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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
6 March 2003 (06.03.2003)

PCT

(10) International Publication Number
WO 03/019321 A2

(51) International Patent Classification[7]: G06F

(21) International Application Number: PCT/IB02/04468

(22) International Filing Date: 27 August 2002 (27.08.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/314,643 27 August 2001 (27.08.2001) US

(71) Applicant: **E-BASE LTD.** [IL/IL]; 3 Ha'barzel St., Ramat Ha'hayal, 69710 Tel-Aviv (IL).

(72) Inventors: **SEGAL, Irit, Haviv**; Ha'barzel Street 3, Ramat Ha'hayal, 69710 Tel-Aviv (IL). **WINDER, Amir**; Ha'barzel Street 3, Ramat Ha'hayal, 69710 Tel-Aviv (IL).

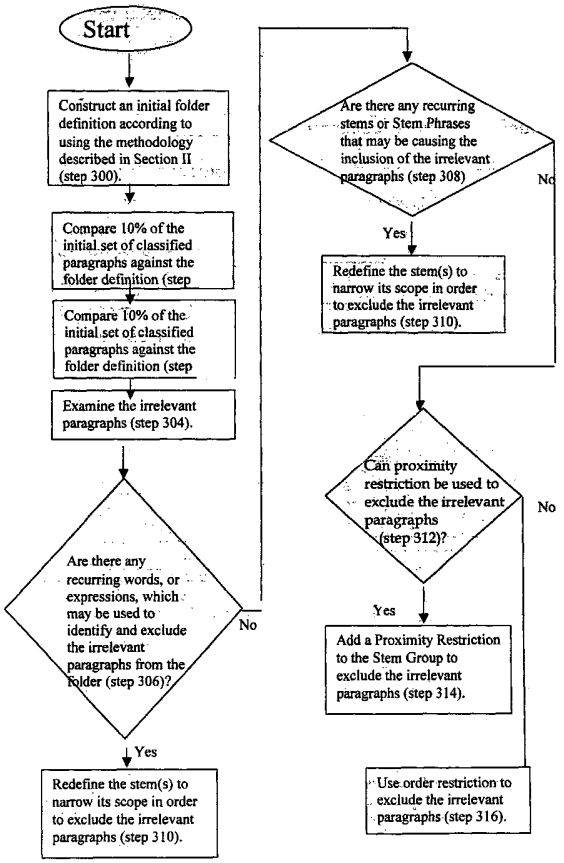
(81) Designated States *(national)*: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR,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

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

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

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

(54) Title: METHODOLOGY FOR CONSTRUCTING AND OPTIMIZING A SELF-POPULATING DIRECTORY



(57) Abstract: A systematic method for detecting meta-ideas used to expanding a skeletal structure. The folder label for each individual first level skeletal folder is placed in a separate collection, and predefined noise words are removed therefrom. A table is tabulated for each collection counting the single word frequency of each word. Words whose frequency falls below a predetermined threshold are removed from the each frequency table. A combined frequency table is created by joining the individual frequency tables wherein meta-ideas are extrapolated from the results of the combined frequency table.

WO 03/019321 A2

METHODOLOGY FOR CONSTRUCTING AND OPTIMIZING A SELF-POPULATING DIRECTORY

Related Application(s)

This patent is related to U.S. Application Serial No. 09/845,196 filed May 1, 2001 entitled "METHOD FOR CREATING CONTENT ORIENTED DATABASES AND CONTENT FILES" which was submitted by the assignee of the present invention.

Claim for Priority

This application claims priority under 35 U.S.C. 120 of U.S. Provisional Application Serial No. 60/314,643 filed August 27, 2001, and which is entitled AUTOMATED FORMATION OF A MODULAR STRUCTURE OF KNOWLEDGE USING MULTI-LINGUAL WORD STEMS".

Field of the Invention:

The present invention relates to a method for constructing and optimizing a directory structure and tools facilitating the same.

Background

The utility of a directory is determined in relation to its breadth and its depth. The granularity of a directory is reflected in the number and length of the branches. If a directory does not have sufficient granularity it will not segregate relevant records from irrelevant records. If the number or length of the branches in the directory exceeds a critical number it may become unwieldy for the user to use.

Conventionally, directory structures are created manually by dividing a topic or field of knowledge into sub-topics, and then subdividing each sub-topic into further sub-topics until a desired level of granularity is reached. An improper selection of topics or

sub-topics will result in the loss of information which is not mapped onto any sub-topic, or the mapping of the information to an overly general topic. Moreover, the list of topics or sub-topics must be dynamic to capture ongoing developments in the field of knowledge.

Unfortunately, the prior art fails to disclose or suggest a systematic way for defining a directory structure or for detecting topics or sub-topics which should be added to a directory structure.

Brief Description of the Drawings

FIG. I- 1 is a screen shot of a sample directory;

FIG. I- 2 is a schematic drawing of a directory;

FIG. I- 3 is a stem phrase according to the present invention;

FIG. I- 4 is a stem group according to the present invention;

FIG. I- 5 is a sample Proximity Restriction according to the present invention;

FIG. I- 6 is a sample paragraph in which words satisfying the stem group of FIG. I- 5 are highlighted;

FIG. I- 7 is an Order Restriction according to the present invention;

FIG. I- 8 is a Combined Order-Proximity Restriction;

FIG. I- 9 is Multi-Stem Group according to the present invention;

FIG. I- 10 is a NOT Phrase according to the present invention;

FIGs. I - 11A-1 and I-11a-2 depict the folder definition for three folders;

FIG. I- 11B is a sample directory for explaining the property of inheritance;

FIG. I- 12 is a Directory constructed from folders created using the methodology of the present invention;

FIG. I- 13 is a flow diagram of the algorithm used to optimize the precision level of the folder definition;

FIG. I- 14 shows two paragraphs, which satisfy the folder definition of FIG. I- 4;

FIG. I- 15 is a flow diagram of the algorithm used to optimize the recall level of the folder definition;

FIG. I- 16 contains a sample noise list for a legal directory;

FIG. I- 17 shows a collection of sentences containing the Concept Stems from FIG. I- 4;

FIG. I- 18 is a table of the frequency of occurrence of combinations of one, two, three and four adjacent words taken from the sentences in FIG. I- 17;

FIG. II-1 is a directory;

FIG. II-2A is a skeletal structure;

FIG. II-2B is a framework structure;

FIG. II-3 is a flow diagram for expanding and optimizing a skeletal structure;

FIG. II-4 is a flowchart for creating framework structure;

FIGs. II- 5A and II-5B are collections of labels;

FIG. II-6 is a sample compilation of noise words;

FIGs. II-7 shows a pointer linking a paragraph to folder;

FIG. II-8 shows the coordinates of paragraph within a file;

FIG. II-9 is a frequency table;

FIG. II-10 is a sample thesaurus;

FIG. II-11 shows the framework structure (FIG. II-2B) appended to the skeletal structure (FIG. II-2A);

FIG. II-12 is a flow diagram of the process for further expanding the skeletal structure;

FIG. II-13A shows a sample folder label;

FIG. II-13B shows a redacted label created by removing noise words from the label of FIG. II-13A;

FIG. II-14 shows the label and definition for an expansion folder;

FIG. II-15 is table showing the rules for replacing prefixes and suffixes for the duplicated stems;

FIG. II-16 is a Venn diagram showing the overlap between two folders;

FIG. II-17 is a flow diagram of the process for organizing the files into a more logical hierarchy;

FIG. II-18 shows an unmatched folder added to a directory for detecting missing skeletal folders.

Detailed Description of the Preferred Embodiments

The methodology of the present invention is the fundamental building block to the construction of an improved self-populating directory. The present invention is used to define the folders which are used to construct the improved self-populating directory. The method for constructing the directory is disclosed in a related application whose disclosure is incorporated by reference.

Every folder in a directory according to the present invention is linked to a collection of paragraphs. To be more precise, paragraphs are automatically classified onto the taxonomy (directory structure). The methodology of the present invention is used to automatically identify paragraphs (textual fragments) which convey a given idea.

According to the present invention a file is a document, web site or the like containing at least one paragraph of text. A paragraph is defined as a text string terminated by paragraph termination symbol such as "¶" or the like, or one or more blank lines. If the text in the file does not contain any recognized paragraph notation then the entire text string is considered to be a single paragraph.

The methodology of the present invention is used to detect paragraphs that convey a particular concept or idea within an appropriate context. As will be explained, the present invention pinpoints the precise paragraph within a multi-paragraph file conveying the specified idea or concept. However, the methodology may be readily adapted to operate on a different unit of text.

The methodology of the present invention reduces the burden to create a self-populating directory.

Moreover, the methodology of the present invention facilitates the mapping of paragraphs whereas conventional directories have difficulties mapping files.

FIG. I- 2 is a sample directory 100 having a root folder 102-A and sub-folders 102-B. Reference numeral 102 is a generic reference to folders 102-A, 102-B.

Each folder 102 in the directory 100 is associated with a label 106 and a definition 108. The label 106 is a description of the folder's concept, and the definition 108 is the criteria used to detect the concept within a paragraph.

An important aspect of the methodology of the present invention relates to the unit of text which is interrogated for a concept. As noted previously according to the present invention the preferred unit of text is the paragraph. However, for some applications the preferred unit of text may be two or more paragraphs.

Roadmap

For the sake of comprehension, the present disclosure is split into four sections. Section I discloses the tools used to specify a folder definition 108, Section I discloses how to create a folder definition 108 using the aforementioned tools, Section II discloses an algorithm for optimizing the precision level of the folder definition; and Section IV discloses an algorithm for optimizing the recall level of the folder definition.

Section I Tools for Specifying the Folder Definition 108

The definition 108 is specified using word stems I-110, where a word stem is an expression ("health care"), a word ("evaluation") or a word fragment ("valu"). A word fragment is a word whose beginning (prefix) or end (suffix) has been truncated.

A word stem I-110 is used to detect words (terms) in which the stem appears at the beginning, end or in middle of the word. The methodology of the present invention uses a series of special operators to specify the manner in which stems I-110 are matched to words within the paragraph. Moreover, the invention uses special operators for specifying stem combinations within a paragraph.

Symbols key:

A hyphen ("-") appended to the end of a stem I-110 signifies a stem which captures only words starting with the stem, e.g., "duty-".

A hyphen ("-") appended to the front of a stem I-110 signifies a stem which captures only words ending with the stem, e.g., "-duty".

A hyphen ("-") appended to both the front and end of a stem I-110 signifies a stem which captures words in which the stem appears in the beginning, middle or end, e.g., "-valu-".

An exact phase is designated through the use of dollar signs ("$") appended to the front and end of a stem, e.g. "act".

Stem Phrase (FIG. I- 3)

As used herein, a Stem Phrase I-120 is a collection of word stems I-110 that pertain to a given idea. FIG. I- 3 is a sample Stem Phrase I-120 used to detect the legal concept "disclosure".

As shown in FIG. I- 3, an OR operator, denoted by the symbol "|" interposed between two stems designates alternative stems, e.g., "duty | duties".

A NOT operator denoted by an exclamation point "!", e.g., "!health care", is used to assure that a certain word stem I-110 does not appear within the paragraph. The appearance of the stem I-110 causes the paragraph to be disqualified from being mapped to a folder 102.

Stem Group (FIG. I- 4)

As used herein, a Stem Group I-130 is a collection of one or more Stem Phrase(s) I-120 that must appear within a paragraph in order to satisfy the folder definition 108. In the event that the Stem Group I-130 contains two or more Stem Phrases I-120, the criterion is the Boolean AND of the respective Stem Phrases I-120.

As will be explained below the Stem Group I-130 may optionally include a Proximity Restriction I-132, an Order Restriction I-134, and a Combined Order/Proximity Restriction 136.

Proximity Restriction (FIG. I- 5)

The Proximity Restriction I-132 enables the user to define the maximal distance between stems from two Stem Phrases I-120. The Proximity Restriction I-132 may be defined by the number of words or characters between stems from the respective Stem Phrases I-120.

According to a preferred embodiment, the Proximity Restriction I-132 is used on the paragraph level, meaning that each paragraph is evaluated to determine whether it satisfies the Proximity Restriction I-132. However, it is possible to specify a different unit of text for evaluation.

In FIG. I- 5, P1, P2 and P3 are Stem Phrases I-120, and the Proximity Restriction I-132 uses the notation "P1-15-P2" to specify a 15 word proximity within a given

paragraph between at least one term from Stem Phrase P1 and at least one term from Stem Phrase P2.

FIG. I- 6 is a sample paragraph in which the stems I-110 from each of the stem phrases I-120 from FIG. I- 5 are underlined showing that the Proximity Restriction I-132 is satisfied.

Order Restriction (FIG. I- 7)

The Order Restriction I-134 is used to define the order in which stems I-110 from corresponding Stem Phrases I-120 appear within a paragraph.

According to a preferred embodiment, the Order Restriction I-134 is used on the paragraph level, meaning that each paragraph is evaluated to determine whether it satisfies the Order Restriction I-134. However, as will be described below, it is possible to specify a different unit of text for evaluation.

FIG. I- 7 shows an Order Restriction I-134 specifying that at least one stem from Stem Phrase P1 (I-120-a) should occur in the paragraph before at least one stem from Stem Phrase P2 (I-120-b).

Combined Order-Proximity Restriction (FIG. I- 8)

The Order Restriction I-134 may be combined with the Proximity Restriction I-132 to form a Combined Order-Proximity Restriction 136. FIG. I- 8 shows a Combined Order-Proximity Restriction 136 which specifies that at least one stem from Stem Phrase P1 (I-120-c) should occur in the paragraph before a term from Stem Phrase P2 (I-120-d).

Multi Stem Group (FIG. I- 9)

A Multi Stem Group I-138 is a union (Boolean OR) of two or more Stem Groups I-120. A paragraph satisfying the criteria of at least one of the Stem Groups I-120-a, I-120-b, . . . , I-120-n will satisfy the criteria of the Multi Stem Group I-138.

FIG. I- 9 shows a sample Multi Stem Group I-138 including Stem Groups I-120-a, I-120-b, I-120-c which pertain to the subject of defenses to defamation torts.

Not Phrase (FIG. I- 10)

A NOT phrase I-140 (FIG. I- 10) is a special type of Stem Phrase I-120 used to disqualify paragraphs which otherwise would be mapped or linked to a folder. The Not Phrase I-140 over-rides the inclusion of a given paragraph specified by a Stem Phrase I-120.

Master Phrase (FIG. I- 11A-1 and 11A-2)

A Master Phrase 142 (FIG. I- 11A-1) is a special type of Stem Phrase I-120 used to define inherited criteria. Like the Stem Phrase I-120, the Master Phrase I-140 is the Boolean OR of a collection of word stems I-110. However, the criteria specified by a Stem Phrase I-120 only applies to the immediate folder 102, and does not affect any other folder in the directory 100. In contrast, the criteria specified in the Master Phrase I-140 are inherited by hierarchically subordinate folders 102 in the directory 100.

The use of a Master Phrase I-140 simplifies the task of specifying a folder definition. The Master Phrase I-140 is most advantageously used to define the context of hierarchically subordinate concepts. In this manner the folder definition 108 of a hierarchically subordinate folder 102 need only contain criteria for detecting the concept, since the context is inherited from a hierarchically superior folder 102.

The inheritance property of the Master Phrase I-140 carries through to each hierarchically subordinate folder 102, i.e., the children, grand-children, great grant children etc of the folder 102. Moreover, changes to the Master Phrase I-140 will change the inclusion criteria of the immediate folder and each of the hierarchically subordinate (child) folders.

FIG. I- 11A shows the definition 108-A of folders 172-A (Negligent Hiring and Supervision), 172-B (Elements of Negligent Hiring) and 172-C (Damages).

FIG. I- 11B is a sample schematic diagram of a directory 170 including folders 172-A, 172-B and 172-C.

The folder definition 108 for folder 172-A includes Master Phrases P1, P2 and P3.

The folder definition 108 for folder 172-B includes Stem Phrases A and B, and inherits Master Phrases P1, P2 and P3.

The folder definition 108 for folder 172-C includes Stem Phrases C, D and E, and inherits Master Phrases P1, P2 and P3..

In directory 170 (FIG. I- 11B) folders 172-B and 172-C are both hierarchically subordinate to folder 172-A. As such, folders 172-B and 172-C inherit the Master Phrases P1, P2 and P3 from the folder 172-A.

Section I Creating a Folder Definition (FIG. I- 12)

The full advantages of folders 102 created using the methodology of the present invention is most apparent when the folders are used to construct a self-populating directory I-500 (FIG. I- 12) of the type described in U.S. Application Serial No. xx/xxx,xxx, entitled "METHODOLOGY FOR CONSTRUCTING AND OPTIMIZING A SELF-POPULATING DIRECTORY" which was filed concurrent with the present invention., hereinafter the 'SELF-POPULATING DIRECTORY specification.

As described in the SELF-POPULATING DIRECTORY specification, the self-populating directory I-500 is constructed from skeletal folders I-502, framework folders I-504 and combined skeletal-framework folders I-506 which are all created using the methodology of the present invention. Thus each of these folders I-502, I-504, I-506 include a label 106 and a definition 108.

As explained in the SELF-POPULATING DIRECTORY specification, the directory I-500 includes a single root skeletal folder $I\text{-}502_{root}$.and plural subordinate skeletal folders I-502. With exception of the root skeletal folder $I\text{-}502_{root}$, each folder I-502, I-504 and I-506 is directly subordinate to only one folder.

The directory I-500 includes one or more hierarchical levels of subordinate skeletal folders I-502.

Framework folders I-504 on a given branch B of the directory I-500 are hierarchically subordinate to all other skeletal folders I-502 on branch B.

Combined skeletal-framework folders I-506 on a given branch B of the directory I-500 are hierarchically subordinate to all other skeletal folders I-502 and framework folders I-504 on branch B.

As described above, the label 106 describes the concept which is being detected, and the definition 108 contains the word stems I-110 etc used to detect the concept within the paragraph.

For ease of comprehension, the method for specifying the definition 108 for the skeletal folders, framework folders and combined skeletal-framework folders will be explained with reference to the following terminology.

Folders I-502-a, I-502-b, . . ., I-502-n are skeletal folders;

Folders I-504-a, I-504-b, . . ., I-504-n are framework folders, where a framework folder is hierarchically subordinate to at least one skeletal folder;

Folders I-506-a, I-506-b, . . ., I-506-n are combined skeletal-framework folders;

Folder definition $108_{skeletal}$ is the combination of stems used to detect the concept specified in the label 106 of a selected skeletal folder I-502-a, I-502-b, . . ., I-502-n.

Folder definition $108_{framework}$ is the Boolean AND of:

[A] the combination of stems used to detect the concept specified in the label 106 of a selected framework folder I-504-a, I-504-b, . . ., I-504-n; and

[B] the combination of stems used to detect the concept specified in the parent (most closely related) skeletal folder I-502-a, I-502-b, . . ., I-502-n.

Folder definition $108_{combined}$ is the Boolean AND of:

[A] the combination of stems used to detect the concept specified in the label 106 of a selected combined skeletal-framework folder I-506-a, I-506-b, . . ., I-506-n; and

[B] the combination of stems used to detect the concept specified in the grandparent skeletal folder I-502-a, I-502-b, . . ., I-502-n, i.e. the parent of the most closely related skeletal folder 520.

In the directory I-500 shown in FIG. I- 12, folder I-502-c is the parent skeletal folder for framework folder I-504-f, because it is the most closely related skeletal folder I-502. Correspondingly, folder I-502-a is the grant-parent skeletal folder for framework folder I-504-f, because it is parent of skeletal folder I-502-c.

Mapping Paragraphs to a Directory

The folder definition 108 is used to detect paragraphs which convey the concept contained in the label 106. A directory I-500 is populated by iteratively comparing each paragraph against each of the folder definitions 108 in the directory I-500. Paragraphs which satisfy the criterion of a given folder definitions 108 are mapped to the folder. This process is described in U.S. Application Serial No. 09/845,196 filed May 1, 2001

entitled "METHOD FOR CREATING CONTENT ORIENTED DATABASES AND CONTENT FILES".

Multilingual Capabilities

As described above, the folder definition 108 is essentially a collection of word stems I-110, where the stem phrases I-120, stem groups I-130, and multi-stem groups I-138 specify the manner in which the stems I-110 must appear within a paragraph for the paragraph to be mapped to the folder. The folder definition 108 is used to detect the concept specified in the folder's label 106.

The methodology of the present invention may be used to create a multi-lingual directory simply by providing additional stem groups I-130 within the folder definition 108. Notably, the multi stem group I-138 may be provided with stem groups I-130 in any number of different languages.

As described previously, each folder is associated with a particular concept, the concept is universal to all languages. A multi-lingual directory eliminates the need to provide separate directory for each language. In a multi-lingual directory according to the present invention the language the user uses to navigates through the directory is independent of the language of the paragraphs mapped to the directory. Thus, a user may use English to locate a desired folder within the multi-lingual directory, and then may retrieve paragraphs mapped to the folder in English, French, German etc.

Section II Optimization of Precision Level (FIG. I- 13)

FIG. I- 13 is a flow diagram of the algorithm for improving the precision of a folder definition 108 according to the present invention.

The process begins with the construction of an initial folder definition 108 using the methodology described in Section I (step I-300).

A sample of 10% from the initial set of classified paragraphs are compared against the folder definition 108, and paragraphs satisfying the criteria of the definition 108 are presented to the user (step I-302).

The user examines the paragraphs to detect irrelevant paragraphs (step I-304), where irrelevelant paragraphs are paragraphs which are not contextually relevant. The displayed paragraph matched all the requisite stem combinations, but the concept detected is used in an irrelevant context. Thus, the folder definition 108 needs to be adjusted to exclude the irrelevant context.

Examine the irrelevant paragraphs to detect recurring words, or expressions, which may be used to identify and exclude the irrelevant paragraphs from the folder (step I-306). These words or expressions are then used to create Not phrases to exclude the irrelevant paragraphs from the folder.

In FIG. I- 14 paragraphs PAR-1 and PAR-2 both satisfy the folder definition 108 of FIG. I- 4. The context of the concept detected in PAR- 1 differs from the context of the context of the concept detected in PAR -2. In step I-306 the user is attempting to identify particular words which signal the irrelevant context.

If no recurring word or expression is detected for excluding the irrelevant paragraphs from the folder in step I-306, then examine the irrelevant paragraphs to detect recurring stems or Stem Phrases that may be causing the inclusion of the irrelevant paragraphs (step I-308).

If a recurring stem or Stem Phrase is detected (in step I-308), then redefine the stem(s) to narrow its scope in order to exclude the irrelevant paragraphs (step I-310).

By manner of example, the Stem Phrase may be changed to include a restriction on the stem so that it is unable to capture the initial set of words or expressions. The Stem

Phrase may also be changed to include restrictions regarding the positioning of the stem within the words (Starting only, Ending only, and Exact phrase).

If no recurring word or expression is detected in steps I-306 or I-308, then examine whether a Proximity Restriction may be used to exclude the irrelevant paragraphs (step I-312). If so, then add a Proximity Restriction to the Stem Group to exclude the irrelevant paragraphs (step I-314).

If no recurring word or expression is detected in steps I-306, I-308 or I-312, examine whether an Order Restriction may be used to exclude the irrelevant paragraphs (step I-316). If so, then add an Order Restriction to the Stem Phrases to exclude the irrelevant paragraphs.

It should be appreciated that if any of the steps I-306, I-308, I-312, I-314 or I-316 drastically reduces the number of paragraphs identified as containing the target concept or idea, then the restriction must be reevaluated to determine whether the restriction has eliminated relevant paragraphs, i.e. caused a recall level decrease.

In the preceding explanation of the methodology of the present invention, the paragraph was used the fundamental unit for capturing an idea. However, one of ordinary skill in the art will appreciate that circumstances may exist in which the use of a paragraph may not prove to be an appropriate unit for capturing an idea. In such cases the methodology of the present invention may be adapted to utilize a Textual Fragment whose length may be defined in terms of a number of sentences it contains, or it may be defined as one or more paragraphs.

Section IV Optimization of Recall Level (FIG. I- 15)

FIG. I- 15 is a flow diagram of the algorithm used to optimize the recall level of the folder definition 108.

The algorithm of FIG. I- 15 is performed on a folder-by folder basis for each folder in the directory. In the case of a multi-lingual directory, the algorithm is separately executed for each language in every folder of the directory.

The process begins with the construction of an initial folder definition 108 using the methodology described in Section I (step I-200).

A sample set of paragraphs are compared against the folder definition 108, and paragraphs satisfying the criteria of the definition 108 are mapped to a folder (step I-202) using the methodology disclosed in U.S. Application Serial No. 09/845,196 filed May 1, 2001 entitled "METHOD FOR CREATING CONTENT ORIENTED DATABASES AND CONTENT FILES".

A list of noise words is compiled (step I-204), where noise words are defined as words that do not have relevance to the directory as a whole. Such noise words typically include digits, dates, seasons, punctuation, single letters, symbols such as "&", currency symbols, participles such as "a", an", "the", and the like. FIG. I- 16 contains a sample noise list for an English language legal directory.

In the case of a multi-lingual directory, separate noise lists are compiled for each language.

Next, the paragraphs mapped in step I-202 are segregated by language (step I-206).

The noise words are removed from each of the paragraphs (step I-208).

As described above each folder definition 108 must contain Stem Phrases I-120 used to detect the concept (label 106) of the folder. In addition, the folder may include stem phrases I-120 for detecting the context of the concept, e.g. Master Phrases 142.

The stems I-110 which collectively form the Stem Phrase(s) I-120 used to detect the folder concept are termed Concept Stems I-110-a. See FIG. I- 11A.

Each of the paragraphs mapped to the folder satisfies the criteria of the definition 108. Consequently, the Concept Stems I-110-a must appear within each of the mapped paragraphs. Sentences containing the Concept Stems I-110-a are extracted and stored in a temporary storage area (step I-210). See FIG. I- 17.

The frequency of occurrence of combinations of one, two, three and four adjacent words is tabulated (step I-212). See FIG. I- 18.

The user visually examines the frequency lists to find terms or expressions which are not already detected by the existing stem phrases I-120, and adds new stem(s) I-110 to the Stem Phrases I-120 as needed to capture the missing term(s) or expressions in the future (step I-214).

It should be appreciated that a high frequency of occurrence is likely to indicate an expression relevant to the idea or concept of the folder.

The present invention provides a methodology for automatically expanding and optimizing a directory of a field of knowledge. A directory 100 (FIG. II-1) is a hierarchical collection of content folders 102 to which text expressing a specified concept is mapped. Notably, each content folder 102 is associated with a particular concept or idea (label 106) and with criteria (definition 108) for detecting the concept within a paragraph or textual fragment, where a textual fragment is a unit of text which is defined in terms of a number of sentences or paragraphs. Textual fragments are compared against the criteria (definition 108) of the respective folders 102 according to pre-defined rules, with textual fragments satisfying the criteria being mapped to the folder(s).

The position of the content folder 102 within the directory 100 defines the context for interpreting the concept. The methodology of the present invention provides a one-to-one function between the definition 108 of a content folder 102 and the contextual meaning of the folder's concept.

Definitions of Textual Units - As used herein, a file is a document, web site or the like containing at least one paragraph of text. A paragraph is defined as a text string terminated by paragraph termination symbol such as "¶" or the like, or one or more blank lines. If the text in the file does not contain any recognized paragraph notation then the entire text string is considered to be a single paragraph. A textual fragment is the basic unit of text mapped to the directory. A textual fragment may be defined in terms of a number of words, sentences or paragraphs. According to a presently preferred embodiment, a paragraph is the basic unit of text which is interrogated to locate a desired concept.

Definition of a Directory - A directory 100 is a hierarchical structure of content folders to which files or textual fragments containing specific concepts have been mapped. Thus, a directory structure becomes a directory after the paragraphs or textual fragments are mapped to the content folders 102. As used in the present disclosure, the initial unmapped directory structure is known as a skeletal structure II-110.

FIG. II-1 is a sample directory 100 of content folders 102, including a root folder 102-A and plural sub-folders 102-B. The last folder 102 on a particular branch 104 is termed an end folder, e.g., folder 102-B_{end}.

The methodology of the present invention is used to expand and optimize the granularity of the skeletal structure II-110. The skeletal structure II-110 is simply a rudimentary arrangement of topics and sub-topics for a given subject or field of knowledge.

Skeletal Structure Definition - FIG. II-2A is a skeletal structure II-110 having plural content folders II-112 in which folder II-112-A is a root folder, folders II-112-B are sub-folders, and folders II-112-B_{end} are end-folders. The folders II-112 are arranged in

branches II-114; each folder II-112 has a single parent folder except the root folder which has no parent folder.

Each skeletal folder II-112 is associated with a label 106 and a definition 108. The label 106 describes the concept or topic of the folder II-112, and definition 108 contains criterion for detecting the expression of the concept within a paragraph.

It is important to appreciate that concepts are detected on a paragraph by paragraph basis, enabling the user to hone in on the precise paragraph conveying a desired concept.

Each skeletal folder II-112 has a unique label 106 to reflect the fact that the concept associated with the skeletal folder II-112 is unique within the directory.

The skeletal folder definition 108 is specified using the methodology disclosed in U.S. Application Serial No. XX/XXX,XXX entitled "METHOD FOR DEFINING AND OPTIMIZING CRITERIA USED TO DETECT A CONTEXTUALY SPECIFIC CONCEPT WITHIN A PARAGRAPH" which was filed concurrent with the present application.

Framework Structure Definition - A separate structure known as a framework structure II-120 is used to expand the granularity of the skeletal structure II-110. The framework structure II-120 is a set of sub-topics used to expand the topics of the skeletal structure II-110. The subtopics within the framework structure II-120 represent the complete set of meta-ideas necessary to define the characteristics of any concept within the skeletal structure II-110. As will be explained below, the framework structure II-120 is automatically generated from the paragraphs mapped to the skeletal folders II-122.

FIG. II-2B is a framework structure II-120 having plural framework (content) folders II-122 in which framework folder II-122-A is a root folder, framework folders II-122-B are sub-folders, and framework folders II-122-B_{end} are end-folders. The

framework folders II-122 are arranged in branches II-114, each folder II-122-B has a single parent folder, and the root folder II-122-A has no parent folder.

Each framework folder II-122 is associated with a label II-126 and a definition II-128. The label II-126 describes the concept or topic of the folder II-122, and definition II-128 contains criterion for detecting the expression of the concept within a paragraph.

The framework folder definition II-128 is specified using the methodology disclosed in U.S. Application Serial No. XX/XXX,XXX entitled "METHOD FOR DEFINING AND OPTIMIZING CRITERIA USED TO DETECT A CONTEXTUALY SPECIFIC CONCEPT WITHIN A PARAGRAPH" which was filed concurrent with the present application.

It should be appreciated that while the same methodology is used to specify the folder definitions 108 and II-128, there is a basic conceptual difference between the two types of folders which is expressed in the way the definition 108, II-128 is specified.

The skeletal folders II-112 are used to define the different subjects or categories of the field of knowledge, whereas the framework folders II-122 are used define characteristics of the skeletal folder II-112.

The characteristics or concepts associated with each of the framework folders II-122 generically describe the concepts associated with the skeletal folders II-112. The "generic" concept of the framework folders II-122 only becomes specific when a context is supplied. As will be explained below, the framework folders II-122 inherit the contextual criterion from the skeletal folders II-112.

The methodology for specifying the folder definition disclosed in U.S. Application Serial No. XX/XXX,XXX entitled "METHOD FOR DEFINING AND OPTIMIZING CRITERIA USED TO DETECT A CONTEXTUALY SPECIFIC CONCEPT WITHIN A PARAGRAPH", includes a concept of inheritance. Inheritance

refers to the situation in which selected criterion (Master Phrases) provided in the skeletal folder definition 108 is inherited by hierarchically subordinate framework folders II-122.

As described in the methodology of the related application, Master Phrases are advantageously used to specify the context criterion. The use of Master Phrases in the folder definition 108 of the skeleton folders II-112 eliminates the need to individually specify context criterion in each of the hierarchically subordinate framework folders II-122. Thus, the context of hierarchically subordinate framework folders II-122 is dynamically defined (inherited) when the framework folder II-122 is added to the directory structure.

ROADMAP

FIG. II-3 is a high level flow diagram providing a roadmap of the methodology for expanding and optimizing a skeletal structure (initial directory structure).

STEP II-300 - As shown, the process begins with the creation of the framework structure II-120 which will be explained below with reference to FIGs. II- 4 through II-10.

A step II-302- II-304 – The skeletal structure II-110 is expanded by appending the framework structure to each of the end-folders II-112-B_{end} of the Skeletal Structure (Step II-302), and irrelevant framework folders are deleted (step II-304). The processes associated with each of these steps will be explained below with reference to FIG. II-11.

STEPs II-306 – II-308 - An iterative process is executed to detect potential concepts missing from the skeletal structure II-110 (step II-306) and add expansion folders II-130 to capture the missing concepts (step II-308). The processes associated with these steps will be explained below with reference to FIGs. 12-20.

STEP II-300 - CREATION OF THE FRAMEWORK STRUCTURE

FIG. II-4 is a flow diagram of the algorithm for creating the framework structure. This process is used to detect the characteristics (meta-ideas) which will be used to increase the granularity of the skeletal structure (initial directory structure) II-110. The detected meta-ideas will be organized into a framework structure II-120 which will be used to systematically expand the skeletal structure II-110.

The disclosed process for detecting meta-ideas was determined empirically. Other processes are contemplated and fall within the scope and spirit of the present invention.

According to a presently preferred embodiment, the meta-ideas are determined by performing statistical processes on labels (concept or topic) 106 of the skeletal folders II-112.

As shown in FIG. II-2A, the first level of folders II-112B1, II-112B2, . . . , II-112Bn are hierarchically subordinate to the root folder II-112A and represent the general topics of the skeletal structure II-110. More particularly, the general topics are described in the labels 106 associated with each of the first level of folders II-112B1, II-112B2, . . . , II-112Bn.

Label Collection - The process begins with collecting the (concepts) labels 106 from all of the content folders II-112B1$_1$ through II-112B1$_n$ for all of the branches II-114 hierarchically subordinate to a selected first level folder II-112B1 into a collection 118-1 (step II-300-2). Step II-300-2 is repeated for each of the first level folders II-112B2, II-112B3, . . ., II-112Bn, collecting the labels 106 into separate collections 118-2, 118-3, . . . , 118-n.

In the sample skeletal structure II-110 shown in FIG. II-2A, folders II-112B1$_1$ through II-112B1$_n$ are all hierarchically subordinate to II-112B1. FIGs. 5A and 5B are collections of labels for II-112B1 and II-112B2.

Removal of Noise Words - Noise words are defined as words that do not have relevance to the directory as a whole. Such noise words typically include digits, dates, seasons, punctuation, single letters, symbols such as "&", currency symbols, participles such as "a", an", "the", and the like. Noise words and noise characters are deleted from each of the collections of labels 118-1, 118-2, and 118-3. . . 118-n (step II-300-4) to create a collection of redacted labels. A sample list of noise words is provided in FIG. II-6. In FIGs. 5A and 5B, the noise words within each of the collections of labels are shown circled. The redacted labels 106 each include at least one word.

Statistical Processes - A frequency table 150-1, 150-2. . . 150-n is tabulated for each word in the label collections labels 118-1, 118-2, 118-3, . . . , 118-n. The frequency table 150 counts the number of times each word occurs within a given collection of redacted labels (step II-300-6).

In the frequency table 150, a low frequency signifies a word which is unlikely to represent a meta-idea relevant to the framework structure II-120. Thus, words whose frequency is below a threshold level T1 are removed from further consideration (step II-300-8).

According to a presently preferred embodiment, T1 is calculated by taking the frequency value of the highest combination and dividing it by the average frequency of the top 100 words. However, other ways for determining threshold T1 are contemplated, and are readily appreciated by one of ordinary skill in the art.

A combined frequency table 170 is compiled by combining the frequency rankings from each of the individual frequency tables 150-1, 150-2. . . 150-n from (step II-300-10).

Empirical evidence has shown that the words (which were taken from the folder labels 106) which occur with the highest frequency within the combined frequency table

170 are likely to be associated with issues which should be included in the framework structure II-120.

The user extrapolates meta-ideas 172 or concepts from the words in the combined frequency table 170 based on his/her knowledge of the subject of the directory. In other words, the user knows from experience that selected words (terminology) are used to describe a meta-idea 172. The user determines whether it is necessary to create a new framework folder II-122 for the meta-idea 172, or whether the concept definition II-128 of an existing (meta-idea) framework folder II-122 needs to be optimized to detect the words in the combined frequency table 170 (step II-300-12).

In operation, results of the combined frequency table 170 are presented to the user. The user examines the words to identify a number of unifying concepts or meta-ideas 172 which may be extrapolated from the words in the combined frequency table 170.

A framework folder II-122 is created for each meta-idea 172 (step II-300-14), wherein the folder label 106 is the meta-idea 172. The folder definition II-128 is created to capture the word(s) from which the meta-idea was extrapolated. However, the folder definition II-128 must be expansive because the meta-idea 172 may be associated with other words which were not reflected in the combined frequency table 170.

Again, the concept definition II-128 is specified using the methodology disclosed in U.S. Serial No. XX/XXX,XXX entitled "METHODOLOGY FOR CAPTURING THE CONTEXTUAL MEANING OF CONCEPTS OR IDEAS WITHIN A PARAGRAPH".

The framework structure II-120 is created by hierarchically organizing the framework folders (meta-ideas) II-122 based on the user's knowledge of the subject of the directory (step II-300-16). Since each of the met-ideas is generic, the hierarchy may be flat.

As will be explained below, the framework structure II-120 in FIG. II-2B is used to elaborate the skeletal structure II-110 (initial directory structure) shown in FIG. II-2A. The framework folders II-122 (FIG. II-2B) correspond to the meta-ideas 172.

Validating the Framework Structure

A validation process is used to verify whether the framework structure II-120 is sufficiently robust to capture all the relevant concepts.

A special content folder termed an unmatched folder II-124 is appended to the root folder II-122A of the framework structure II-120 (step II-300-18). See FIG. II-2B. Like any other content folder, the unmatched folder II-124 has a label II-126 and a definition II-128.

The folder definition II-128 of the unmatched folder II-124 is specified to capture all paragraphs (textual fragments) which were not mapped to any other framework folder II-122.

Mapping of a paragraph to a folder II-122 entails associating a pointer II-140 with the paragraph, and linking the folder II-122 with the pointer II-140. See FIG. II-8A. The location of a paragraph within a file is identified by coordinates 142 which identify the file (document) and relative position of paragraph within the file. See FIG. II-8B.

Paragraphs are mapped to the framework structure II-120 by comparing each paragraph with the folder definitions II-128 (II-300-20). Again, the mapping process is disclosed in U.S. Application Serial No. 09/845,196 filed May 1, 2001 entitled "METHOD FOR CREATING CONTENT ORIENTED DATABASES AND CONTENT FILES".

By definition paragraphs which were mapped to the unmatched folder II-124 were not mapped to any other folder II-122 within the framework structure II-120. Thus, it is

necessary to determine whether these paragraphs contain pertinent concepts which should be added to the framework structure II-120.

The process for identifying concepts for inclusion in the framework structure is similar to the process of steps II-300-2 through II-300-12.

A frequency table II-180 (FIG. II-9) is compiled from the paragraphs mapped to the unmatched folder II-124 (step II-300-22). The frequency table II-180 includes one, two, three and four word combinations from each sentence within the paragraphs mapped to the unmatched folder II-124.

Noise combinations in the frequency table II-180 are removed from further consideration (step II-300-24). According to a presently preferred embodiment, noise combinations are determined using first and second threshold values, however, acceptable results may also be obtained using only the second threshold value.

The first threshold is empirically determined as a positional frequency. According to a presently preferred embodiment, the first threshold is defined to exclude the top two most frequently occurring combinations.

A second threshold is calculated by taking the frequency value of the highest combination that is smaller than the first threshold and dividing it by the average frequency of the top 100 combinations.

Extract word combinations whose frequency is lower than a first threshold but higher than a second threshold.

A thesaurus II-160 is table of records II-162, where each record II-162 contains synonymous terminology within the context of a specific field of knowledge. FIG. II-10 is a sample thesaurus II-160 of legal terminology.

The thesaurus II-160 is used to detect synonymous terminology within the frequency table II-180. The synonymous terminology and its associated frequency values

are removed from the frequency table II-180, and replaced by a single synonymous word or word combination with a frequency value calculated as the sum of the individual frequencies of the synonymous terminology (step II-300-26).

It is now necessary to examine the word combinations in the frequency table II-180 to determine whether the combinations are indicative of framework folders (concepts) II-122 missing from the framework structure II-120, or whether the folder definition II-128 of an existing framework folder II-122 should be optimized to detect the word combination. More precisely, the user extrapolates concepts from the word combinations in the frequency table II-180 based on his/her knowledge of the subject of the directory (step II-300-28).

The user knows from experience that selected word combinations are used to describe a selected concept, and then checks whether an existing framework folder II-122 corresponds to the extrapolated concept. If so, the concept definition II-128 of the corresponding framework folder II-122 needs to be optimized to detect the word combination (step II-300-30).

If no framework folder II-122 corresponds to the extrapolated concept, then a new framework folder II-122 may need to be defined whose concept definition detects the word combination (step II-300-32). Alternatively, the word combination may be irrelevant (noise) to the framework structure II-120.

It should be appreciated that the above process for detecting missing framework folders II-122 should be executed periodically to ensure that newly evolving concepts are included in the framework structure II-120 as new framework folders II-122 or existing concept definitions II-128 are optimized to detect new terminology.

Steps II-302, II-304 Creating Initial Directory Structure (FIG. II-11)

At this stage in the process, we have two distinct structures, the skeletal structure II-110 and the framework structure II-120.

The granularity of the skeletal structure II-110 is expanded using the framework structure II-120. More particularly, a copy of the framework structure II-120 is appended to each end-folder II-112B_{end} of the skeletal structure II-110 (II-302-2).

As will be explained below, additional step are necessary to further expand and optimize the skeletal structure II-110.

FIG. II-11 shows the how the skeletal structure II-110 of FIG. II-2A is expanded by appending the framework structure II-110 from FIG. II-2B to each of the end-folder II-112B_{end}.

It is now necessary to remove unnecessary framework folders II-122 from the newly expanded skeletal structure II-110. Notably, some of the framework folders II-122 may not be relevant within the context of a particular skeletal folder II-112. This determination is made by mapping a sample collection of paragraphs to the expanded skeletal structure (step II-304-2).

The number of paragraphs mapped to each of the framework folders II-122 is tabulated (step II-304-4). See FIG. II-3.

If less than a threshold level of paragraphs is mapped to any framework folder II-122 it is judged to be unnecessary and is deleted from the expanded skeletal structure II-110.

STEPs II-306, II-308 Expanding (Elaborating) the Directory Structure

FIG. II-12 is a flow diagram of the process for further expanding the skeletal structure II-110.

Step II-306-02 - The first step in the process involves mapping a collection of paragraphs to the skeletal structure, and tabulating the number of paragraphs mapped to

each of the end-folders II-122B$_{end}$. Folders having more than a critical number of mapped paragraphs are targeted for expansion.

It is now necessary to automatically generate a set of prospective expansion folders II-130 for expanding the targeted framework end-folder II-122B$_{end}$.

Automated Process for Generating Prospective Skeletal Folders II-112

Step II-306-04 - For each of the targeted end-folder II-122B$_{end}$, create a redacted label II-126$_{red}$ by removing noise words (e.g., FIG. II-6) from the folder's label II-126.

By manner of illustration, FIG. II-13A shows a label II-126 and FIG. II-13B shows a redacted label II-126$_{red}$ created by removing noise words (FIG. II-6) from the label II-126.

Step II-306-06 - For each of the paragraphs (textual fragments) mapped to a targeted end-folder II-122B$_{end}$, extract sentences which contain the redacted folder label II-126$_{red}$.

Step II-306-08 - Tabulate a frequency table II-180 of two, three four words combinations that re-occur in the extracted sentences. See FIG. II-9. These word combinations represent concepts which will be used to expand the targeted framework end folder II-122B$_{end}$.

Step II-306-10 - Noise combinations in the frequency table are removed from further consideration. According to a presently preferred embodiment, noise combinations are determined using first and second threshold values, however, acceptable results may also be obtained using only the second threshold value.

Extract word combinations whose frequency is higher than a first threshold or lower than a second threshold. The first and second threshold limits are used to exclude irrelevant combinations (noise).

According to a presently preferred embodiment the first threshold is empirically determined as a positional frequency. For example, the first threshold may be defined to exclude the top two most frequently occurring combinations. Experience has shown that word combinations whose frequency is higher than the first threshold are noise combinations, i.e., irrelevant combinations.

According to a presently preferred embodiment the second threshold is calculated by taking the frequency value of the highest combination that is smaller than the first threshold and dividing it by the average frequency of the top N combinations. If the value of N is too small then the average frequency will be skewed towards the highly occurring combinations, and too many combinations will be excluded. Conversely, if the value of N is too large then the average frequency will be relatively low, and too many combinations will be included. The inventors of the present invention have found that setting N to be 100 produces a manageable number of combinations. However, other values of N may be appropriate depending on the dataset of files being mapped.

Step II-306-10 will be explained with reference to the frequency table II-180 of FIG. II-9. Let us assume that the first positional threshold is the second highest frequency, and N=100. The top two most frequently occurring word combinations are extracted, and then the second threshold is computed as the average frequency of top 100 remaining word combinations. Word combinations whose frequency value falls below the second threshold are extracted.

Again, the word combinations represent concepts which may be used to expand the targeted framework end folder II-$122B_{end}$.

Out of the remaining word combinations (word combinations falling within the two thresholds), retain only the first M combinations. If the value of M is too large then the table II-180 will contain many irrelevant word combinations. Conversely, if the value

of M is too small then the table II-180 will omit many relevant word combinations. The inventors of the present invention have found that setting M to be 100 produces a manageable number of combinations. However, other values of M may be appropriate depending on the dataset of files being mapped.

Step II-308-02– It is now necessary to create an expansion folder II-130 for each of the concepts in the table II-180. Again, each expansion folder II-130 must have a label 136 and a folder definition 138. The label 136 is determined as a word combination from the table II-180, and the folder definition 138 is created using the methodology of the related application.

Each word combination in table II-180 is a combination of two, three or four words. Each word in the combination is set as a stem phrase and proximity and order restrictions are imposed to preserve the appearance of the original word combination.

More particularly, the folder definition 138 includes a first Stem Group created from the word combination and the definition of the parent folder, and a second Stem Group created from the word combination and the definition of the grand-parent folder.

FIG. II-14 shows the label 136 and folder definition 138 for a sample expansion folder II-130 created from the table II-180 (FIG. II-9).

Step II-308-04 - Next the Stem Phrases of each of the newly created Stem Groups of the new Multi-Stem Group are enhanced. The thesaurus II-160 (FIG. II-10) is used to add synonyms of every stem to every Stem Phrase.

At this stage, each of the stems in the Stem Group is a word taken from the framework folder's label II-128. In order to create a more robust Stem Phrase, we duplicate each of the stems with different prefixes and suffixes using predefined. FIG. II-15 is a sample table showing the rules for replacing prefixes and suffixes for the duplicated stems.

Detecting Unnecessary Expansion Folders II-130

The automatically generated expansion folders II-130 include redundant folders, i.e., folders which have the same folder definition 138 but slightly different labels 136. These labels 136 are essentially identical apart from minor differences in prefixes and suffixes.

Step II-308-06 - The prefixes and suffixes from the words comprising the folder label 106 are deleted or replaced using predefined criteria. FIG. II-15 is a table containing sample criteria for deleting or replacing the prefixes and suffixes.

Step II-308-08 - If two or more folders have the same label 138, then only one of the folders is retained. An arbitrary one of the set of redundant folders II-130 may be retained, as it is assumed that an identical label indicates an identical folder definition 138.

Steps II-308-10 - The paragraphs mapped to the parent folder (target end-folder) are re-mapped to the newly created sub-folders.

Step II-308-12 - If the number of paragraphs mapped to an expansion folder II-130 is below a threshold level calculated as a percentage of the total number of paragraphs originally mapped to parent folder, then the sub-folder is deleted.

Still further, duplicative (redundant) expansion folders II-130 may be detected by examining the overlap between a selected pair of folders. To facilitate understanding let us designate one of the folders A and the other B. If the two folders share a large number of paragraphs it indicates that one of the folders is redundant.

Empirical evidence has demonstrated that if the number of mutual paragraphs exceeds a threshold percentage L then one of the folders is deemed to be redundant. For the sake of example, let us assume that L is 75%.

Step II-308-14 - The calculation is performed by checking whether the paragraphs (textual fragments) within the intersection of A and B is greater than 75% of the number of paragraphs within the union of A and B. See FIG. II-16. If so, then one of the skeletal folders II-130 is redundant, and it is now necessary to determine which of the folders should be retained.

The expansion folder II-130 which is most closely related to the paragraphs contained in the intersection of A and B is retained. As will be explained, the redundant folder is deleted, and the definition of the non-redundant folder is modified to map the paragraphs (textual fragments) not included in the intersection.

The skeletal folder to be retained is determined by calculating a relevance factor R for each folder (step II-308-16). The relevance factor is determined by dividing the number of paragraphs within the intersection of A and B by the total number of Paragraphs mapped to the folder. Let us assume that there are 15 paragraphs within the intersection of A and B, 25 paragraphs in A and 35 paragraphs in B. Then folder A is retained since $15/25 > 15/35$.

The folder definition 138 of the redundant expansion folder II-130, i.e., its Multi-Stem Group is added to the folder definition 138 of the retained expansion folder II-130, and the redundant expansion folder II-130 is deleted (II-308-18).

Steps II-308-14 through II-308-18 are repeated until there is no mutual overlap of over 75% between the folders. The end result is a flat arrangement of folders.

Step II-310 Organizing the Expansion Files II-130 into a Hierarchy

FIG. II-17 is a flow diagram of the process for organizing the expansion files II-130 into a more logical hierarchy beneath the target end-folder II-122b_{end}. This process detects which expansion folders II-130 have less than a threshold degree of commonality

(sibling folders) and should remain on the same hierarchical level, and which expansion folders II-130 should be arranged in a parent-child relationship.

It should be appreciated that at this stage, duplicative expansion folders II-130 have been removed. According to the presently preferred embodiment, duplicative folders were defined as folders which have a 75% overlap of mapped paragraphs. The remaining folders are related by less than the threshold (75%) overlap.

Sibling Test

For the purposes of explaining the sibling test, let us designate the newly created expansion folders as D1 through Dn, and designate the target end-folder II-122b_{end} as C.

A collection of paragraphs are mapped to folders D1 through Dn and C (step II-310-02).

Steps II-306-04 through II-306-08 (FIG. II-12) are executed for each of the folders D1 through Dn and C, yielding for each a frequency table II-180 (FIG. II-9) of two, three and four word combinations (step II-310-04).

Part 1 of the Sibling Test

If the number of mutual paragraphs between D1 and D2 is zero, then D1 and D2 are siblings (step II-310-06). This pre-screening is repeated for D1 and D3, D1 and D4 through D1 and Dn.

Part 2 of the Sibling Test

Check whether the label of D2 through Dn matches any of the combinations in the frequency table of D1 (Step II-310-08)

If the label of Dn does not match any of the combinations in the frequency table of D1, then D1 and Dn are regarded as siblings (step II-310-10).

Parent Child Relationship Test

If the folders D1 and Dn are not determined to be siblings using the two part sibling test, then we know that the folders belong in a parent-child relationship, but it remains to be determined which folder is the parent and which the child.

From the second part of the sibling test, we know that the label of D2 through Dn matches one of the combinations in the frequency table of D1.

C_1, C_2, C_n are the ranked frequencies from the frequency table of C.

$D1_1$, $D1_2$. . . $D1_n$ are the first, second and n-th ranked frequencies from the frequency table of D1.

$D2_1$, $D2_2$. . . $D2_n$ are the first, second and n-th ranked frequencies from the frequency table of D2.

CD1 is the frequency value of the name of D1 within the frequency table of C.

D1Dn is the frequency value of the name of Dn within the frequency table of D1.

DnD1 is the frequency value of the name of D1 within the frequency table of Dn.

R1 is defined as C2/CD1.

R2 is defined as D11/D1D2.

R3 is defined as D22/D2D1.

R4 is defined as C2/CD11.

If R1> R2 then (Step II-310-12)

No – D1 is the parent of D2

Yes – If R4 > R3 then (step II-310-14)

No – D2 is the parent of D1

Yes – If CD2 > CD1 then (step II-310-16)

No – D1 is the parent of D2

Yes – D2 is the parent of D1

Using Unmatched Node to Detect Blind Spots

In the present context, blind spots are topics which are not captured by any of the content folders II-112, II-122, II-130 within the directory structure.

As before, blind spots are detected using the unmatched folder II-124, where the unmatched folder is a content folder whose folder definition 108 is constructed to capture paragraphs which are not mapped to any other content folder II-112, II-122, II-130.

As shown in FIG. II-18, the unmatched folders II-124 are attached to the directory 100 on the same hierarchical level as the end-nodes II-112B$_{end}$ of the skeletal framework within the directory structure 100. In other words, an unmatched folder II-124 is attached beside each of the top level framework folders II-122B1, II-122B2, . . II-122Bn.

The content folders of the directory are populated by mapping paragraphs to the directory structure.

By definition paragraphs which were mapped to the unmatched folder II-124 were not mapped to any other folder II-112, II-122, II-130 within the expanded skeletal structure II-110. Thus, it is necessary to determine whether these paragraphs contain pertinent concepts which should be added to the skeletal structure II-120.

The process for identifying concepts for inclusion in the framework structure is identical to the process of steps II-300-22 through II-300-32.

A frequency table II-180 (FIG. II-9) is compiled from the paragraphs mapped to the unmatched folder II-124 (step II-300-22). The frequency table II-180 includes one, two, three and four word combinations from each sentence within the paragraphs mapped to the unmatched folder II-124.

Noise combinations in the frequency table II-180 are removed from further consideration (step II-300-24). According to a presently preferred embodiment, noise combinations are determined using first and second threshold values, however, acceptable results may also be obtained using only the second threshold value (step II-300-26).

Noise combinations in the frequency table II-180 are removed from further consideration (step II-300-24). According to a presently preferred embodiment, noise combinations are determined using first and second threshold values, however, acceptable results may also be obtained using only the second threshold value.

The first threshold is empirically determined as a positional frequency. According to a presently preferred embodiment, the first threshold is defined to exclude the top two most frequently occurring combinations.

A second threshold is calculated by taking the frequency value of the highest combination that is smaller than the first threshold and dividing it by the average frequency of the top 100 combinations.

Extract word combinations whose frequency is lower than a first threshold but higher than a second threshold.

A thesaurus II-160 is table of records II-162, where each record II-162 contains synonymous terminology within the context of a specific field of knowledge. FIG. II-10 is a sample thesaurus II-160 of legal terminology.

The thesaurus II-160 is used to detect synonymous terminology within the frequency table II-180. The synonymous terminology and its associated frequency values are removed from the frequency table II-180, and replaced by a single synonymous word or word combination with a frequency value calculated as the sum of the individual frequencies of the synonymous terminology (step II-300-26).

It is now necessary to examine the word combinations in the frequency table II-180 to determine whether the combinations are indicative of framework folders (concepts) II-122 missing from the framework structure II-120, or whether the folder definition II-128 of an existing framework folder II-122 should be optimized to detect the word combination. More precisely, the user extrapolates concepts from the word combinations in the frequency table II-180 based on his/her knowledge of the subject of the directory (step II-300-28).

The user knows from experience that selected word combinations are used to describe a selected concept, and then checks whether an existing framework folder II-122

corresponds to the extrapolated concept. If so, the concept definition II-128 of the corresponding framework folder II-122 needs to be optimized to detect the word combination (step II-300-30).

If no existing folder II-112, II-122, II-130 corresponds to the extrapolated concept, then a new skeletal folder II-112 may need to be defined whose concept definition detects the word combination (step II-300-32). Alternatively, the word combination may be irrelevant (noise) to the framework structure II-120.

A final yet important aspect of the disclosed invention relates to the framework structure II-120 used to expand the skeletal structure II-110. Notably, changes to the framework structure II-110 will result in corresponding changes throughout the expanded skeletal structure.

For example, if a change is made in the folder definition II-128 within the framework structure II-120 (FIG. II-2B), the change is dynamically reflected in the corresponding framework folders II-122 within the expanded skeletal structure II-110 (FIG. II-11).

Similarly, if a new framework folder II-122 is added to the framework structure II-120, then the change is dynamically reflected in each of the places where the framework structure II-120 was appended.

However, if a change is made to a framework folder II-122 within the expanded skeletal structure II-110, the change is not dynamically reflected back to the framework structure II-120 or to any of the corresponding framework folders II-122 within the expanded skeletal structure II-110.

Moreover, modification of a folder definition II-128 within the framework structure II-120 will not over-ride the local changes to the folder definition II-128 within the expanded skeletal structure II-110.

While the invention has been described with reference to certain preferred embodiments, as will apparent to those of ordinary skill in the art, certain changes and modifications can be made without departing from the scope of the invention as defined by the following claims.

We Claim:

1. A systematic method for creating framework folders used to expanding a skeletal structure, comprising the steps of:

collect the folder label for each individual first level skeletal folder and the folder labels of all hierarchically subordinate skeletal folders into separate collections;

remove predefined noise words from each collection of folder labels;

tabulate a separate frequency table for each collection, counting the single word frequency of each word a given collection of folder labels;

remove words from each frequency table whose frequency falls below a predetermined threshold;

combine the individual frequency tables into a combined frequency table;

output the results of the combined frequency table, wherein a directory editor extrapolates concepts from the results of the combined frequency table and creates a new framework folder for each extrapolated concept.

2. A method for optimizing a framework structure, comprising the steps of:

append an unmatched folder to the framework structure;

map a collection of paragraphs to the framework structure;

compile a frequency table of one, two, three and four words combinations from the paragraphs mapped to the unmatched folder;

remove noise combinations from the frequency table; and

output the results of the combined frequency table, wherein a directory editor does one of:

extrapolates concepts from the results of the frequency table and creates a new framework folder for each extrapolated concept; and

optimizes the framework folder definition(s) to detect the concept conveyed in the paragraphs mapped to the unmatched folder.

3. A method for systematically expanding a skeletal structure:

creating a framework structure from the folder labels of the skeletal structure; and

appending a copy of the framework structure to each skeletal end folder.

4. The method according to claim 3 further comprising the steps of:

mapping a collection of paragraphs to the expanded skeletal structure;

tabulating a number of paragraphs mapped to each end-folder of the expanded skeletal structure; and

deleting a selected end-folder if the number of paragraphs mapped to the selected end-folder is below a predetermined threshold.

5. The method according to claim 4 further comprising the steps of:

mapping a collection of paragraphs to the expanded skeletal structure;

tabulating a number of paragraphs mapped to each end-folder of the expanded skeletal structure;

flagging a selected end-folder if the number of paragraphs mapped to the selected end-folder is above a predetermined threshold;

copy the folder label of each flagged end-folder and redact the copied folder label to remove noise words;

for each of the paragraphs mapped to a flagged end-folder, extract sentences which contain the redacted folder label;

tabulate a frequency table one, two, three and four word combinations that re-occur in the extracted sentences;

remove predefined noise combinations from the frequency table;

retain a predetermined number of the most highest frequency word combinations; and

create an expansion folder for each retained word combination.

6. A method for optimizing a skeletal directory structure, comprising:

append an unmatched folder to the skeletal structure;

map a collection of paragraphs to the skeletal structure;

compile a frequency table of one, two, three and four words combinations from the paragraphs mapped to the unmatched folder;

remove noise combinations from the frequency table; and

output the results of the combined frequency table, wherein a directory editor extrapolates concepts from the results of the frequency table, if the extrapolated concept does not correspond to the label of an existing folder then create a new framework folder for the extrapolated concept(s), otherwise the directory editor optimizes the framework folder definition(s) to detect paragraphs mapped to the unmatched folder.

7. A method for compiling word combinations indicative of concepts for inclusion in a framework structure from the folder labels of a skeletal strcuture:

collect the folder label for each individual first level skeletal folder and the folder labels of all hierarchically subordinate skeletal folders into separate collections;

remove predefined noise words from each collection of folder labels;

tabulate a separate frequency table for each collection, counting the single word frequency of each word a given collection of folder labels;

remove words from each frequency table whose frequency falls below a predetermined threshold;

combine the individual frequency tables into a combined frequency table; and

output the results of the combined frequency table, wherein the combinations in the combined frequency table are indicative of concepts which should be included within the framework structure.

FIG. I-1

Business & Economy
B2B, Finance, Shopping, Jobs...

Computers & Internet
Internet, WWW, Software, Games...

News & Media
Newspapers, TV, Radio...

Entertainment
Movies, Humor, Music...

Recreation & Sports
Sports, Travel, Autos, Outdoors...

Health
Diseases, Drugs, Fitness, Medicine...

Government
Elections, Military, Law, Taxes...

Regional
Countries, Regions, US States...

Society & Culture
People, Environment, Religion...

Education
College and University, K-12...

Arts & Humanities
Photography, History, Literature...

Science
Animals, Astronomy, Engineering...

Social Science
Languages, Archaeology, Psychology...

Reference
Phone Numbers, Dictionaries, Quotations...

Fig. I - 2

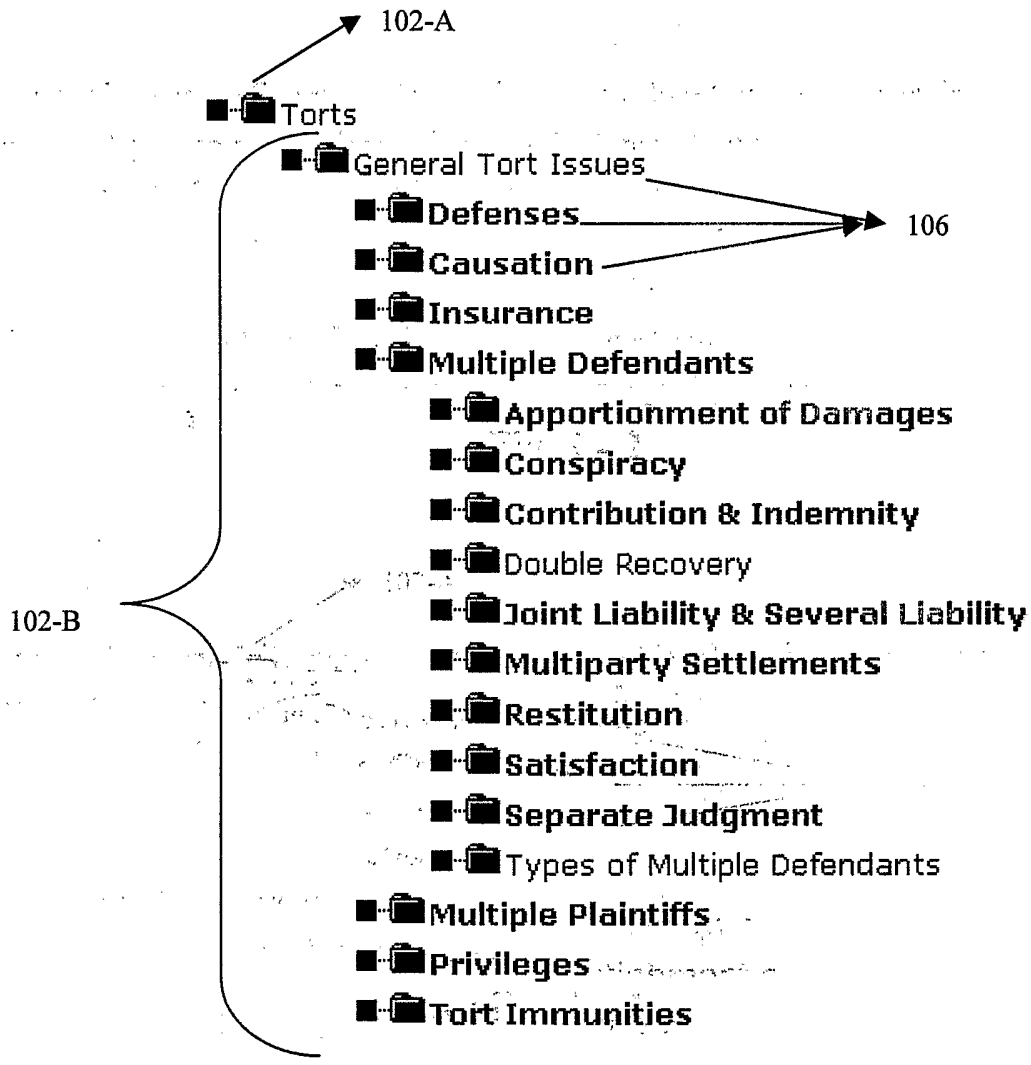

102-A
Torts
General Tort Issues
Defenses
Causation
106
Insurance
Multiple Defendants
Apportionment of Damages
Conspiracy
Contribution & Indemnity
Double Recovery
102-B
Joint Liability & Several Liability
Multiparty Settlements
Restitution
Satisfaction
Separate Judgment
Types of Multiple Defendants
Multiple Plaintiffs
Privileges
Tort Immunities

FIG. I - 3

brochure-|flyer-|hand-out-|bulletin-|make public-|magazine-|novel-|publicat-|dispatch-|journal-|newsletter-|slander-|leaflet-|the scoop-|portray-|circulat-|$advert$|statement-|publish-|disseminat-|advertis-|periodical-|adverts-|disclos-|newspaper-|broadcast-|article-|handout

FIG. I - 4

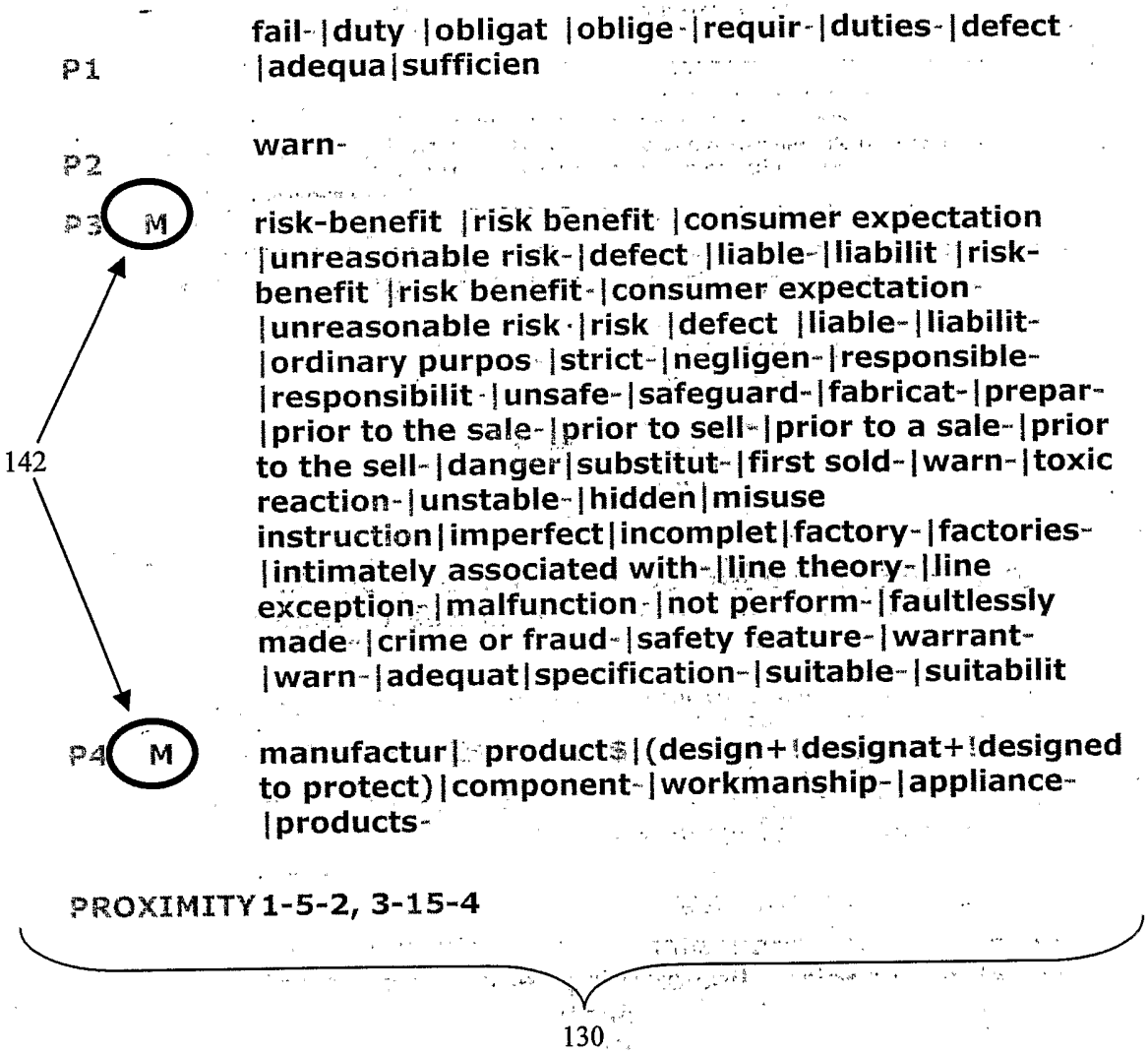
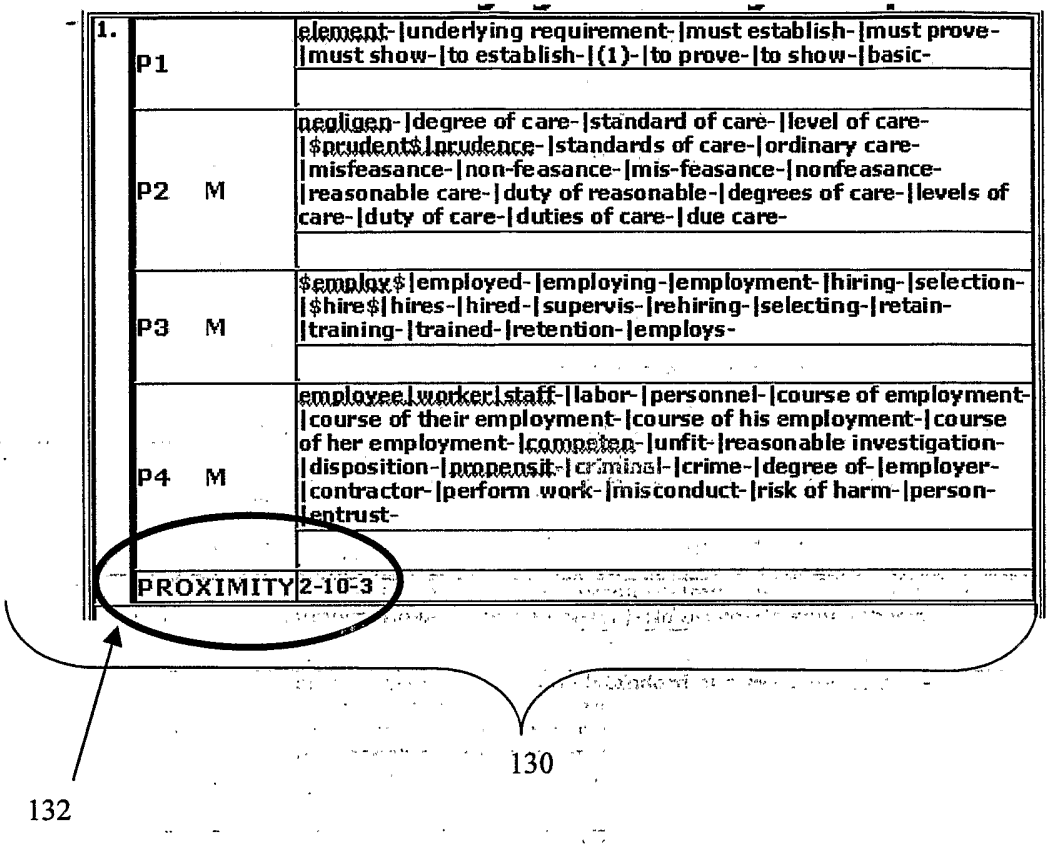

P1
fail-|duty |obligat |oblige-|requir-|duties-|defect |adequa|sufficien
P2
warn-
P3
M
risk-benefit |risk benefit |consumer expectation |unreasonable risk-|defect |liable-|liabilit |risk-benefit |risk benefit-|consumer expectation |unreasonable risk-|risk |defect |liable-|liabilit-|ordinary purpos |strict-|negligen-|responsible-|responsibilit-|unsafe-|safeguard-|fabricat-|prepar-|prior to the sale-|prior to sell-|prior to a sale-|prior to the sell-|danger|substitut-|first sold-|warn-|toxic reaction-|unstable-|hidden|misuse instruction|imperfect|incomplet|factory-|factories-|intimately associated with-|line theory-|line exception-|malfunction-|not perform-|faultlessly made-|crime or fraud-|safety feature-|warrant-|warn-|adequat|specification-|suitable-|suitabilit
142
P4
M
manufactur|product$|(design+!designat+!designed to protect)|component-|workmanship-|appliance-|products-
PROXIMITY 1-5-2, 3-15-4
130

FIG. I - 5



1.		
	P1	element-|underlying requirement-|must establish-|must prove-|must show-|to establish-|(1)-|to prove-|to show-|basic-
	P2 M	negligen-|degree of care-|standard of care-|level of care-|$prudent$|prudence-|standards of care-|ordinary care-|misfeasance-|non-feasance-|mis-feasance-|nonfeasance-|reasonable care-|duty of reasonable-|degrees of care-|levels of care-|duty of care-|duties of care-|due care-
	P3 M	$employ$|employed-|employing-|employment-|hiring-|selection-|$hire$|hires-|hired-|supervis-|rehiring-|selecting-|retain-|training-|trained-|retention-|employs-
	P4 M	employee|worker|staff-|labor-|personnel-|course of employment-|course of their employment-|course of his employment-|course of her employment-|competen-|unfit-|reasonable investigation-|disposition-|propensit-|criminal-|crime-|degree of-|employer-|contractor-|perform work-|misconduct-|risk of harm-|person-|entrust-
	PROXIMITY	2-10-3

130

132

FIG. I - 6

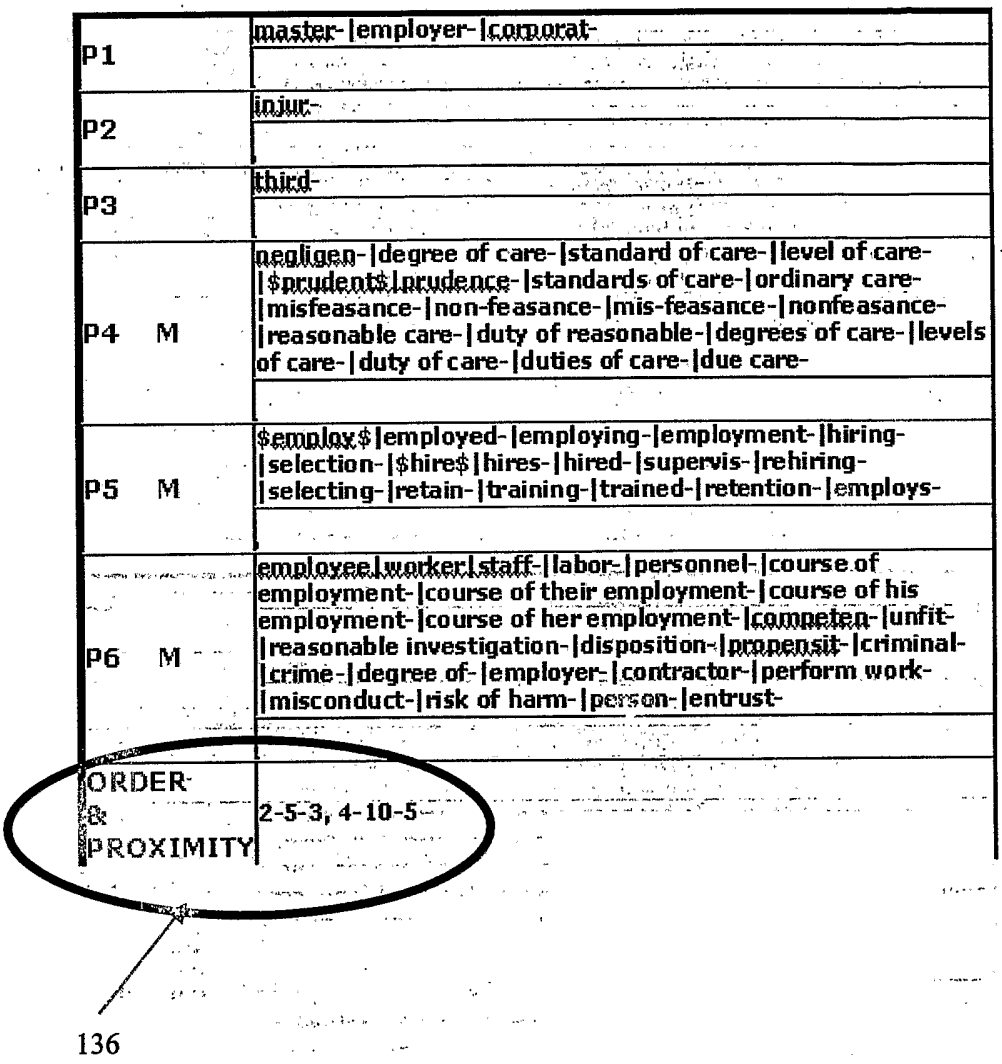
In general, the theory of negligent hiring is that "[a]n employer whose employees are brought into contact with members of the public in the course of their employment is responsible for exercising a duty of reasonable care in the selection . . . of its employees." Di Cosala v. Kay, 91 N.J. 159, 450 A.2d 508 (1982). Therefore, if an employer negligently hires an individual who is incompetent or unfit for the job, and knew or should have known through a reasonable investigation that the employee was unfit, the employer may be liable to third parties whose injuries were proximately caused by the employer's negligence

FIG. I - 7

P1	master-|employer-|corporat-
P2	injur-
P3	third-
P4 M	negligen-|degree of care-|standard of care-|level of care-|$prudent$|prudence-|standards of care-|ordinary care-|misfeasance-|non-feasance-|mis-feasance-|nonfeasance-|reasonable care-|duty of reasonable-|degrees of care-|levels of care-|duty of care-|duties of care-|due care-
P5 M	$employ$|employed-|employing-|employment-|hiring-|selection-|$hire$|hires-|hired-|supervis-|rehiring-|selecting-|retain-|training-|trained-|retention-|employs-
P6 M	employee|worker|staff-|labor-|personnel-|course of employment-|course of their employment-|course of his employment-|course of her employment-|competen-|unfit-|reasonable investigation-|disposition-|propensit-|criminal-|crime-|degree of-|employer-|contractor-|perform work-|misconduct-|risk of harm-|person-|entrust-
PROXIMITY ORDER	2-1-3, 4-5-0-5

132

FIG. I - 8



P1	master-|employer-|comporat-
P2	injur-
P3	third-
P4 M	negligen-|degree of care-|standard of care-|level of care-|$prudent$|prudence-|standards of care-|ordinary care-|misfeasance-|non-feasance-|mis-feasance-|nonfeasance-|reasonable care-|duty of reasonable-|degrees of care-|levels of care-|duty of care-|duties of care-|due care-
P5 M	$employ$|employed-|employing-|employment-|hiring-|selection-|$hire$|hires-|hired-|supervis-|rehiring-|selecting-|retain-|training-|trained-|retention-|employs-
P6 M	employee|worker|staff-|labor-|personnel-|course of employment-|course of their employment-|course of his employment-|course of her employment-|competen-|unfit-|reasonable investigation-|disposition-|propensit-|criminal-|crime-|degree of-|employer-|contractor-|perform work-|misconduct-|risk of harm-|person-|entrust-
ORDER & PROXIMITY	2-5-3, 4-10-5

136

FIG. I - 9

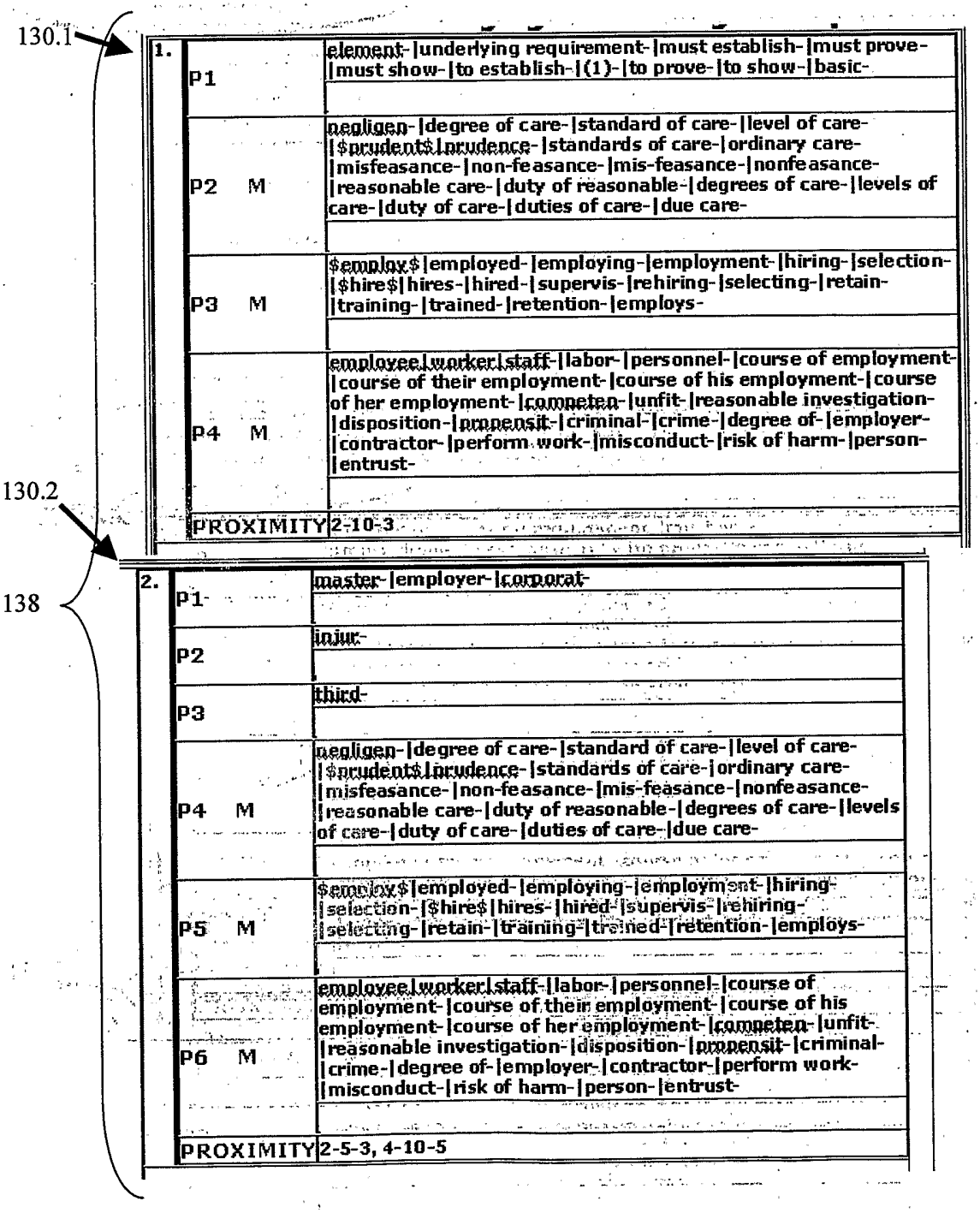

130.1
130.2
138
1.
P1
element-|underlying requirement-|must establish-|must prove-|must show-|to establish-|(1)-|to prove-|to show-|basic-
P2 M
negligen-|degree of care-|standard of care-|level of care-|$prudent$|prudence-|standards of care-|ordinary care-|misfeasance-|non-feasance-|mis-feasance-|nonfeasance-|reasonable care-|duty of reasonable-|degrees of care-|levels of care-|duty of care-|duties of care-|due care-
P3 M
$employ$|employed-|employing-|employment-|hiring-|selection-|$hire$|hires-|hired-|supervis-|rehiring-|selecting-|retain-|training-|trained-|retention-|employs-
P4 M
employee|worker|staff-|labor-|personnel-|course of employment-|course of their employment-|course of his employment-|course of her employment-|competen-|unfit-|reasonable investigation-|disposition-|propensit-|criminal-|crime-|degree of-|employer-|contractor-|perform work-|misconduct-|risk of harm-|person-|entrust-
PROXIMITY 2-10-3
2.
P1-
master-|employer-|corporat-
P2
injur-
P3
third-
P4 M
negligen-|degree of care-|standard of care-|level of care-|$prudent$|prudence-|standards of care-|ordinary care-|misfeasance-|non-feasance-|mis-feasance-|nonfeasance-|reasonable care-|duty of reasonable-|degrees of care-|levels of care-|duty of care-|duties of care-|due care-
P5 M
$employ$|employed-|employing-|employment-|hiring-|selection-|$hire$|hires-|hired-|supervis-|rehiring-|selecting-|retain-|training-|trained-|retention-|employs-
P6 M
employee|worker|staff-|labor-|personnel-|course of employment-|course of their employment-|course of his employment-|course of her employment-|competen-|unfit-|reasonable investigation-|disposition-|propensit-|criminal-|crime-|degree of-|employer-|contractor-|perform work-|misconduct-|risk of harm-|person-|entrust-
PROXIMITY 2-5-3, 4-10-5

FIG. I - 10

P1		element-|underlying requirement-|must establish-|must prove-|must show-|to establish-|(1)-|to prove-|to show-|basic-
P2	M	negligen-|degree of care-|standard of care-|level of care-|prudent-|prudence-|standards of care-|ordinary care-|misfeasance-|non-feasance-|mis-feasance-|nonfeasance-|reasonable care-|duty of reasonable-|degrees of care-|levels of care-|duty of care-|duties of care-|due care-
P3	M	$employ$|employed-|employing-|employment-|hiring-|selection-|hire-|hire-|hired-|supervis-|rehiring-|selecting-|retain-|training-|trained-|retention-|employs-
P4	M	employee|worker|staff-|labor-|personnel-|course of employment-|course of their employment-|course of his employment-|course of her employment-|competen-|unfit-|reasonable investigation-|disposition-|propensit-|criminal-|crime-|degree of-|employer-|contractor-|perform work-|misconduct-|risk of harm-|person-|entrust-
Same Sent		2-10-3
NOT1		Hiring
NOT2		Prudential | Prudentially

FIG. I – 11A-1

Torts >>> Types of Torts >>> Employment Torts >>> *Negligent Hiring and Supervision*

P1	M	negligen-|degree of care-|standard of care-|level of care- |$prudent$|prudence-|standards of care-|ordinary care-|misfeasance-|non-feasance-|mis-feasance-|nonfeasance-|reasonable care-|duty of reasonable-|degrees of care-|levels of care-|duty of care-|duties of care-|due care-
P2	M	$employ$|employed-|employing-|employment-|hiring-|selection- |$hire$|hires-|hired-|supervis-|rehiring-|selecting-|retain-|training-|trained-|retention-|employs-
P3	M	employee|worker|staff-|labor-|personnel-|course of employment-|course of their employment-|course of his employment-|course of her employment-|competen-|unfit-|reasonable investigation-|disposition-|propensit-|criminal-|crime-|degree of-|employer-|contractor-|perform work-|misconduct-|risk of harm-|person-|entrust-
proximity		1-10-2

Torts >>> Types of Torts >>> Employment Torts >>> Negligent Hiring and Supervision >>> *Damages*

A		indemnification-|recovery-|economic relief-|financial remed-|Compensation-|compensat-|indemnity-|reparation-|indemnify-|reimburs-|damage-|injur-|harm-|loss-|Remedy-|remedies-|Judicial Sanction-|remed-|injunction-|relief-|remed-|relief-|financial relief-|monetary remed-|monetary relief-|monetary award-|monetary award-|injunctive-
B		measure-|special-|amount-|valuation-|account-|pay-|paid-|responsible-|award-|grant-|ground-|responsibilit-|subject-|determin-|fix-|amount-|measur-|apprais-|value-|valua-|decid-|decision-|determin-
P1	M	negligen-|degree of care-|standard of care-|level of care- |$prudent$|prudence-|standards of care-|ordinary care-|misfeasance-|non-feasance-|mis-feasance-|nonfeasance-|reasonable care-|duty of reasonable-|degrees of care-|levels of care-|duty of care-|duties of care-|due care-
P2	M	$employ$|employed-|employing-|employment-|hiring-|selection- |$hire$|hires-|hired-|supervis-|rehiring-|selecting-|retain-|training-|trained-|retention-|employs-
P3	M	employee|worker|staff-|labor-|personnel-|course of employment-|course of their employment-|course of his employment-|course of her employment-|competen-|unfit-|reasonable investigation-|disposition-|propensit-|criminal-|crime-|degree of-|employer-|contractor-|perform work-|misconduct-|risk of harm-|person-|entrust-
PROXIMITY		A-10-B, P1-10-P2

FIG. I - 11A-2

Torts >>> **Types of Torts** >>> **Employment Torts** >>> **Negligent Hiring and Supervision** >>> **Damages** >>> ***Punitive Damages***

C	punitive-|disciplinary-|exemplary
D	damages-|award-|relief-|remed-|recover-|function-
E	amount-|liable-|pay-|responsible-|responsibilit-|subject-|value-|valuat-|measure-|special-|account-|liabilit-|paid-|grant-|ground-|valuation-|award-
P1 M	negligen-|degree of care-|standard of care-|level of care-|$prudent$|prudence-|standards of care-|ordinary care-|misfeasance-|non-feasance-|mis-feasance-|nonfeasance-|reasonable care-|duty of reasonable-|degrees of care-|levels of care-|duty of care-|duties of care-|due care-
P2 M	$employ$|employed-|employing-|employment-|hiring-|selection-|$hire$|hires-|hired-|supervis-|rehiring-|selecting-|retain-|training-|trained-|retention-|employs-
P3 M	employee|worker|staff-|labor-|personnel-|course of employment-|course of their employment-|course of his employment-|course of her employment-|competen-|unfit-|reasonable investigation-|disposition-|propensit-|criminal-|crime-|degree of-|employer-|contractor-|perform work-|misconduct-|risk of harm-|person-|entrust-
PROXIMITY	C-10-D, D-10-E, P1-10-P2

FIG. I - 11B

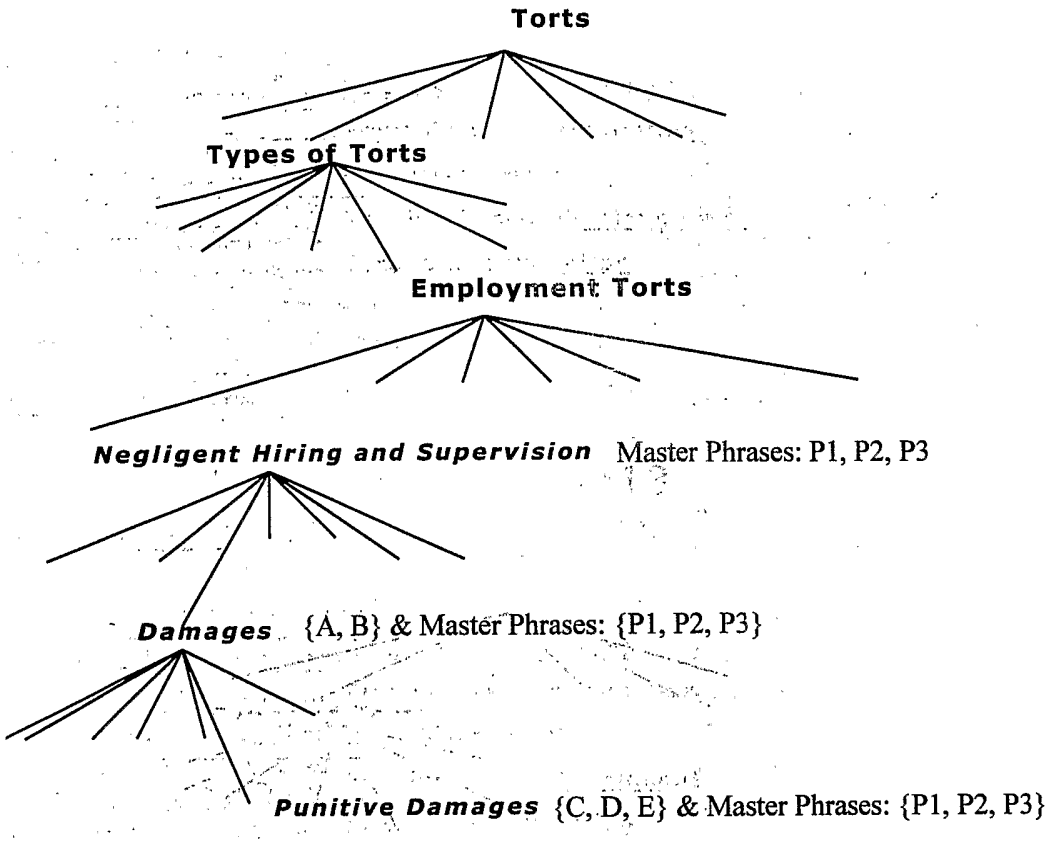

Torts
Types of Torts
Employment Torts
Negligent Hiring and Supervision
Master Phrases: P1, P2, P3
Damages
{A, B} & Master Phrases: {P1, P2, P3}
Punitive Damages
{C, D, E} & Master Phrases: {P1, P2, P3}

FIG. I - 12

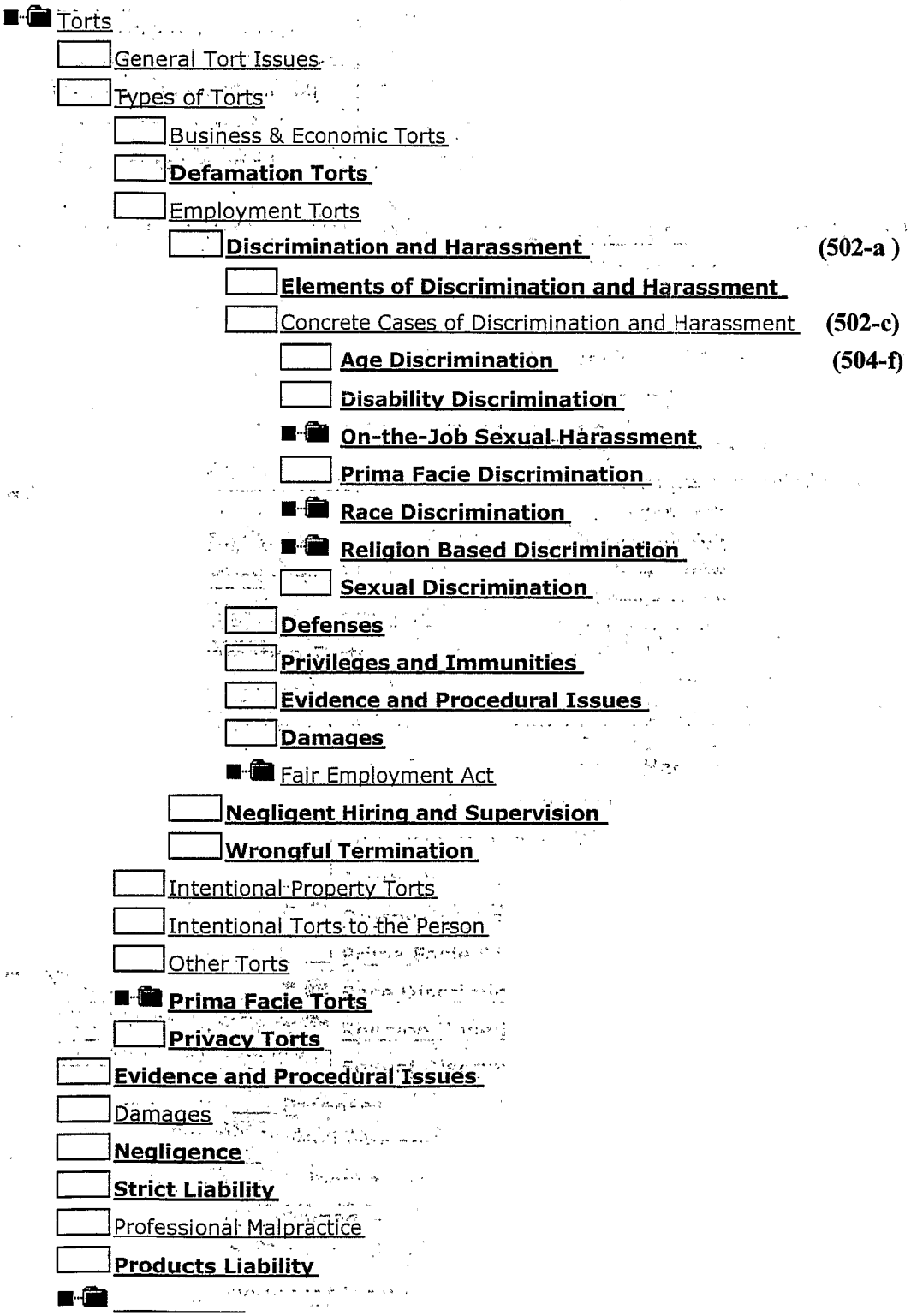

Torts
General Tort Issues
Types of Torts
Business & Economic Torts
Defamation Torts
Employment Torts
Discrimination and Harassment
(502-a)
Elements of Discrimination and Harassment
Concrete Cases of Discrimination and Harassment
(502-c)
Age Discrimination
(504-f)
Disability Discrimination
On-the-Job Sexual Harassment
Prima Facie Discrimination
Race Discrimination
Religion Based Discrimination
Sexual Discrimination
Defenses
Privileges and Immunities
Evidence and Procedural Issues
Damages
Fair Employment Act
Negligent Hiring and Supervision
Wrongful Termination
Intentional Property Torts
Intentional Torts to the Person
Other Torts
Prima Facie Torts
Privacy Torts
Evidence and Procedural Issues
Damages
Negligence
Strict Liability
Professional Malpractice
Products Liability

FIG. I - 13

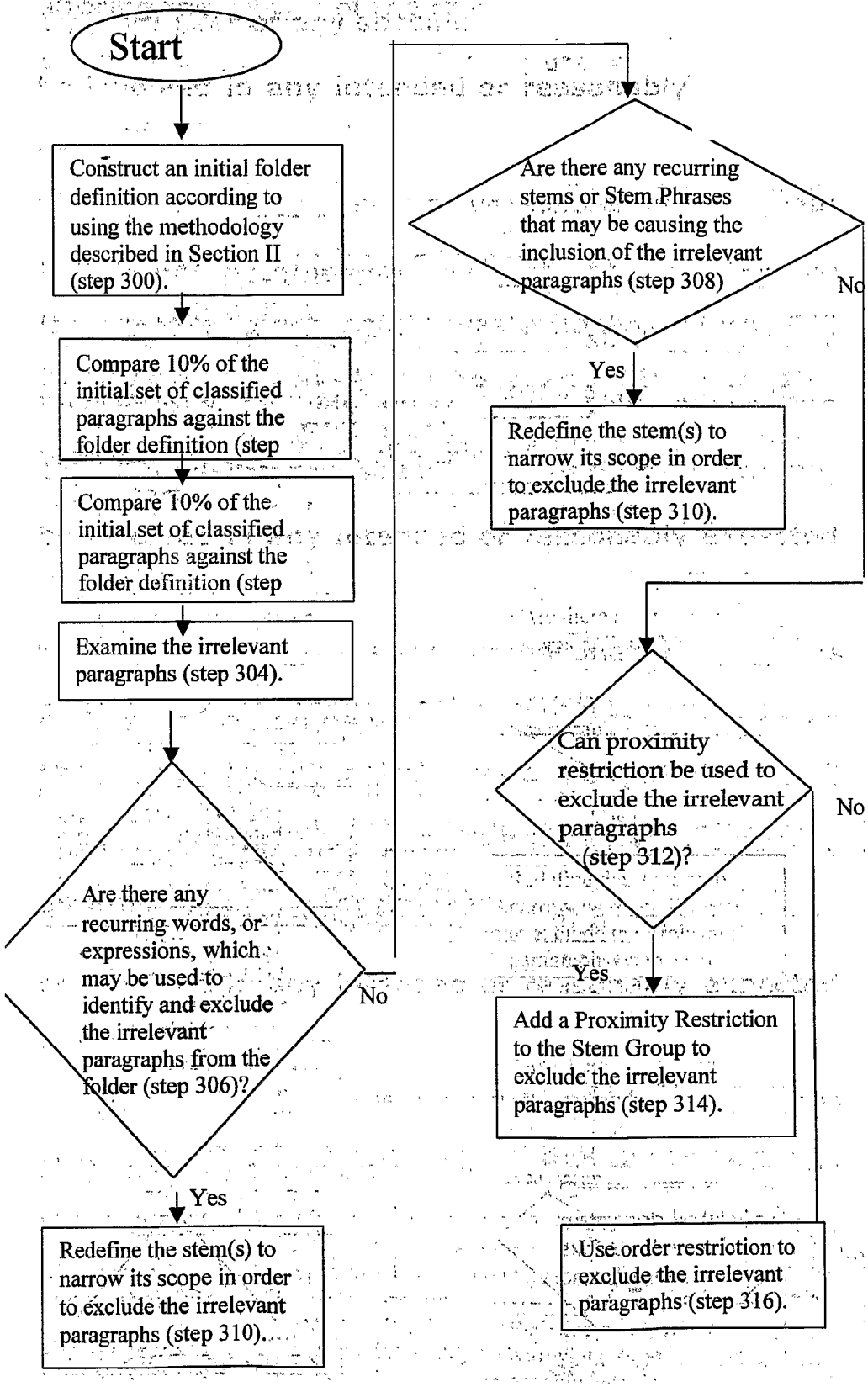

Start
Construct an initial folder definition according to using the methodology described in Section II (step 300).
Compare 10% of the initial set of classified paragraphs against the folder definition (step
Compare 10% of the initial set of classified paragraphs against the folder definition (step
Examine the irrelevant paragraphs (step 304).
Are there any recurring words, or expressions, which may be used to identify and exclude the irrelevant paragraphs from the folder (step 306)?
No
Yes
Redefine the stem(s) to narrow its scope in order to exclude the irrelevant paragraphs (step 310).
Are there any recurring stems or Stem Phrases that may be causing the inclusion of the irrelevant paragraphs (step 308)
No
Yes
Redefine the stem(s) to narrow its scope in order to exclude the irrelevant paragraphs (step 310).
Can proximity restriction be used to exclude the irrelevant paragraphs (step 312)?
No
Yes
Add a Proximity Restriction to the Stem Group to exclude the irrelevant paragraphs (step 314).
Use order restriction to exclude the irrelevant paragraphs (step 316).

FIG. I - 14

EXAMPLE1:

"Thereafter, Lynn Armentrout and his wife Tina Armentrout filed an action against FMC asserting claims of strict liability for failure to warn, strict liability for design defect, negligent failure to warn and negligent design defect. The Armentrouts requested relief under the theory that the existence of the crane's "pinch point" was a hazard which FMC should have Page 179 warned against or removed by altering the design of the crane."

EXAMPLE2:

""A product is defective and unreasonably dangerous if it is not accompanied by sufficient warnings or instructions for use. To be sufficient, such warnings or instructions for use must adequately inform the ordinary user of any specific risk of harm which may be involved in any intended or reasonably expected use."

FIG. I-15

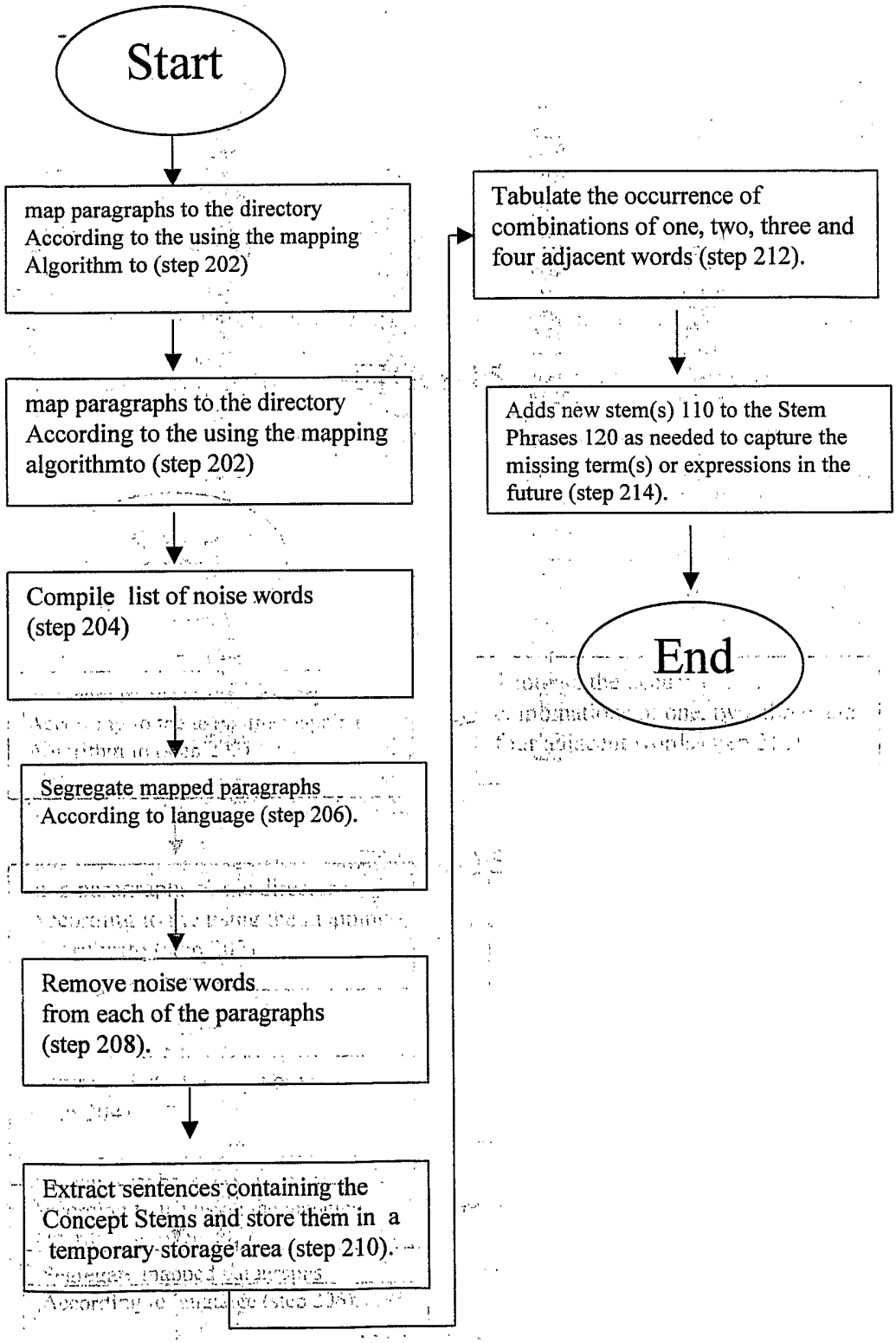

Start
map paragraphs to the directory According to the using the mapping Algorithm to (step 202)
map paragraphs to the directory According to the using the mapping algorithmto (step 202)
Compile list of noise words (step 204)
Segregate mapped paragraphs According to language (step 206).
Remove noise words from each of the paragraphs (step 208).
Extract sentences containing the Concept Stems and store them in a temporary storage area (step 210).
Tabulate the occurrence of combinations of one, two, three and four adjacent words (step 212).
Adds new stem(s) 110 to the Stem Phrases 120 as needed to capture the missing term(s) or expressions in the future (step 214).
End

FIG. I - 16

-	As	in	SECOND	were	WHOM
#	at	into	see	what	WHOSE
$	be	is	SEEK	when	will
%	because	it	SEEM	where	with
&	been	just	shall	whether	within
*	before	KNEW	she	which	without
^	being	KNOWN	should	While	witness
a	between	made	so	who	would
b	both	make	some	WHOM	you
c	but	many	special	WHOSE	your
d	by	may	still	will	
e	came	me	such	with	
f	can	might	take	within	
g	cannot	MILLION	taken	without	
h	could	MONTH	than	witness	
I	did	more	that	would	
j	do	most	the	you	
k	does	much	their	your	
l	each	must	them	were	
n	EITHER	my	then	what	
o	else	NECESSARILY	there	when	
p	EVEN	NEEDLESS	THEREBY	where	
q	for	NEITHER	therefore	whether	
r	FOUR	NEVER	these	which	
s	FROM	NEVERTHELESS	they	While	
t	further	new	this	who	
u	get	no	those	WHOM	
v	got	NO.	THOUGH	WHOSE	
w	had	Nor	THREE	will	
x	has	not	through	with	
y	has been	NOW	Thus	within	
z	have	OBEY	to	without	
about	he	of	too	witness	
after	her	OFTEN	under	would	
against	here	on	UNTIL	you	
All	HERE.	on the	up	your	
also	him	one	upon	were	
ALTHOUGH	HIM.	only	very	what	
an	HIMSELF	or	want	when	
and	his	other	was	where	
another	how	our	way	whether	
any	however	out	we	which	
are	if	right	well	While	
As				who	
at					

FIG. I - 17

"The rationale for limiting manufacturer liability under the facts of this case is similar to that set forth by those jurisdictions which have precluded a DUTY TO WARN, for strict liability purposes, where the existence of a dangerous condition is generally known and recognized."

"Because, based on the evidence adduced at trial, the jury might have concluded that Montgomery Ward had the DUTY TO WARN the buyers, and its failure to do so was the proximate cause of the accident resulting in injury, the trial court's refusal to instruct the jury on this theory of the case was reversible error."

"Accordingly, the district court held that the defendant had no DUTY TO WARN of such inherent dangers and risks of skiing under section 33-44-107(2)(d), and that the plaintiff's claims were barred by section 33-44-112, which prohibits claims against ski area operators for injuries resulting from such inherent dangers and risks."

"A manufacturer's DUTY TO WARN encompasses both the duty to provide adequate instructions for safe use and the duty to provide warnings as to dangers inherent in improper use."

"It is undisputed that the label on the product at issue complies with the FHSA, and we conclude that the FHSA preempts plaintiff's DUTY TO WARN claims."

"Moreover, the purposes underlying both legislative regulation of labeling and a state common law DUTY TO WARN are the same."

"Therefore, we find it only logical to hold that the common law DUTY TO WARN is subjected to the same federal preemptive constraints as a state statute."

FIG. I - 18

Multiple Words	Freq
Duty To Warn	56
Duty To Warn Cases	3
Common Law Duty	3
Warn Of Dangers	3
Warn Cases	3
Warn Of These Dangers	3
Warn P No Duty	2
Warn Plaintiffs	2
Warn The User	2
Warn Theory	2
Warn Under The Knowledge	2
Warn Exists	2
Warn Of Danger	2
Warn Of Its Danger	2
Admissible Because Duty	2
Bb Gun Because Danger	2
Breach Of This Duty	2
Duty To Warn Exists	2
Duty To Warn Plaintiffs	2
Gun Because Danger	2
Dangers And A Breach	2
Dangers Of Bb Gun	2
Danger Is Obvious	2
Danger Of No Duty	2
Knowledge Or Duty	2

FIG. II - 1

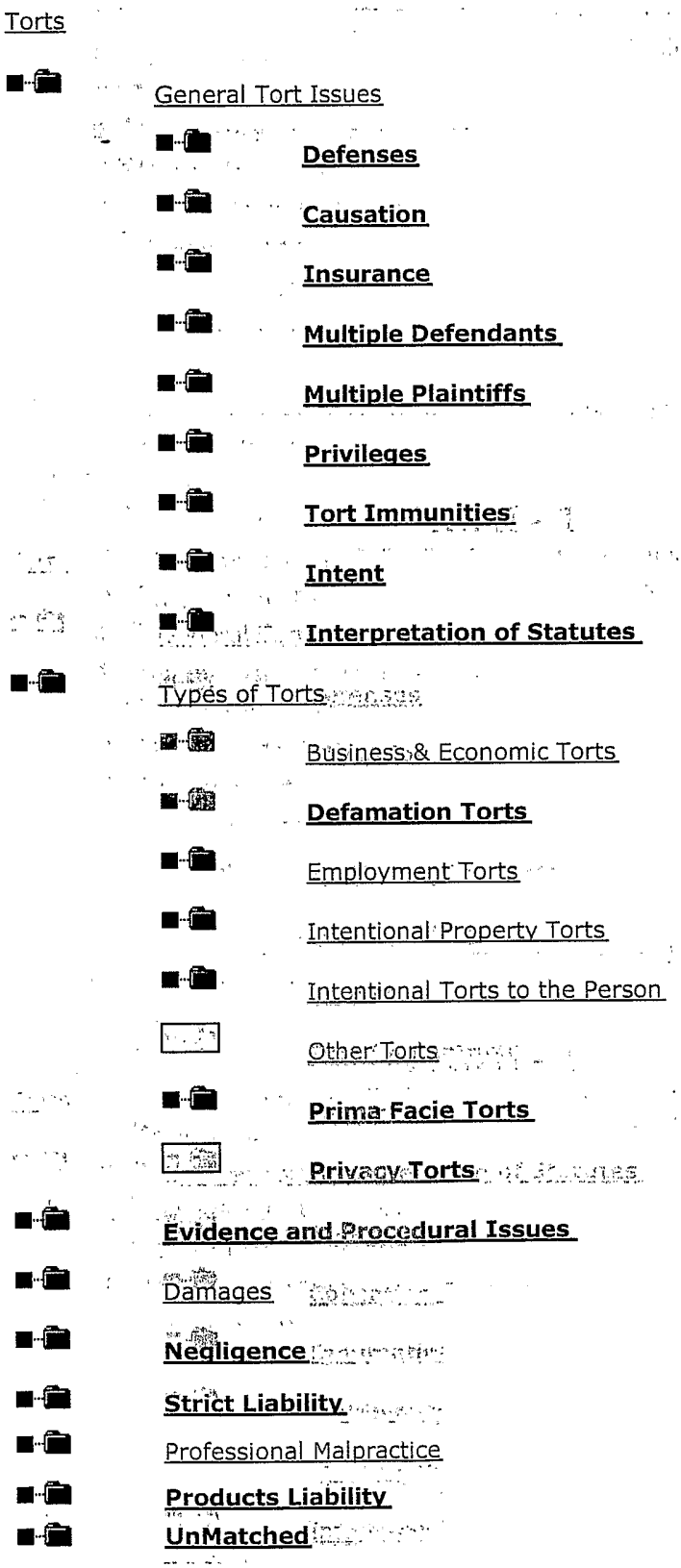

Torts
General Tort Issues
Defenses
Causation
Insurance
Multiple Defendants
Multiple Plaintiffs
Privileges
Tort Immunities
Intent
Interpretation of Statutes
Types of Torts
Business & Economic Torts
Defamation Torts
Employment Torts
Intentional Property Torts
Intentional Torts to the Person
Other Torts
Prima Facie Torts
Privacy Torts
Evidence and Procedural Issues
Damages
Negligence
Strict Liability
Professional Malpractice
Products Liability
UnMatched

FIG. II - 2A

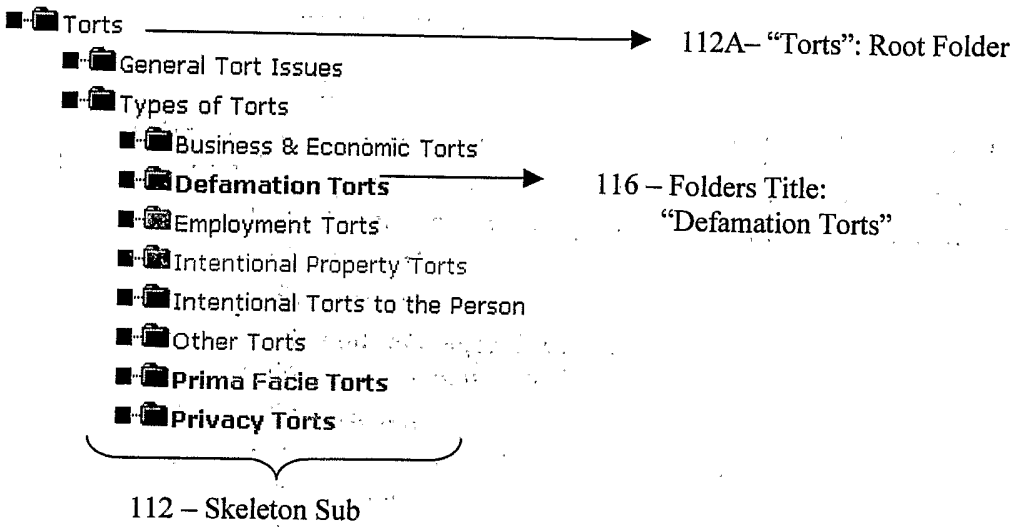

Torts
General Tort Issues
Types of Torts
Business & Economic Torts
Defamation Torts
Employment Torts
Intentional Property Torts
Intentional Torts to the Person
Other Torts
Prima Facie Torts
Privacy Torts
112A– "Torts": Root Folder
116 – Folders Title: "Defamation Torts"
112 – Skeleton Sub

FIG. II - 2B

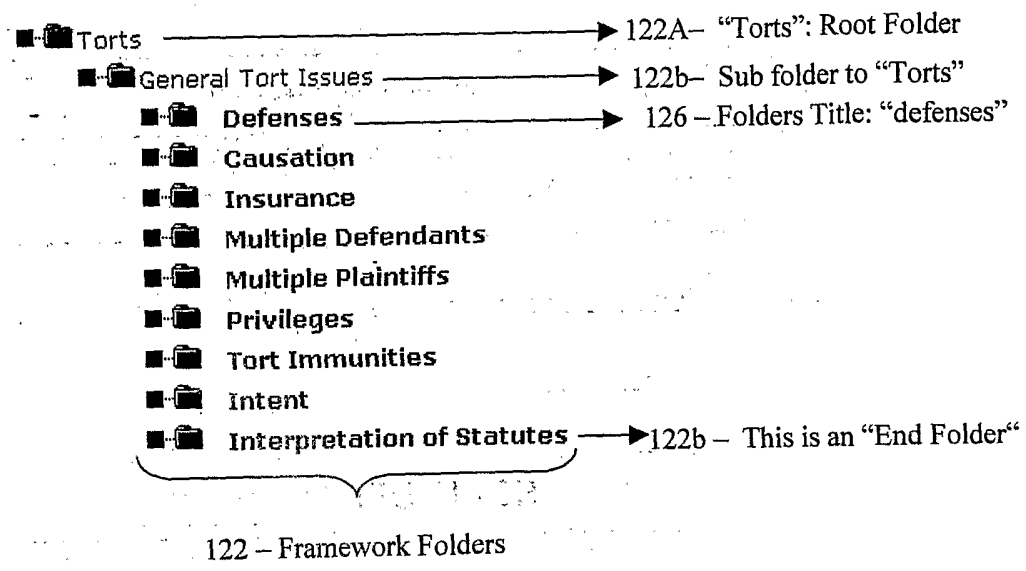

Torts
122A– "Torts": Root Folder
General Tort Issues
122b– Sub folder to "Torts"
Defenses
126 – Folders Title: "defenses"
Causation
Insurance
Multiple Defendants
Multiple Plaintiffs
Privileges
Tort Immunities
Intent
Interpretation of Statutes
122b – This is an "End Folder"
122 – Framework Folders

FIG. II - 3

FIG. 3

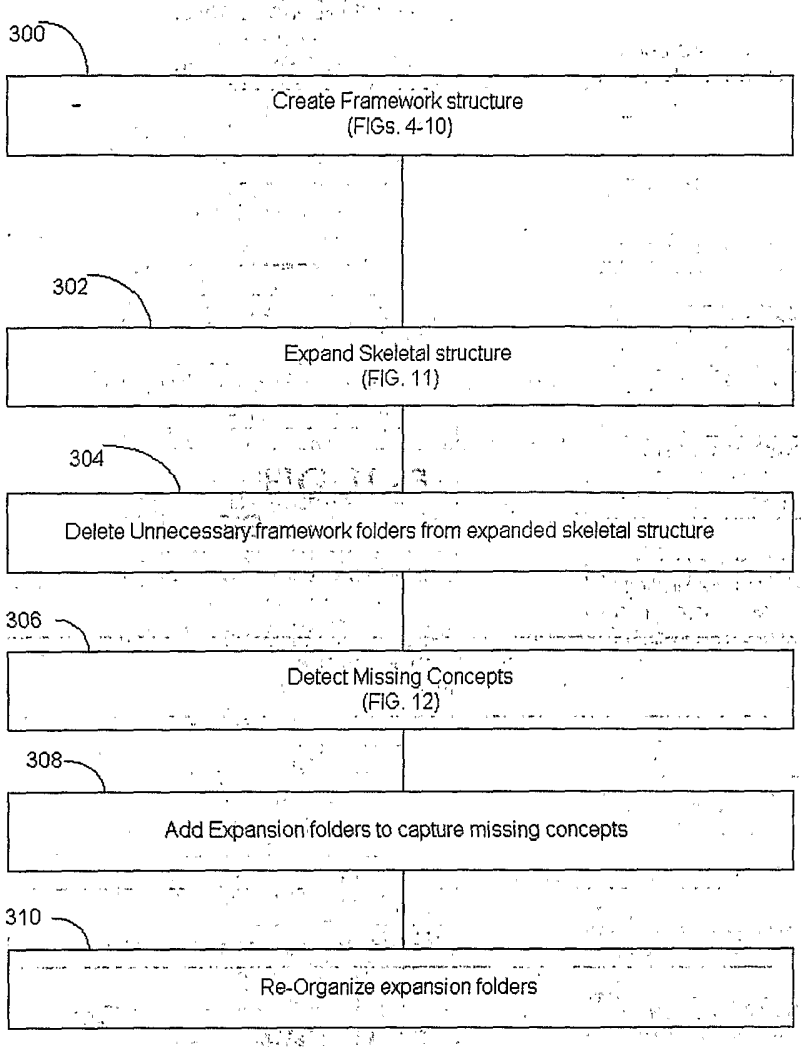

300
Create Framework structure
(FIGs. 4-10)
302
Expand Skeletal structure
(FIG. 11)
304
Delete Unnecessary framework folders from expanded skeletal structure
306
Detect Missing Concepts
(FIG. 12)
308
Add Expansion folders to capture missing concepts
310
Re-Organize expansion folders

FIG. 4

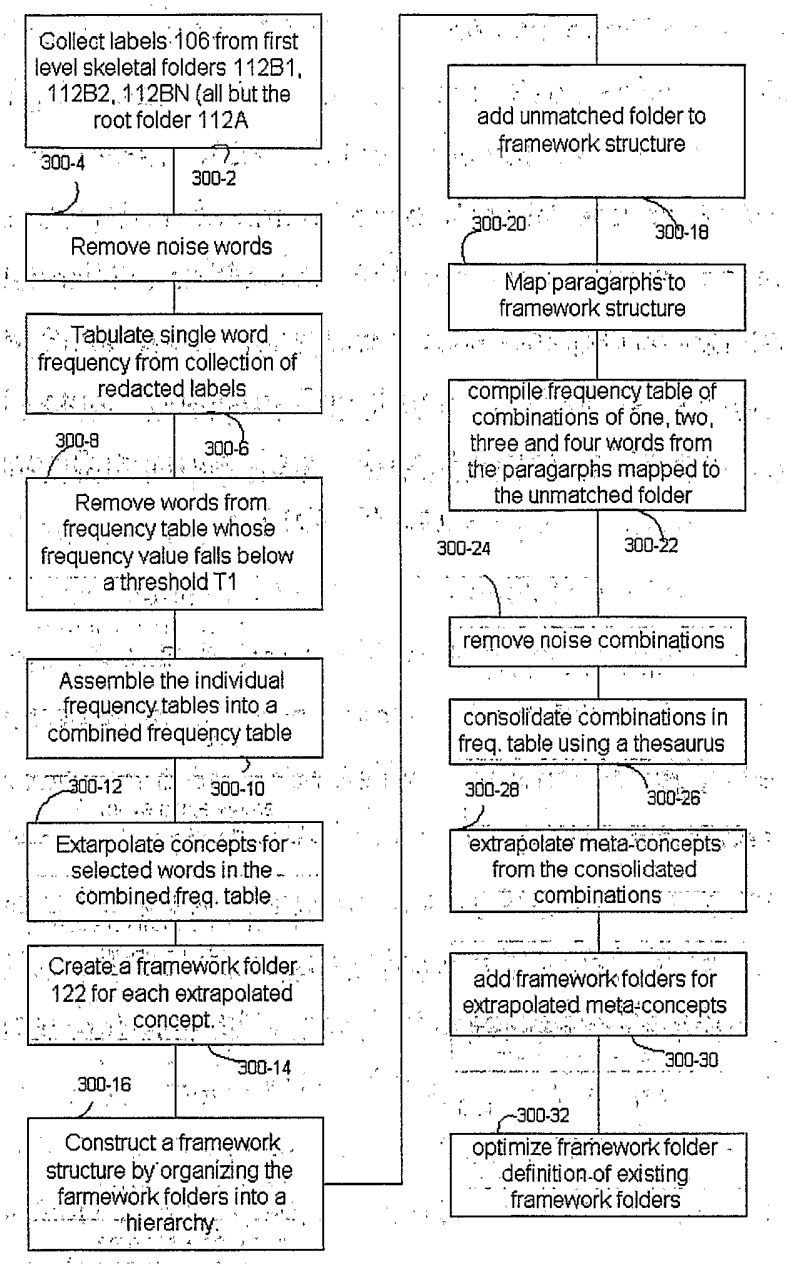

Collect labels 106 from first level skeletal folders 112B1, 112B2, 112BN (all but the root folder 112A
300-4
300-2
Remove noise words
Tabulate single word frequency from collection of redacted labels
300-8
300-6
Remove words from frequency table whose frequency value falls below a threshold T1
Assemble the individual frequency tables into a combined frequency table
300-12
300-10
Extarpolate concepts for selected words in the combined freq. table
Create a framework folder 122 for each extrapolated concept.
300-14
300-16
Construct a framework structure by organizing the farmework folders into a hierarchy.
add unmatched folder to framework structure
300-20
300-18
Map paragarphs to framework structure
compile frequency table of combinations of one, two, three and four words from the paragarphs mapped to the unmatched folder
300-24
300-22
remove noise combinations
consolidate combinations in freq. table using a thesaurus
300-28
300-26
extrapolate meta-concepts from the consolidated combinations
add framework folders for extrapolated meta-concepts
300-30
300-32
optimize framework folder definition of existing framework folders

FIG. II - 5A

Negligence, Strict Liability, Products Liability, Professional Malpractice, Exculpatory Clauses, Assumption of Risk, Contributory Negligence, Comparative Fault, Defense of Consent, Self-Defense, Defense of Truth, Absolute Privileges, Qualified Privileges, Constitutional Privileges, Wrongful Death & Survival, Cause in Fact, Proximate Cause, Failure to Warn, Personal Motive is not Connected, Probability and Common Sense, Preponderance of the Evidence, Expert Opinion, Surrounding Circumstances, Facts Sufficient to Constitute Causation, Instructing the Jury on Factual Causation, The "But For..." Test, Joint Causes/Substantial Factor Test, Alternative Causes

FIG. II – 5B

Direct Threats of Force, Indirect Threats of Force, Failure to Provide Means of Escape, Invalid Use of Legal Authority, Moral Pressure, Future Threats, Amount of Force Allowable, Shopkeeper's Privilege, Privileged Arrest, Felony Arrest Without a Warrant, Misdemeanor Arrests Without a Warrant, Arrests Without a Warrant to Prevent a Crime, Felony Arrest, Misdemeanor Arrest, Reasonable Belief of Theft, Reasonable Manner of Detention, Reasonable Period of Time, Non-Deadly Force, All Directions, Reasonable Means of Escape, Apportionment of Damages, Conspiracy, Contribution & Indemnity, Separate Judgment, Double Recovery, Types of Multiple Defendants, Joint Liability & Several Liability, Multiparty Settlements, Satisfaction

FIG. II - 6

-	As	in	SECOND	were	WHOM
#	at	into	see	what	WHOSE
$	be	is	SEEK	when	will
%	because	it	SEEM	where	with
&	been	just	shall	whether	within
*	before	KNEW	she	which	without
^	being	KNOWN	should	While	witness
a	between	made	so	who	would
b	both	make	some	WHOM	you
c	but	many	special	WHOSE	your
d	by	may	still	will	
e	came	me	such	with	
f	can	might	take	within	
g	cannot	MILLION	taken	without	
h	could	MONTH	than	witness	
I	did	more	that	would	
j	do	most	the	you	
k	does	much	their	your	
l	each	must	them	were	
n	EITHER	my	then	what	
o	else	NECESSARILY	there	when	
p	EVEN	NEEDLESS	THEREBY	where	
q	for	NEITHER	therefore	whether	
r	FOUR	NEVER	these	which	
s	FROM	NEVERTHELESS	they	While	
t	further	new	this	who	
u	get	no	those	WHOM	
v	got	NO.	THOUGH	WHOSE	
w	had	Nor	THREE	will	
x	has	not	through	with	
y	has been	NOW	Thus	within	
z	have	OBEY	to	without	
about	he	of	too	witness	
after	her	OFTEN	under	would	
against	here	on	UNTIL	you	
All	HERE.	on the	up	your	
also	him	one	upon	were	
ALTHOUGH	HIM.	only	very	what	
an	HIMSELF	or	want	when	
and	his	other	was	where	
another	how	our	way	whether	
any	however	out	we	which	
are	if	right	well	While	
As				who	
at					

FIG. II - 7

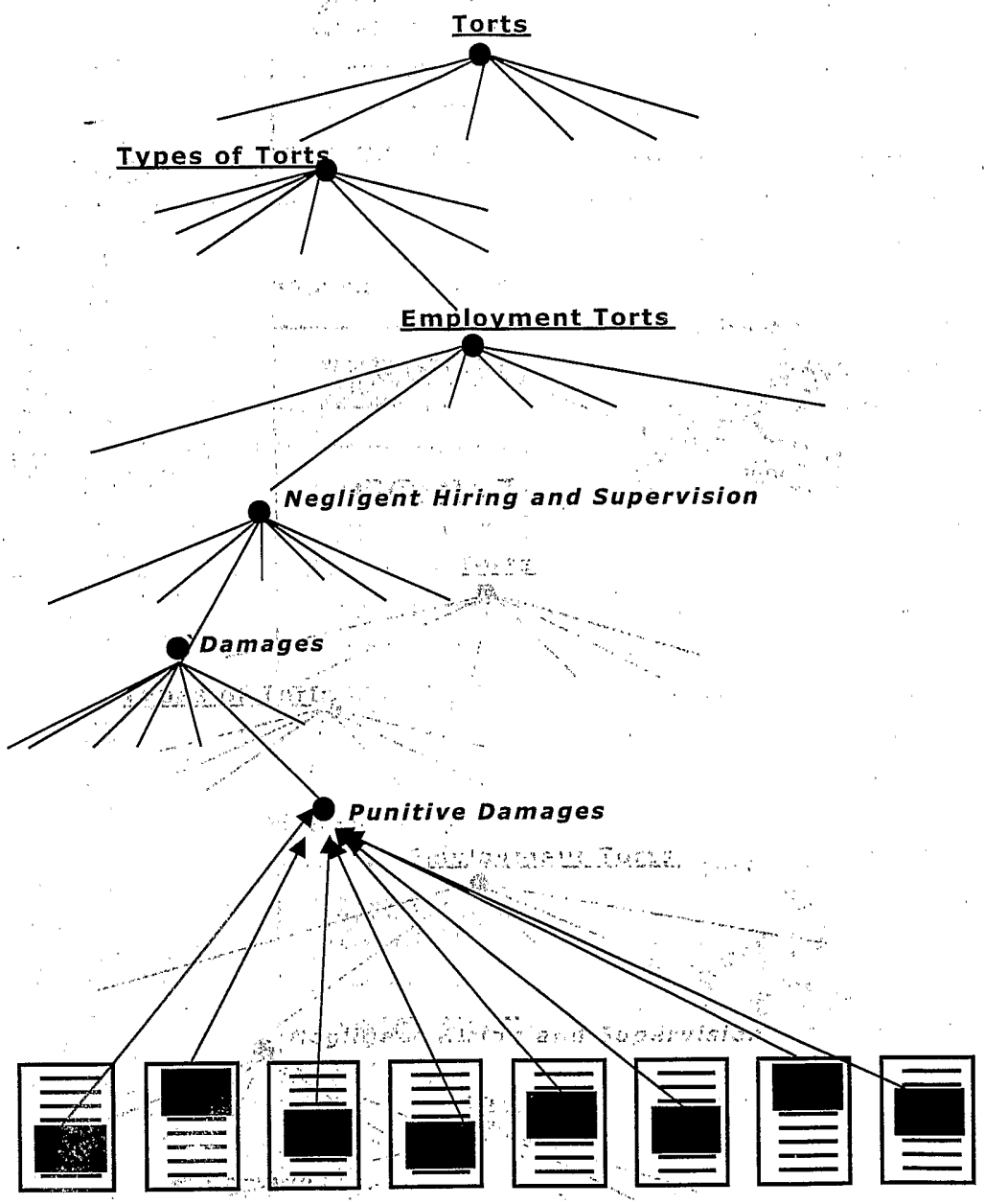

Torts
Types of Torts
Employment Torts
Negligent Hiring and Supervision
Damages
Punitive Damages

FIG. II - 8

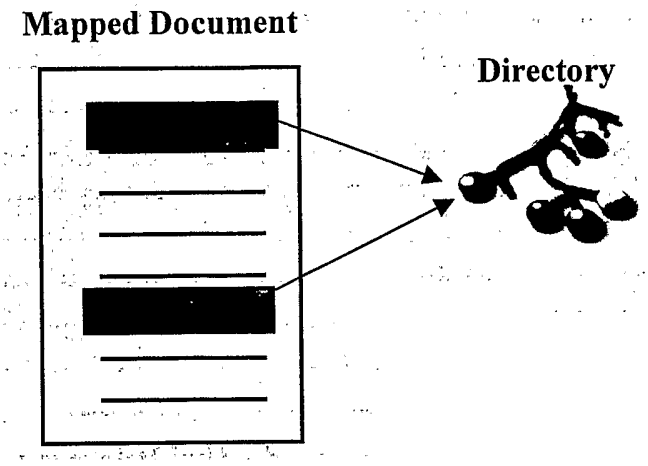


FIG. II - 9

Multiple Words		
Strict Products	59	
Causes Of Action	35	
\ul Tort \ulnone	25	
Negligence Strict	19	
Implied Warranty	18	
Breach Of Warranty	16	
Breach Of Implied	15	
Strict Liability In \ul	15	
Negligence And Strict	13	
Breach Of Implied Warranty	11	
Recover Damages	10	
Action Sounding	9	
Failure To Warn	9	
Negligence And Strict Products	9	
Inter Alia	8	
Action To Recover	6	
Causes Of Action Sounding	6	
City Of New York	6	
Commenced This Action	6	
Action To Recover Damages	5	
Economic Loss	5	
Defective Design	5	
Defective Product	5	
Held Liable	5	
Negligence Breach	5	
Negligence Strict Products	5	
Public Nuisance	5	
Products Liability Or Negligence	5	
Products Liability And Breach	5	
Plaintiff Commenced	5	
Sounding In Strict	5	
Summary Judgment	5	

FIG. II - 10

role	Starting Only	2
function	Starting Only	2
duty	Starting Only	2
duties	Starting Only	2
obligat	Starting Only	2
oblige	Starting Only	2
responsibilit	Starting Only	2
requir	Starting Only	2
tax	Starting Only	2
duty	Starting Only	2
levy	Starting Only	2
levies	Starting Only	2
dues	Starting Only	2
exise	Starting Only	2
customs	Starting Only	2
toll	Starting Only	2
duties	Starting Only	2
duty	Starting Only	2
accident	Starting Only	2
crash	Starting Only	2
collision	Starting Only	2
tragedies	Starting Only	2
tragedy	Starting Only	2
colliding	Starting Only	2
collide	Starting Only	2
amputat	Starting Only	2
dismember	Starting Only	2
lose	Starting Only	2
cuts	Exact Phrase	2

Loss	Starting Only	2
lost	Starting Only	2
damage	Exact Phrase	2
damaging	Starting Only	2
losing	Starting Only	2
harm	Starting Only	2
suffer	Starting Only	2
injur	Starting Only	2
grievance	Starting Only	2
hurt	Starting Only	2
impair	Starting Only	2
maim	Starting Only	2
pain	Starting Only	2
ache	Starting Only	2
casualty	Starting Only	2
casualties	Starting Only	2
careless	Starting Only	2
reckless	Starting Only	2
death	Starting Only	2
died	Starting Only	2
dying	Starting Only	2
die	Exact Phrase	2
fatalities	Starting Only	2
fatality	Starting Only	2
loss of life	Starting Only	2
operat	Starting Only	2
duty	Starting Only	2
duties	Starting Only	2
duty	Starting Only	2
duties	Starting Only	2
degree	Starting Only	2

FIG. II - 11

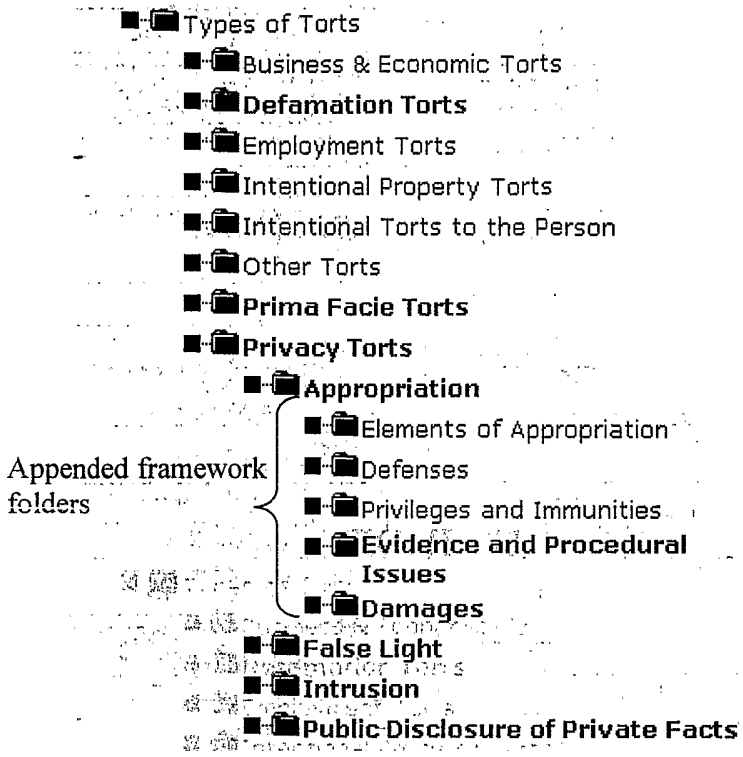

Types of Torts
Business & Economic Torts
Defamation Torts
Employment Torts
Intentional Property Torts
Intentional Torts to the Person
Other Torts
Prima Facie Torts
Privacy Torts
Appropriation
Elements of Appropriation
Defenses
Privileges and Immunities
Evidence and Procedural Issues
Damages
False Light
Intrusion
Public Disclosure of Private Facts
Appended framework folders

FIG. 12

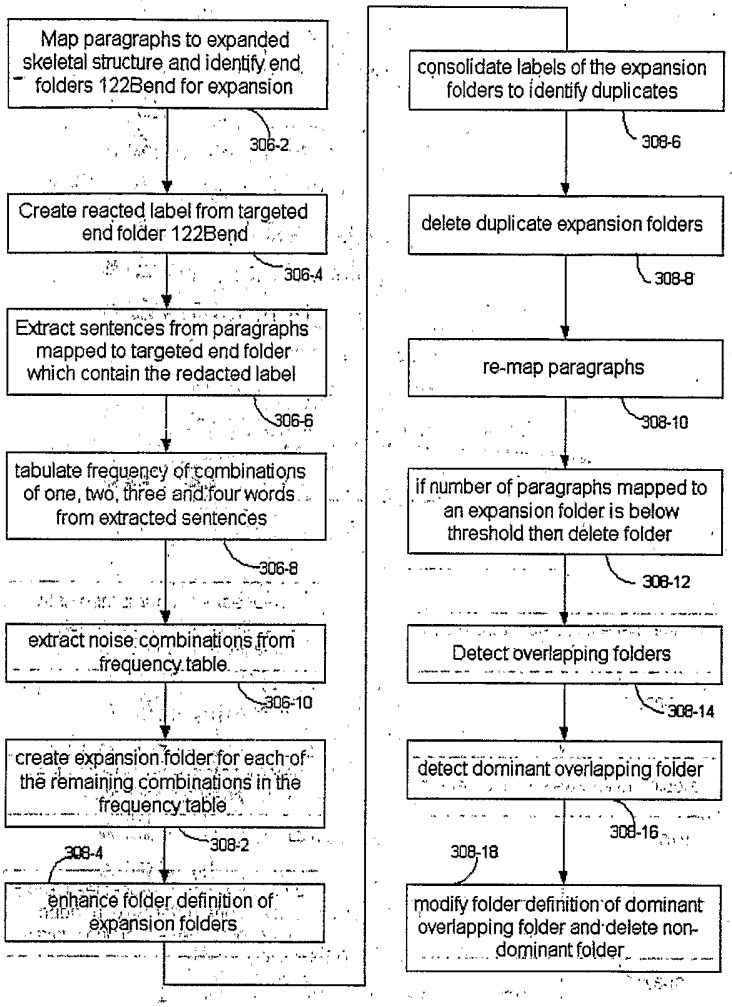

Map paragraphs to expanded skeletal structure and identify end folders 122Bend for expansion
306-2
Create reacted label from targeted end folder 122Bend
306-4
Extract sentences from paragraphs mapped to targeted end folder which contain the redacted label
306-6
tabulate frequency of combinations of one, two, three and four words from extracted sentences
306-8
extract noise combinations from frequency table
306-10
create expansion folder for each of the remaining combinations in the frequency table
308-2
308-4
enhance folder definition of expansion folders
consolidate labels of the expansion folders to identify duplicates
308-6
delete duplicate expansion folders
308-8
re-map paragraphs
308-10
if number of paragraphs mapped to an expansion folder is below threshold then delete folder
308-12
Detect overlapping folders
308-14
detect dominant overlapping folder
308-16
308-18
modify folder definition of dominant overlapping folder and delete non-dominant folder

FIG. II - 13A

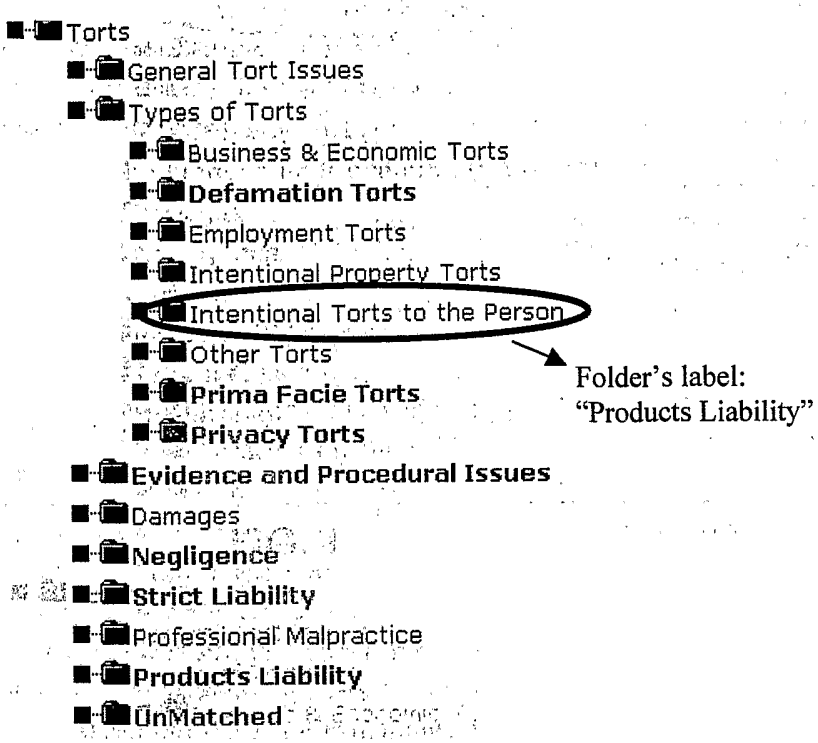

Torts
General Tort Issues
Types of Torts
Business & Economic Torts
Defamation Torts
Employment Torts
Intentional Property Torts
Intentional Torts to the Person
Other Torts
Prima Facie Torts
Privacy Torts
Evidence and Procedural Issues
Damages
Negligence
Strict Liability
Professional Malpractice
Products Liability
UnMatched
Folder's label:
"Products Liability"

FIG. II - 13B

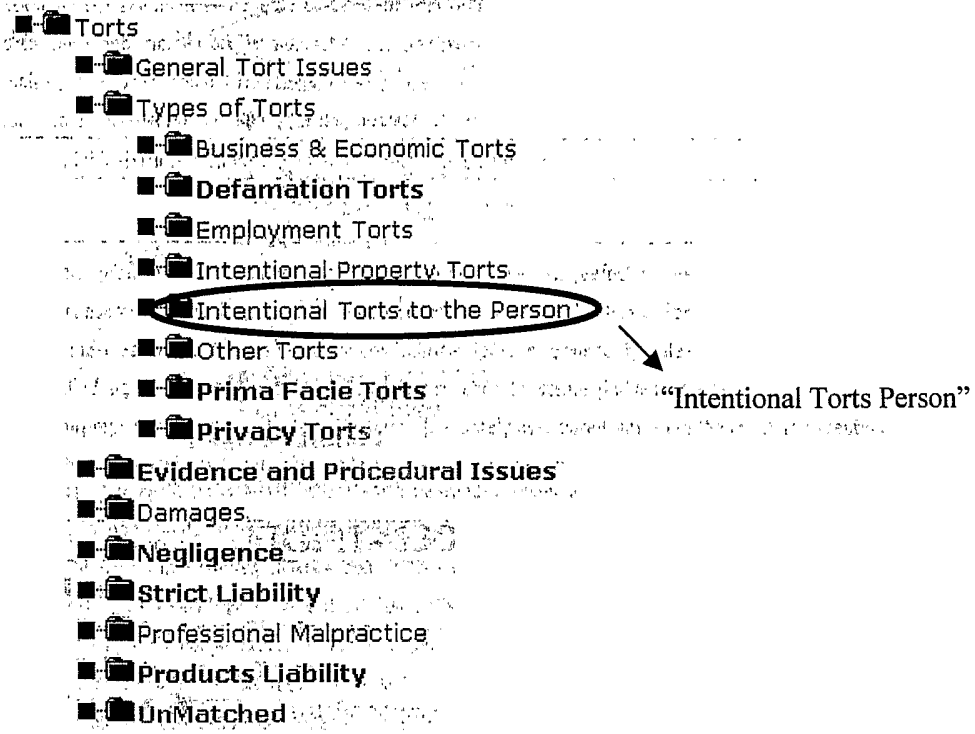
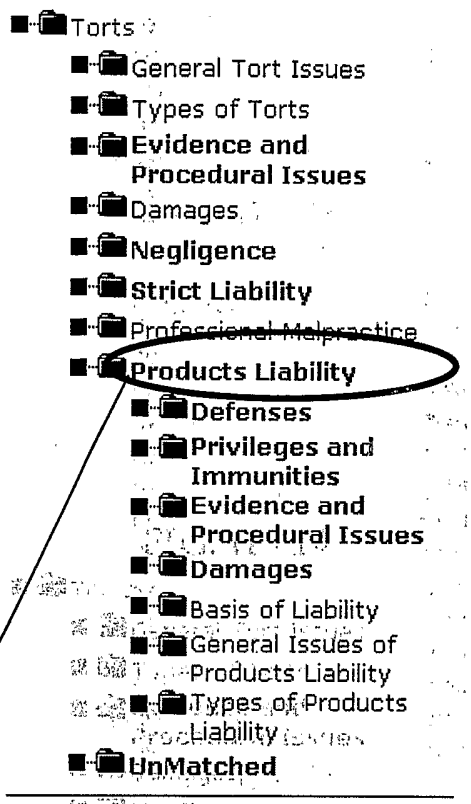

Torts
General Tort Issues
Types of Torts
Business & Economic Torts
Defamation Torts
Employment Torts
Intentional Property Torts
Intentional Torts to the Person
Other Torts
Prima Facie Torts
Privacy Torts
Evidence and Procedural Issues
Damages
Negligence
Strict Liability
Professional Malpractice
Products Liability
UnMatched
"Intentional Torts Person"

FIG. II - 14



1.	P1	consumer|crash-worth-|crashworth-|crash worth-|learned-intermediar-|learned intermediar-|learnedintermediar-|malfunction-|merchantability-|$product$|products-|product's-|risk-benefit-|risk-utilit-|cost-benefit-|costbenefit-|riskbenefit-|riskutilit-|risk benefit-|cost benefit-|risk utilit-|sophisticated buyer-|sophisticated user-|sophisticated purchaser-|sophisticated-buyer-|sophisticatedbuyer-|sophisticatedpurchaser-|sophisticated-purchaser-|sophisticated-user-|sophisticateduser-|design defect-|design-defect-|latent defect-|latent-defect-|common-defect-|common defect-|warning defect-|warning-defect-
	P2	liability-|liable-|strict-|responsible-|mean-|defect|manufactur
	Same Sent	1-20-2
2.	P1 M	risk-benefit-|risk benefit-|consumer expectation-|unreasonable risk-|defect-|liable-|liabilit-|risk-benefit-|risk benefit-|consumer expectation-|unreasonable risk-|risk-|defect-|liable-|liabilit-|ordinary purpos-|strict-|negligen-|responsible-|responsibilit-|unsafe-|safeguard-|fabricat-|prepar-|prior to the sale-|prior to sell-|prior to a sale-|prior to the sell-|danger|substitut-|first sold-|warn-|toxic reaction-|unstable-|hidden|misuse instruction|imperfect|incomplet|factory-|factories-|intimately associated with-|line theory-|line exception-|malfunction-|not perform-|faultlessly made-|crime or fraud-|safety feature-|warrant-|warn-|adequat|specification-|suitable-|suitabilit-
	P2 M	manufactur|$product$|(design+!designat+!designed to protect)|component-|workmanship-|appliance-|products-
	Same Sent	1-15-2

FIG. II - 15

Identify	Replace	ReplaceWith
%by	by	bied
%cy-	cy	cied
%dy	dy	died
%fy	fy	fied
%gy	gy	gied
%hy	hy	hied
%jy	jy	jied
%ky	ky	kied
%ly	ly	lied
%my	my	mied
%ny	ny	nied
%py	py	pied
%qy	qy	qied
%ry	ry	ried
%sy	sy	sied
%ty	ty	tied
%vy	vy	ved
%wy	wy	wied
%xy	xy	xied
%zy	zy	zied
%by	by	bies
%cy	cy	cies
%dy	dy	dies
%fy	fy	fies
%gy	gy	gies
%hy	hy	hies

FIG. 16

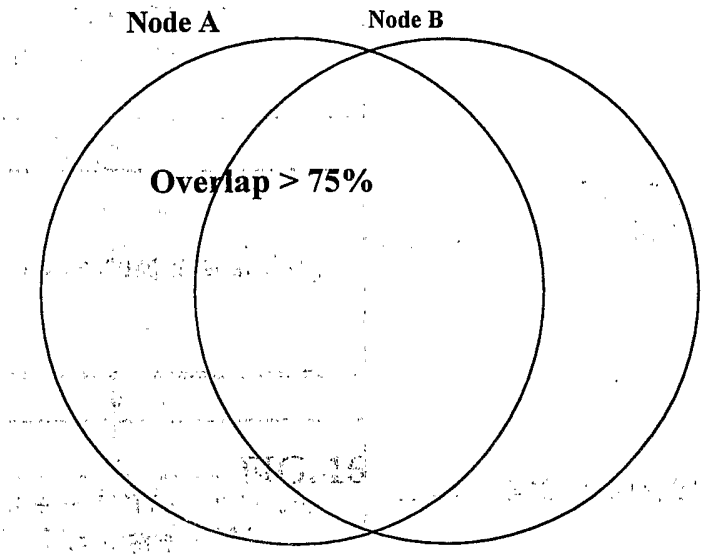


FIG. 17

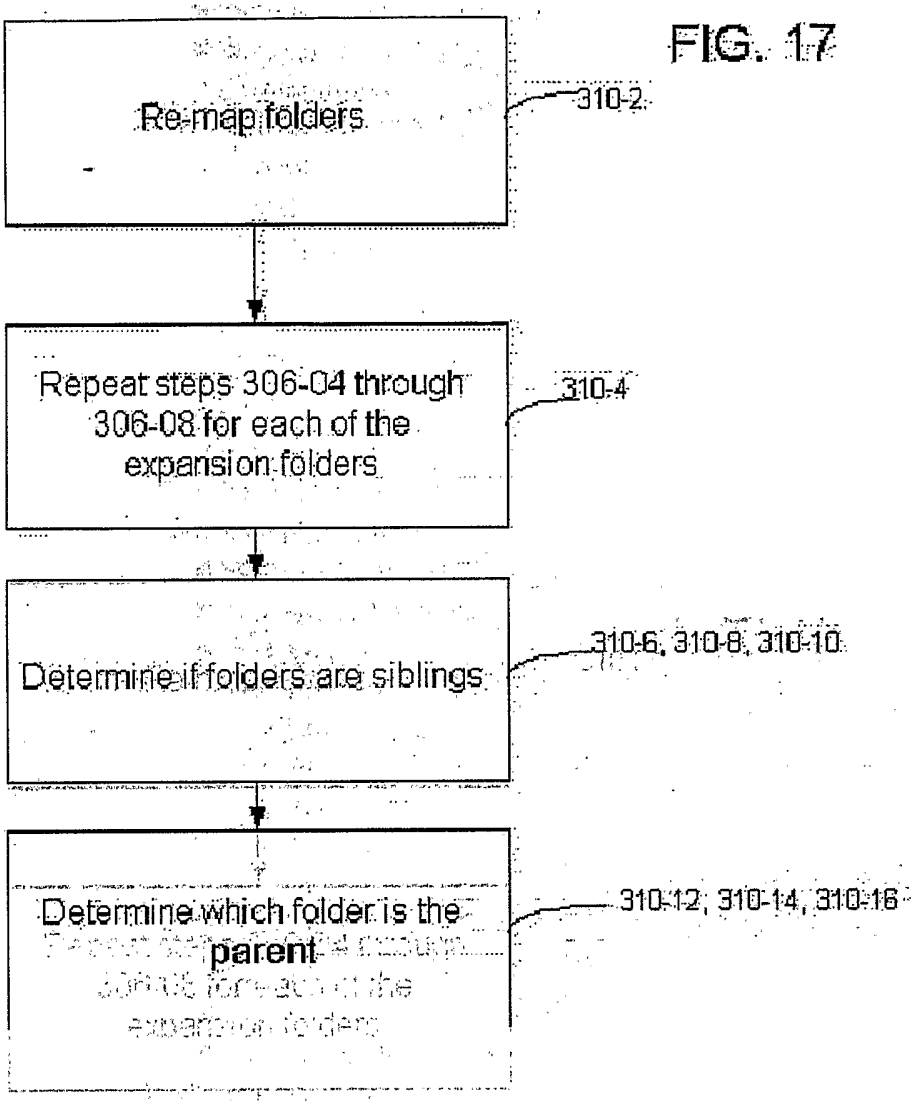
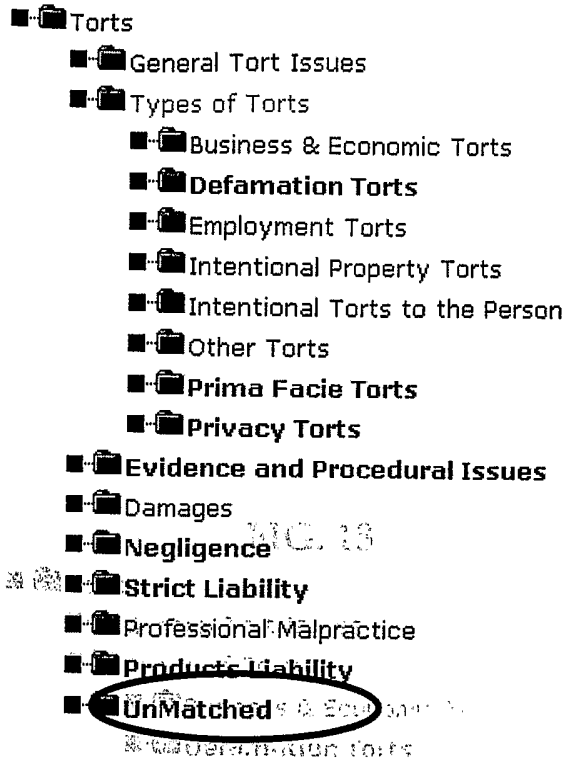

Re-map folders
310-2
Repeat steps 306-04 through 306-08 for each of the expansion folders
310-4
Determine if folders are siblings
310-6, 310-8, 310-10
Determine which folder is the parent
310-12, 310-14, 310-16

FIG. 18


Torts
General Tort Issues
Types of Torts
Business & Economic Torts
Defamation Torts
Employment Torts
Intentional Property Torts
Intentional Torts to the Person
Other Torts
Prima Facie Torts
Privacy Torts
Evidence and Procedural Issues
Damages
Negligence
Strict Liability
Professional Malpractice
Products Liability
UnMatched