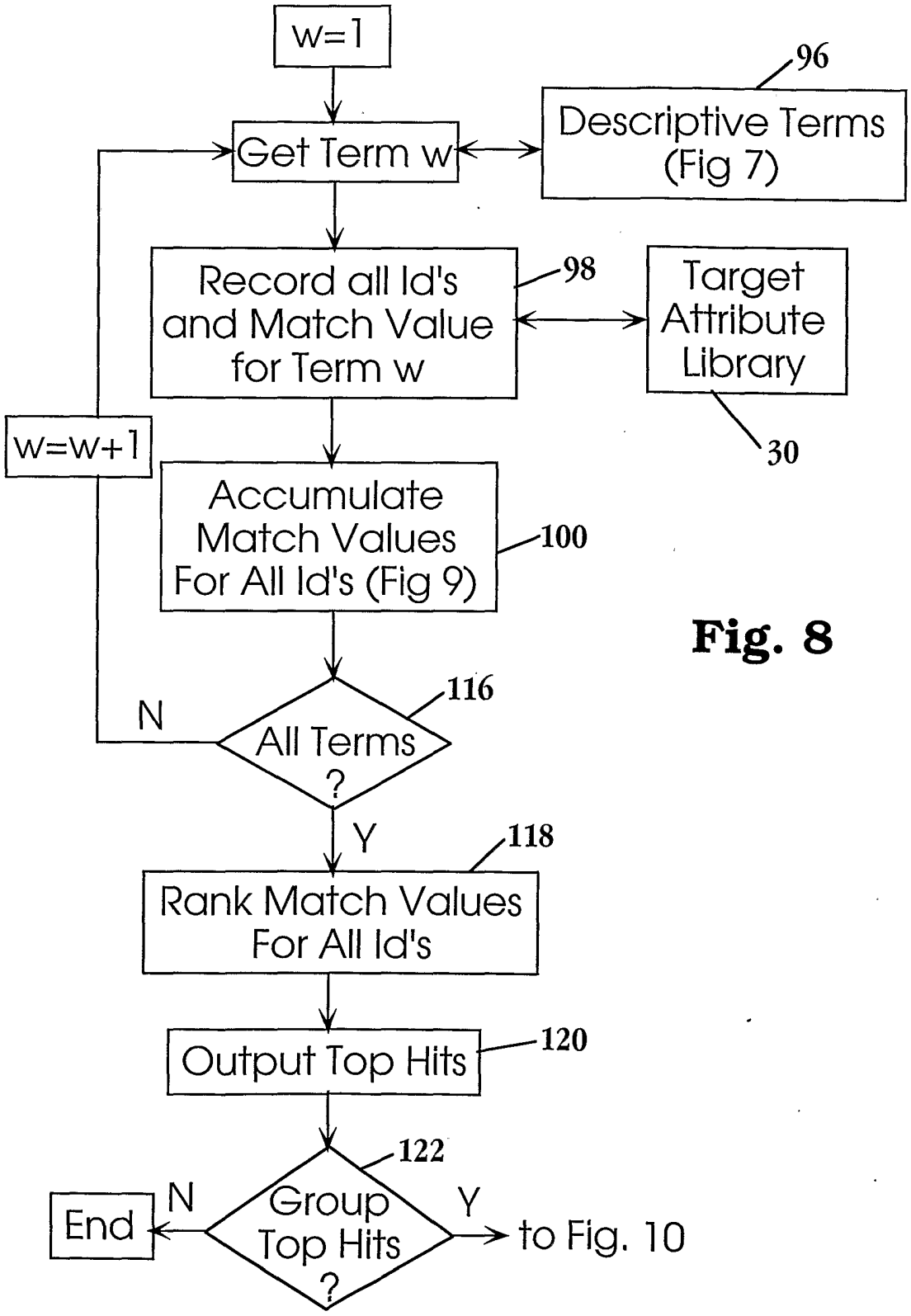


Fig. 8

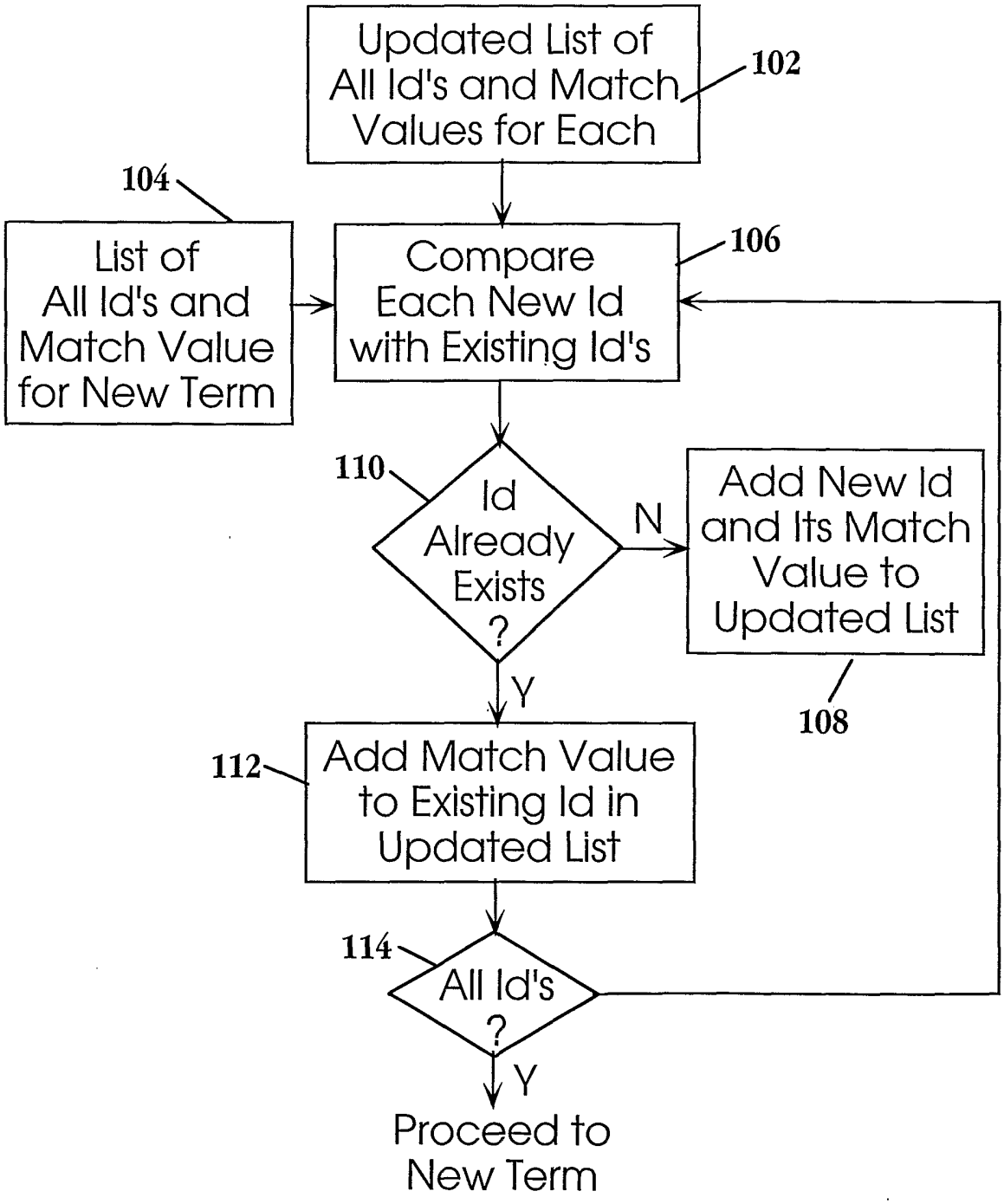

Updated List of All Id's and Match Values for Each
102
104
List of All Id's and Match Value for New Term
Compare Each New Id with Existing Id's
106
110
Id Already Exists ?
N
Add New Id and Its Match Value to Updated List
108
Y
112
Add Match Value to Existing Id in Updated List
114
All Id's ?
Y
Proceed to New Term

Fig. 9

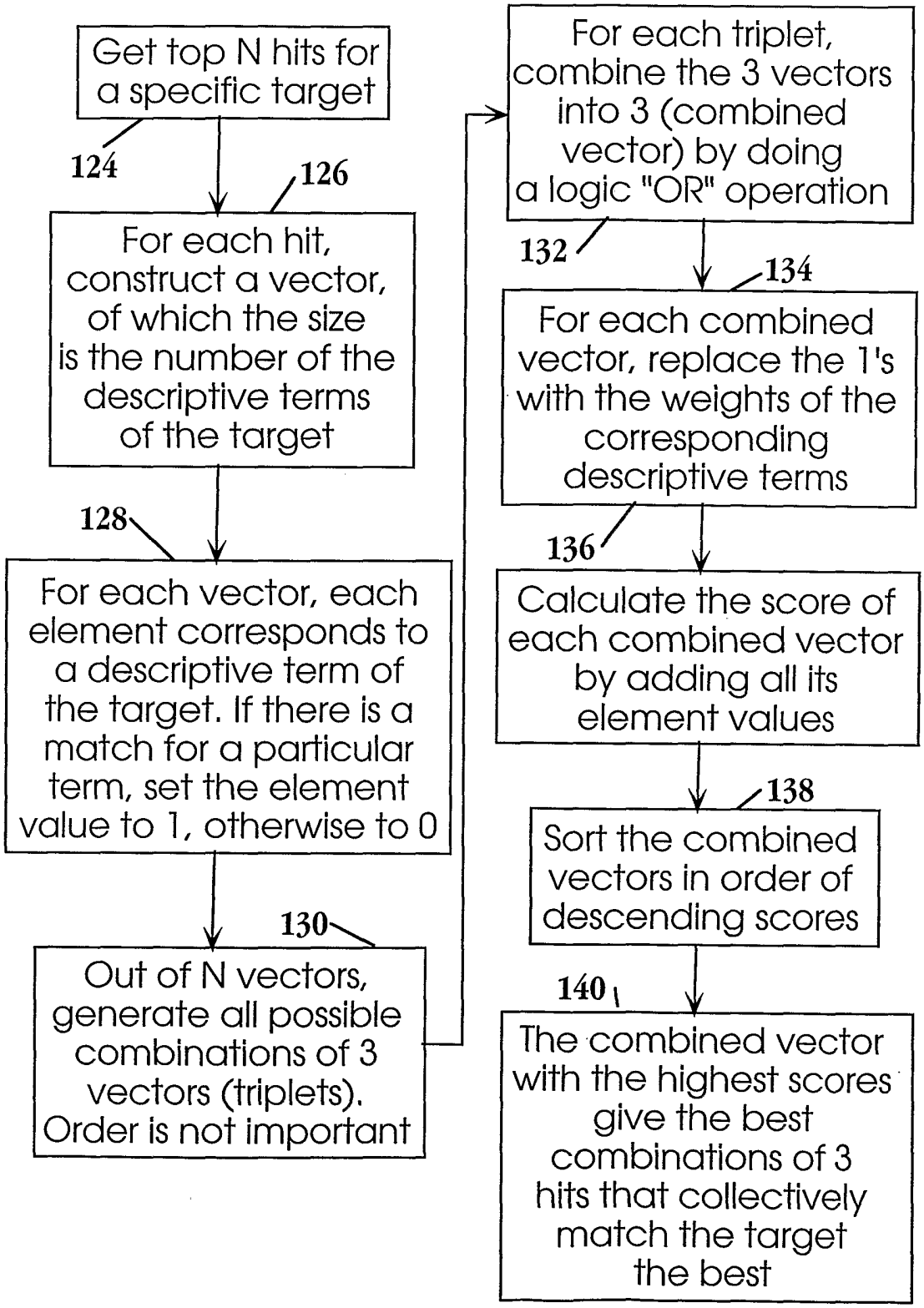

Get top N hits for a specific target
124
126
For each hit, construct a vector, of which the size is the number of the descriptive terms of the target
128
For each vector, each element corresponds to a descriptive term of the target. If there is a match for a particular term, set the element value to 1, otherwise to 0
130
Out of N vectors, generate all possible combinations of 3 vectors (triplets). Order is not important
For each triplet, combine the 3 vectors into 3 (combined vector) by doing a logic "OR" operation
132
134
For each combined vector, replace the 1's with the weights of the corresponding descriptive terms
136
Calculate the score of each combined vector by adding all its element values
138
Sort the combined vectors in order of descending scores
140
The combined vector with the highest scores give the best combinations of 3 hits that collectively match the target the best

Fig. 10

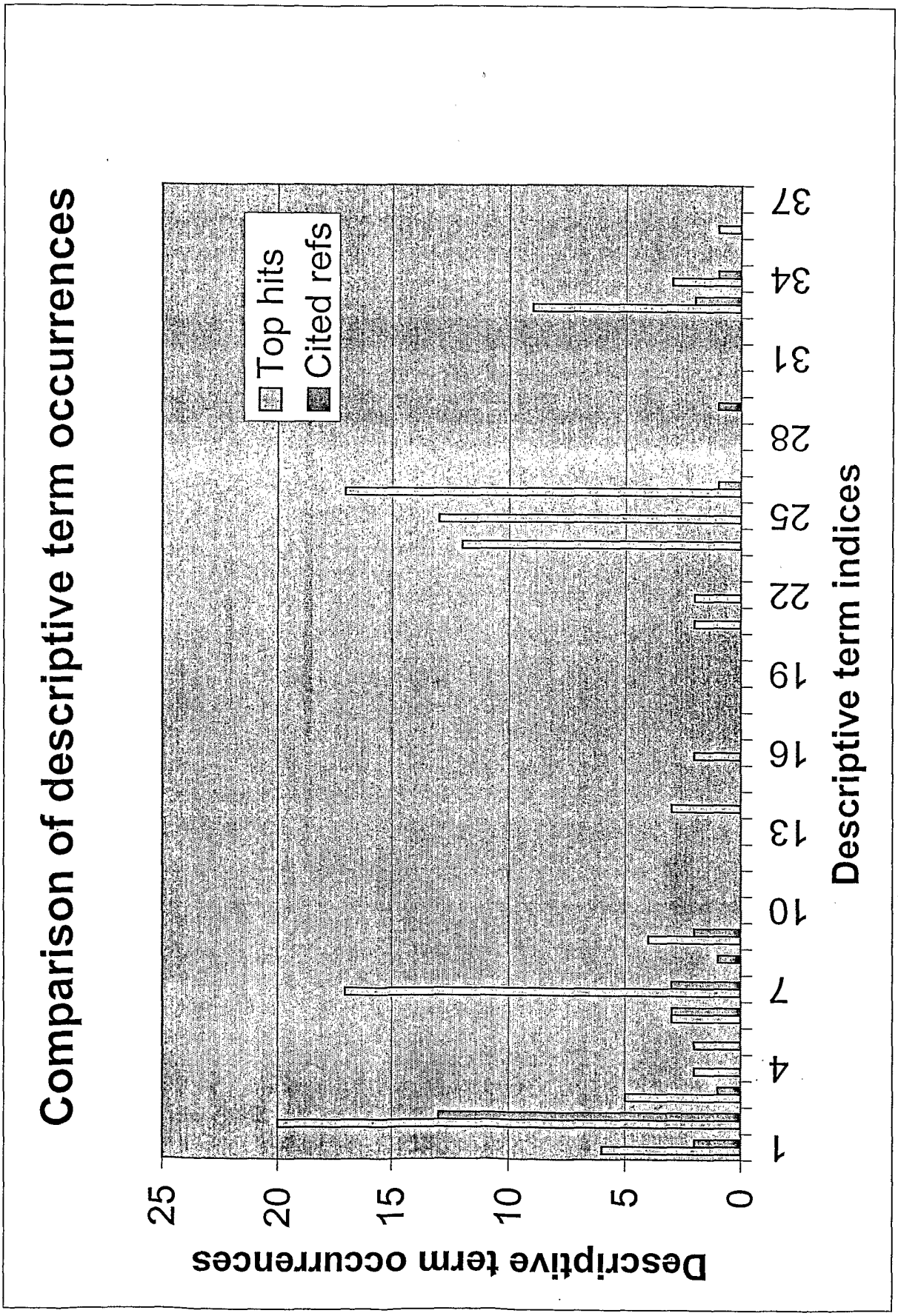

Comparison of descriptive term occurrences
Top hits
Cited refs
Descriptive term occurrences
0
5
10
15
20
25
Descriptive term indices
1
4
7
10
13
16
19
22
25
28
31
34
37

Fig. 11

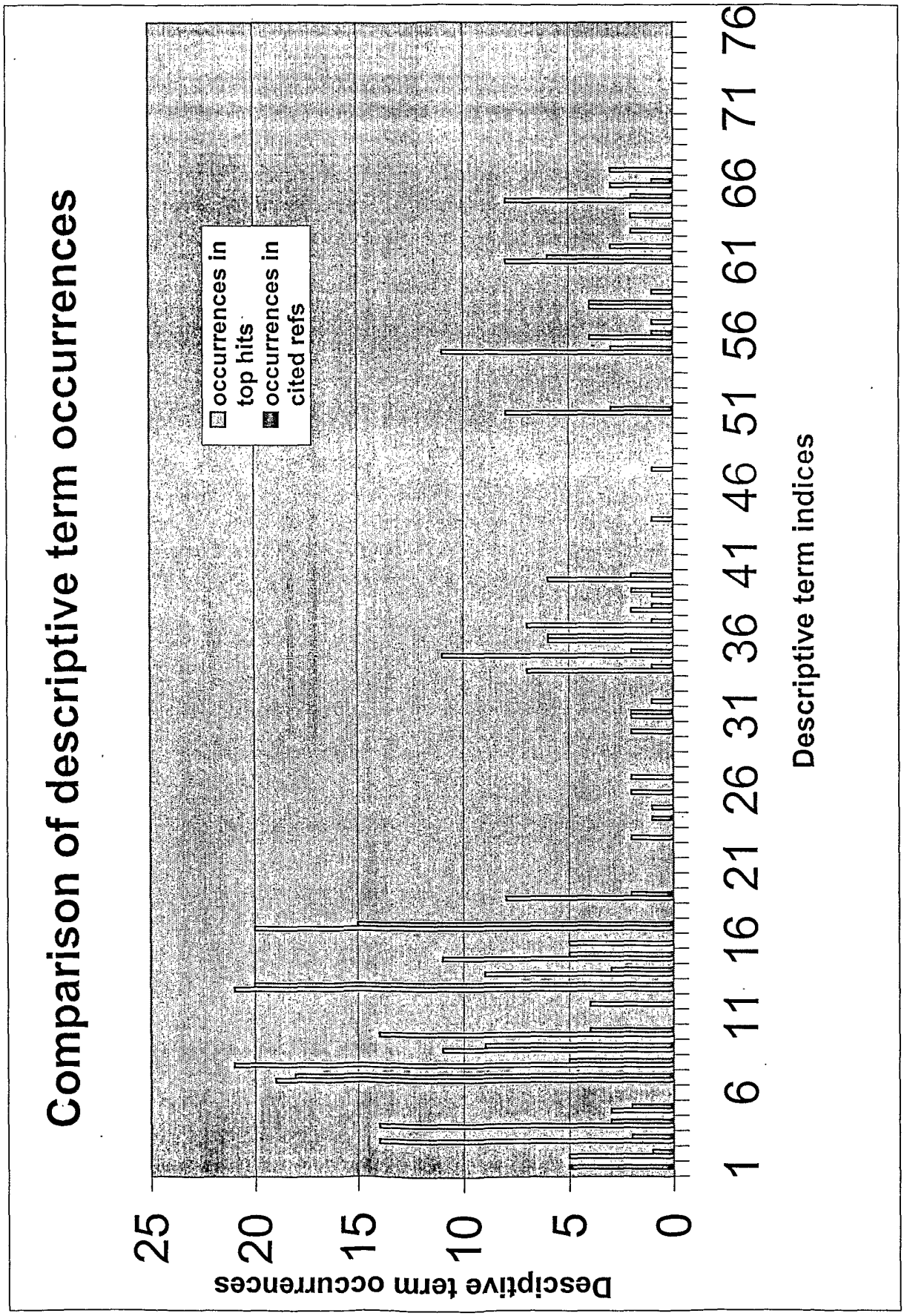

Comparison of descriptive term occurrences
occurrences in top hits
occurrences in cited refs
Descriptive term occurrences
0
5
10
15
20
25
Descriptive term indices
1
6
11
16
21
26
31
36
41
46
51
56
61
66
71
76

Fig. 12

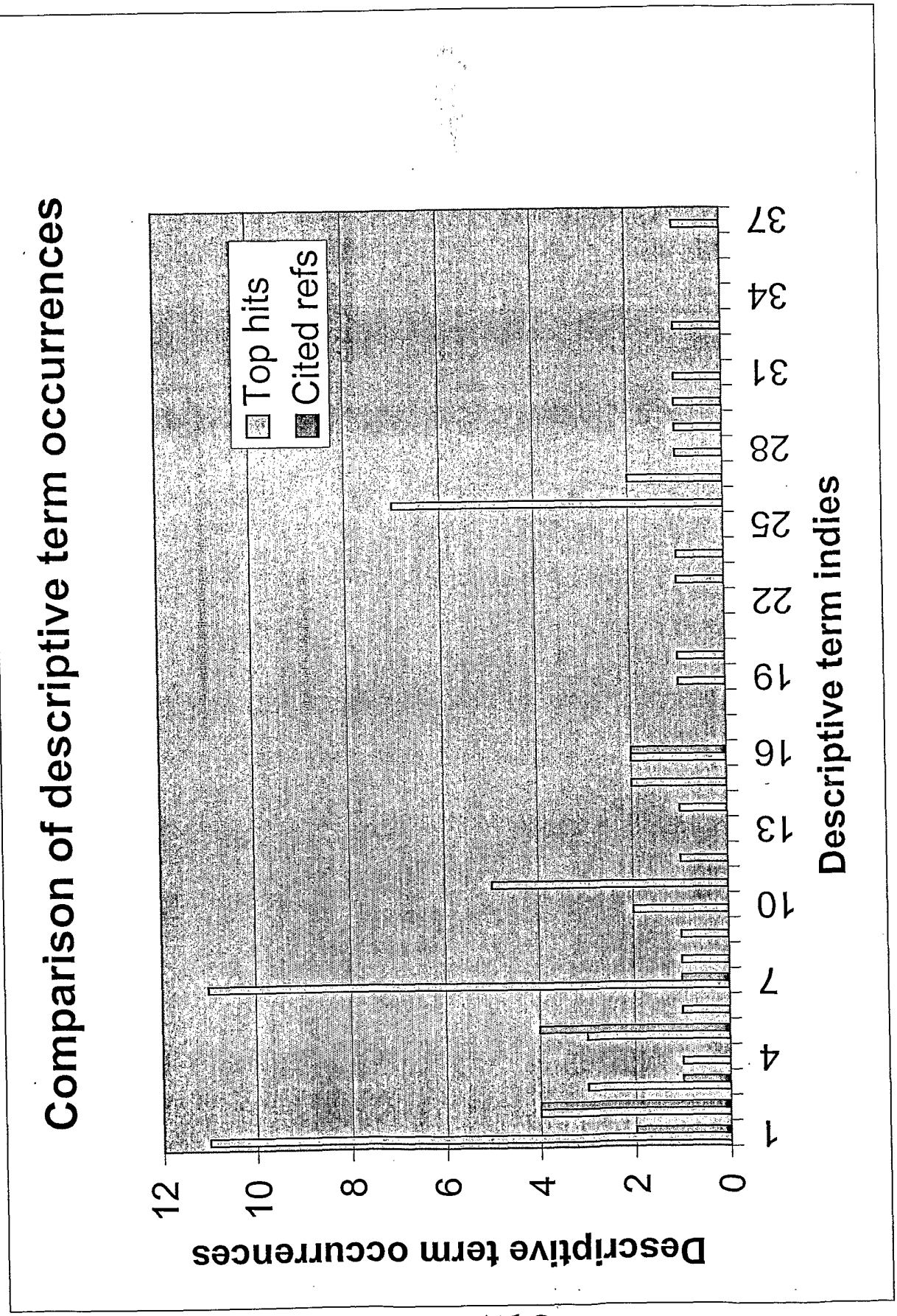

Comparison of descriptive term occurrences
Top hits
Cited refs
Descriptive term occurrences
0
2
4
6
8
10
12
Descriptive term indies
1
4
7
10
13
16
19
22
25
28
31
34
37

Fig. 13

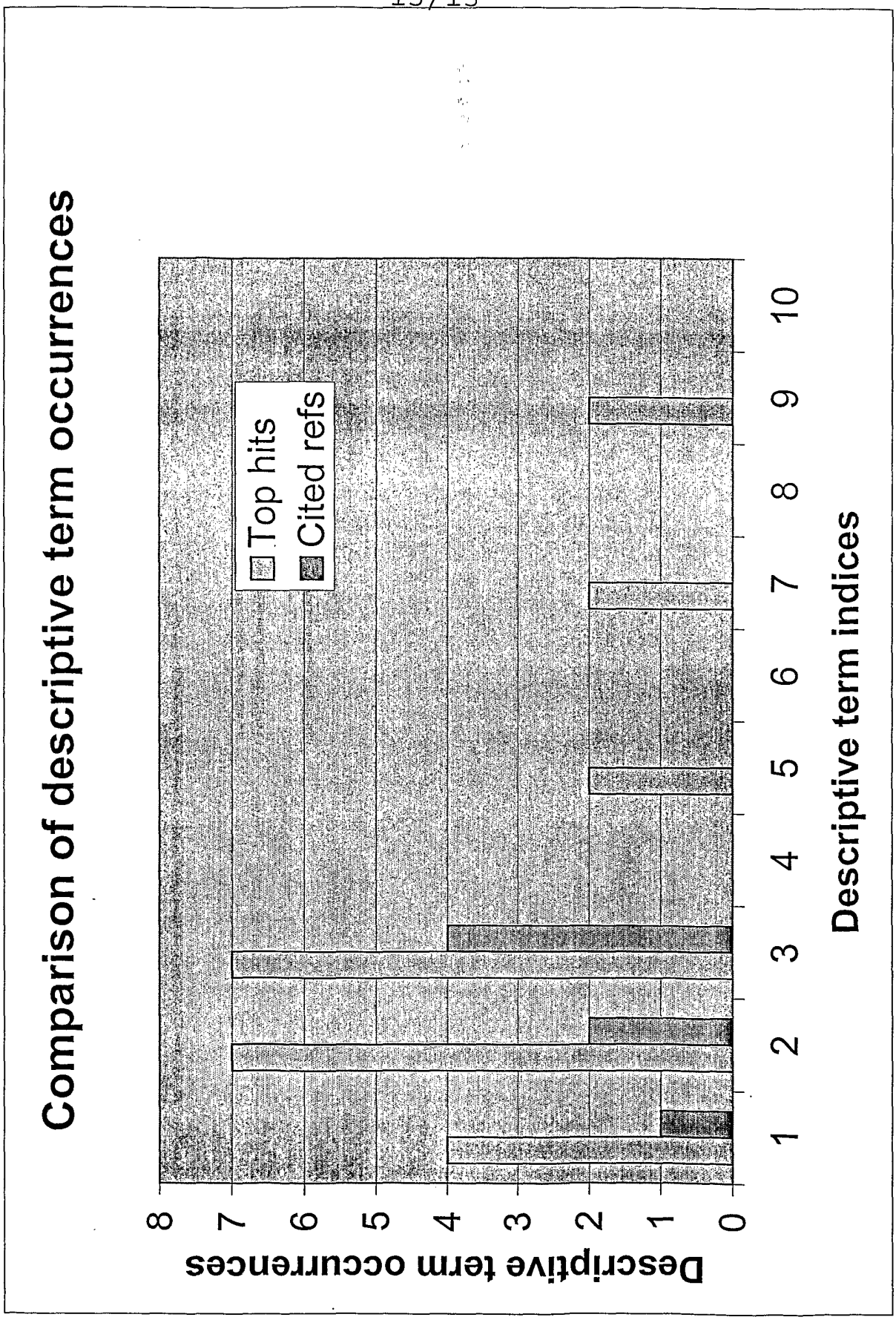

Comparison of descriptive term occurrences
Top hits
Cited refs
Descriptive term occurrences
0
1
2
3
4
5
6
7
8
Descriptive term indices
1
2
3
4
5
6
7
8
9
10

Fig. 14

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US02/21198

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : G06F 17/30
US CL : 707/6

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)
U.S. : 707/6, 707/3-5

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US, 6,374,210 B1 (Chu) 16 April 2002, columns 3-4.	1-28.
A	US, 5,873,056 (Liddy et al) 16 February 1999, see entire reference.	1-28.
A	US, 6,006,221 (Liddy et al) 21 December 1999, see entire reference.	1-28.

☐ Further documents are listed in the continuation of Box C.

☐ See patent family 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 application or patent 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	Date of mailing of the international search report
19 September 2002 (19.09.2002)	29 OCT 2002
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703)305-3230	Authorized officer JOHN BREENE Telephone No. 703-305-3900

Form PCT/ISA/210 (second sheet) (July 1998)