

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
31 May 2007 (31.05.2007)

PCT

(10) International Publication Number
WO 2007/061451 A1

(51) International Patent Classification:
G06F 17/30 (2006.01)

(21) International Application Number:
PCT/US2006/025101

(22) International Filing Date: 28 June 2006 (28.06.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/273,568 14 November 2005 (14.11.2005) US
11/314,835 21 December 2005 (21.12.2005) US

(71) Applicant (for all designated States except US): **Make Sence, Inc.**; 3076 Sir Francis Drake's Highway, P.O. Box 3463, Roadtown, Tortola (VG).

(71) Applicants and

(72) Inventors: **BOBICK, Mark** [US/US]; 236 Wayne Avenue, Indialantic, Florida 32903 (US). **WIMMER, Carl** [CA/BB]; 2002 Worthy Down, Graeme Hall, Christ Church (BB).

(74) Agents: **STEWART, David** et al.; 255 South Orange Avenue, Suite 1401, Orlando, Florida 32801 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

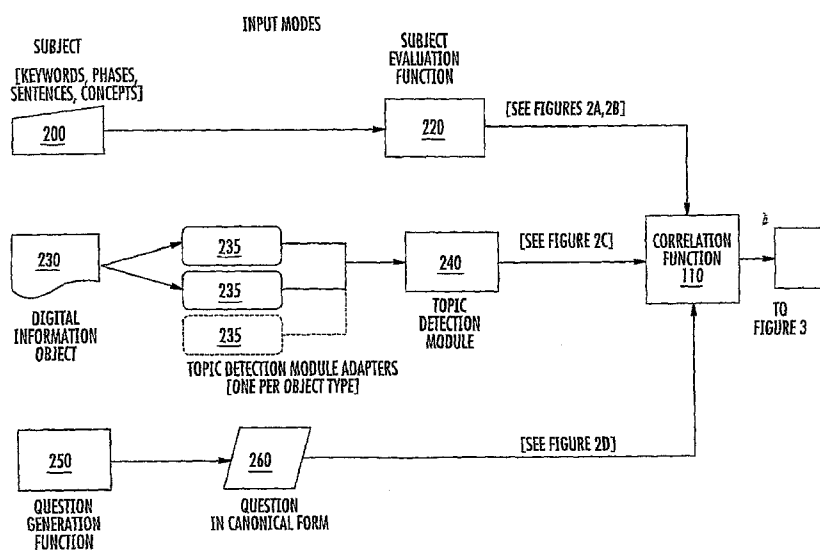
(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

- with international search report
- before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments

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: A KNOWLEDGE CORRELATION SEARCH ENGINE



(57) **Abstract:** Clauses are important for a variety of NLP tasks such as predicting phrasing in text-to-speech synthesis and inferring text alignment for machine translation (Ejherhed 1988, Leffa 1998, Papageorgiou 1997). The Computational Natural Language Learning 2001 shared task (Sang & Déjean 2001) set the goal of identifying clause boundaries in text using machine learning methods. Systems created for the task predicted a label for each word specifying the number of clauses starting and ending at that position in the sentence without differentiating between clause types. This work extends that of the shared task in several ways: (1) performance bounds are explored, (2) an attempt is made to distinguish 'main' and 'subordinate' clauses, and (3) Winnow and maximum entropy, model classes proven effective in similar domains yet not previously employed for the task, are applied to the problem.

WO 2007/061451 A1

A KNOWLEDGE CORRELATION SEARCH ENGINE

Cross Reference To Related Applications

This application is a continuation-in-part of and claims priority to each of (1) U.S. Serial No. 11/273,568, filed November 14, 2005 entitled Techniques For Knowledge Discovery By Constructing Knowledge Correlations Using Concepts or Terms, by inventors Mark Bobick and Carl Wimmer (Attorney docket no. 83071), and (2) U.S. Serial No. 11/314,835, filed December 21, 2005, entitled Techniques For Knowledge Discovery By Constructing Knowledge Correlations Using Concepts or Terms, by inventors Mark Bobick and Carl Wimmer (Attorney docket no. 83071_CIP), the contents of each of which are hereby incorporated into this application by reference in their entirety.

Background Of The Invention

Field Of The Invention

The invention is related to information technology and, more particularly, to a search engine that utilizes the results of knowledge correlation to identify network and/or Internet resources significant to any given user question, subject, or topic of a digital information object.

Description Of The Related Art

Search engines are widely acknowledged to be part of the Information Retrieval (IR) domain of knowledge. IR methods are directed to locating resources (typically documents) that are relevant to a question called a query. That query can take forms ranging from a single search term to a complex sentence composed in a natural language such as English. The collection of potential resources that are searched is called a corpus (body), and different techniques have been developed to search each type of corpus. For example, techniques used to search the set of articles contained in a digitized encyclopedia differ from the techniques used by a web search engine. Regardless of the techniques utilized, the core issue in IR is relevance – that is, the relevance of the documents retrieved to the original query. Formal metrics are applied to compare the

effectiveness of the various IR methods. Common IR effectiveness metrics include precision, which is the proportion of relevant documents retrieved to all retrieved documents; recall, which is the proportion of relevant documents retrieved to all relevant documents in the corpus; and fall-out, which is the proportion of irrelevant documents retrieved to all irrelevant documents in the corpus. Post retrieval, documents deemed relevant are (in most IR systems) assigned a relevance rank, again using a variety of techniques, and results are returned. Although most commonly the query is submitted by – and the results returned to – a human being called a user, the user can be another software process.

Text retrieval is a type of IR that is typically concerned with locating relevant documents which are composed of text, and document retrieval is concerned with locating specific fragments of text documents, particularly those documents composed of unstructured (or “free”) text.

The related knowledge domain of data retrieval differs from IR in that data retrieval is concerned with rapid, accurate retrieval of specific data items, such as records from a SQL database.

Information extraction (IE) is another type of IR which is has the purpose of automatic extraction of information from unstructured (usually text) documents into data structures such as a template of name/value pairs. From such templates, the information can subsequently correctly update or be inserted into a relational database.

Search engines that have been described in the literature or released as software products use a number of forms of input, ranging from individual keywords, to phrases, sentences, paragraphs, concepts and data objects. Although the meanings of *keyword*, *sentence*, and *paragraph* conform to the common understanding of the terms, the meanings of *phrase*, *concept*, and *data object* varies by implementation. Sometimes, the word *phrase* is defined using its traditional meaning in grammar. In this use, types of phrases include Prepositional Phrases (PP), Noun Phrases (NP), Verb Phrases (VP), Adjective Phrases, and Adverbial Phrases. For other implementations, the word *phrase* may be defined as any proper name (for example “New York City”). Most definitions require that a phrase contain multiple words, although at least one definition permits even a single word to be considered a phrase. Some search engine implementations utilize a

lexicon (a pre-canned list) of phrases. The WordNet Lexical Database is a common source of phrases.

When used in conjunction with search engines, the word *concept* generally refers to one of two constructs. The first construct is *concept* as a cluster of related words, similar to a thesaurus, associated with a keyword. In a number of implementations, this cluster is made available to a user - via a Graphic User Interface (GUI) for correction and customization. The user can tailor the cluster of words until the resulting *concept* is most representative of the user's understanding and intent. The second construct is *concept* as a localized semantic net of related words around a keyword. Here, a local or public ontology and taxonomy is consulted to create a semantic net around the keyword. Some implementations of *concept* include images and other non-text elements.

Topics in general practice need to be identified or "detected" from a applying a specific set of operations against a body of text. Different methodologies for identification and/or detection of topics have been described in the literature. Use of a topic as input to a search engine therefore usually means that a body of text is input, and a required topic identification or topic detection function is invoked. Depending upon the format and length of the resulting topic, an appropriate relevancy function can then be invoked by the search engine.

Data objects as input to a search engine can take forms including a varying length set of free form sentences, to full-length text documents, to meta-data documents such as XML documents. The Object Oriented (OO) paradigm dictates that OO systems accept objects as inputs. Some software function is almost always required to process the input object so that the subsequent relevance function of the search engine can proceed.

Ranked result sets have been the key to marketplace success for search engines. The current dominance of the Google search engine (a product of Google, Inc.) is due to far more to the PageRank system used in Google that lets (essentially) the popularity of a given document dictate result rank. Popularity in the Google example applies to the number of links and to the preferences of Google users who input any given search term or phrase. These rankings permit Google to optimize searches by returning only those documents with ranks above a certain threshold (called *k*). Other methods used by web search engines to rank results include "Hubs & Authorities" which counts links into and

out of a given web page or document, Markov chains, and random walks.

Brief Summary Of The Invention

The present invention discloses a new and novel form of search engine which utilizes a computer implemented method to identify at least one resource, referenced by that resource's unique URI (*Uniform Resource Identifier*) or referenced by that resource's URL (*Uniform Resource Locator*), such resource being significant to any given user question, subject, or topic of a digital information object. For the present invention, the user question or subject or topic acts as input. The input is utilized by a software function which attempts to construct or discover logical structures within a collection of data objects, each data object being associated with the resource that contributed the data object, and the constructed or discovered logical structures being strongly associated with the input. For a preferred embodiment, that software function is a knowledge correlation function as described in said Serial No. 11/273,568 and the logical structure is a form of directed acyclic graph termed a quiver of paths. If such logical structures strongly associated with the input are in fact constructed or discovered, the data object members of such logical structures become an answer space. Using the answer space, another software function is then able to determine with a high degree of confidence which of the resources that contributed to the answer space are the most significant contributors to the answer space, and thereby identify URLs and URIs most significant to the input question, subject or topic. Finally, a software function is used to rank in significance to the input each of the URL and URI referenced resources that contributed data objects to the answer space.

The present invention differs from existing search engines because the Knowledge Correlation process as described in said Serial No. 11/273,568, which is used in this invention, attempts to construct an exhaustive collection of paths describing all connections - called correlations - between one term, phrase, or concept referred to as X (or "origin") and a minimum of a second term, phrase or concept referred to as Y (or "destination"). If one or more such correlations can in fact be constructed, the present invention identifies as relevant all resources which contributed to constructing the correlation(s). Unlike existing search engines, relevancy in the present invention applies

not to individual terms, phrases or concepts in isolation but instead to the answer space of correlations that includes not only the X and the Y, but to all the terms, phrases and concepts encountered in constructing the correlations. Because of these novel characteristics, the present invention is uniquely capable of satisfying user queries for which cannot be answered using the content of a single web page or document.

Input to the present invention differs from current uses because all input modes of the present invention must present a minimum of two (2) non-identical terms, phrases, or concepts. "Non-identical" in this usage means lexical or semantic overlap or disjunction is required. As described in said Serial No. 11/273,568, the minimum two terms, phrases, or concepts are referred to as X and Y (or "origin" and "destination"). No input process can result in synonymy, identity, or idempotent X and Y term, phrases or concepts. As with existing art, text objects and data objects can be accepted (in the present invention, as either X or Y) and the topics and/or concepts can be extracted prior to submission to the Knowledge Correlation process. However, unlike most (if not all) existing search engines, the form of the input (term, phrase, concept, or object) is not constrained in the present invention. This is possible because the relevancy function (Knowledge Correlation) does not utilize similarity measures to establish relevancy. This characteristic will allow the present invention to be seamlessly integrated with many existing IR applications.

Regardless of the forms or methods of input, the purpose of Knowledge Correlation in the present invention is to establish document relevancy. Currently, relevancy is established in IR using three general approaches: set-theoretic models which represent documents by sets; algebraic models which represent documents as vectors or matrices; and probabilistic models which use probabilistic theorems to learn document attributes (such as topic). Each model provides a means of determining if one or more documents are similar and thereby, relevant, to a given input. For example, the most basic set-theoretic model uses the standard Boolean approach to relevancy – does an input word appear in the document? If yes, the document is relevant. If no, then the document is not relevant. Algebraic models utilize techniques such as vector space models where documents represented as vectors of terms are compared to the input query represented as a vector of terms. Similarity of the vectors implies relevancy of the

documents. For probabilistic models, relevancy is determined by the compared probabilities of input and document.

As described above, the present invention establishes relevancy by an entirely different process, using an entirely different criteria than any existing search engine. However, the present invention is dependent upon Discovery and Acquisition of “relevant” sources within the corpus (especially if that corpus is the WWW). For this reason, any form of the existing art can be utilized without restriction during the Discovery phase as described in said Serial No. 11/273,568 to assist in identifying candidate resources for input to the Knowledge Correlation process.

For all search engines, simply determining relevancy of a given document to a given input is necessary but not sufficient. After all – using the standard Boolean approach to relevancy as an example – for any query against the WWW, which contained the word “computer”, tens of millions of documents would qualify as relevant. If the user was actually interested only in documents describing a specific application of “computer”, such a large result set would prove unusable. As a practical matter, users require that search engines rank their results from most relevant to least relevant. Typically, users prefer to have the relevant documents presented in order of decreasing relevance – with the most relevant result first. Because most relevance functions produce real number values, a natural way to rank any search engine result set is to rank the members of the result set by their respective relevance scores.

The present invention utilizes a ranking method that is novel because it is a function of the degree to which a given document or resource contributed to the correlation “answer space”. As described in said Serial No. 11/273,568, that answer space is constructed from data structures called nodes, which in turn are created by decomposition of relevant resources. Even the most naïve ranking function of the present invention – which counts the frequency of node occurrence in the answer space – can identify documents that uniquely or strongly relevant to the original user query. More sophisticated ranking mechanisms of the current invention as described more hereinafter improve that outcome.

Brief Description Of The Drawings

Figure 1 is a block diagram showing functional components of a search engine in accordance with one aspect of the invention.

Figure 2 is a block diagram of the pre-search block of Figure 1.

Figure 2A is a block diagram of part of an exemplary subject evaluation function for keywords, phrases, sentences and concepts in accordance with one aspect of the invention.

Figure 2B is a block diagram of the remaining part of an exemplary subject evaluation function for compound, complex or orthogonal subjects and for a simple web query in accordance with one aspect of the invention.

Figure 2C is a block diagram of an exemplary topic detection module and related adapter(s) in accordance with one aspect of the invention.

Figure 2D is a block diagram of a question generation function in accordance with one aspect of the invention.

Figure 3 is a copy of Figure 1A of Serial No. 11/273,568.

Figure 4 is a copy of Figure 1B of Serial No. 11/273,568.

Figure 5 is a copy of Figure 1C of Serial No. 11/273,568.

Figure 6 is a copy of Figure 2A of Serial No. 11/273,568.

Figure 7 is a copy of Figure 2E of Serial No. 11/273,568.

Figure 8 is a block diagram of the post search block 120 of Figure 1.

Description Of The Invention

Figure 1 is a block diagram of three examples of input accepted by the correlation function 110. A subject 200 is evaluated by the subject evaluation function 220. A digital information object 230 is examined for a topic by an adapter 235 of the topic detection module 240. A canonical form question generation function 250 generates a question 260 as input.

In a preferred embodiment, a minimum of two inputs in any form and from any source as described more hereinafter must be submitted to the correlation function 110. There is a first such input, called the X or "origin" input and there is a second such input,

called the Y or "destination" input. Accordingly, acceptable inputs may include any combination of two subjects 200, digital information objects 230, or questions 260.

In another embodiment as described more hereinafter, a minimum of one X input which is a subject 200, digital information object 230 or question 260 is submitted to the correlation function 110. A second input, called a stop correlation condition, is passed to the correlation function 110. The actual value of the required Y which is a correlation destination remains unknown until the stop correlation function is satisfied by the correlation function 110. No actual Y input need be processed as input, but the requirement for a correlation destination is satisfied.

The first example illustrated in Figure 1 is illustrated in greater detail in Figure 1A. In one embodiment, the subject 200 may be an individual keyword, a phrase, a sentence, or a concept. When the subject 200 is an individual keyword, the subject 200 is passed by the subject evaluation function 220 directly to the correlation function 110 without further processing. Likewise, when the subject 200 is a phrase, the subject 200 is passed by the subject evaluation function 220 directly to the correlation function 110 without further processing. When the subject 200 is a sentence, a natural language parser (NLP) 133 will be invoked to perform a syntactic analysis of the sentence to extract the actual subject 200 of the sentence in the form of words and/or phrases. Such words or phrases will then be passed to the correlation function 110. Additional words or phrases may be extracted from the sentence and submitted to the correlation function 110 as context. As described in said Serial No. 11/273,568, any number of context words or phrases which are in addition to the X or Y words or phrases can be submitted to the correlation function 110 to improve said function. The selection of what words or phrases (if any) that are to be extracted from a sentence is based upon the membership of the word or phrase in any lexicon of the NLP 133, and the absence of the word from a common list of stop words. Stop words are well known in IR. Such words cannot be used to establish relevance in set-theoretic models of IR, so are never added to the indexes built for such models.

In the event that the sentence is a question 250 which matches a canonical form, the subject evaluation function 220 will extract from the sentence both the X and Y words or phrases and submit them to the correlation function 110. When the subject 200

is a concept, the concept word or phrase will be submitted to the correlation function 110 as either X or Y, and the remaining terms in the concept cluster or map will be submitted to the correlation function 110 as context words or phrases.

In a preferred embodiment, the subject will be provided by a user by means of Graphical User Interface such as Serial No. 11/273,568 Figure 2A. In other embodiments, any well known input interface will be utilized (e.g. text input field, spoken input, etc.).

In one embodiment, referring to FIGURE 1A, the subject 200 shall take the form of a complex subject, that is, a subject that consists of one independent clause, and one or more dependent clauses. For example, "regulation of pollution, given the effects of automobile pollution". In other embodiments, the subject 200 shall take the form of a compound subject, that is, a subject that consists of two or more independent clauses connected using logical operators such as "and" "or" "not". For example, "the Trilateral Commission and international NGOs not World Bank". Alternatively, the subject 200 shall take the form of a multi-part orthogonal subject, that is, a subject that consists of two or more independent clauses which are not connected, and which may be orthogonal with respect to each other. For example, "poaching, endangered species, men's health, government intervention". Alternatively, the subject 200 shall take the form of a multi-part orthogonal subject, that is, a subject that consists of two or more independent clauses which are not connected, and which may be orthogonal with respect to each other. For example, "poaching, endangered species, men's health, government intervention". In these embodiments, advanced NLP methods for clause recognition (see Hachey, B.C. 2002. Thesis: *Recognising Clauses Using Symbolic and Machine Learning Approaches*. University of Edinburgh) will be applied to the subject 200 to first decompose the subject 200 into clauses and from there, by means of syntactic analysis, into keywords and phrases. Clause recognition techniques will be used to discriminate between X, Y, and context inputs to the correlation function 110.

In one embodiment, the subject evaluation function 220 will determine if the user-provided subject 200 would produce as response from the present invention a listing as the most appropriate response. For example, referring to FIGURE 1B, is the user-provided subject is "Italian restaurants Dover DE", the subject evaluation function 220

will recognize that a listing of Italian restaurants in Dover, Delaware is sought. In this event, the subject evaluation function 220 will either direct the user to use one of the well known simple web search engines such as Google (a product of Google, Inc.) or Yahoo (a product of Yahoo, Inc.), or will directly invoke one of those simple search engines. Alternatively, the subject evaluation function 220 will determine if the user-provided subject would produce as response a single web page as the most appropriate response. For example, if the user-provided subject is “show times rialto theatre”, the subject evaluation function 220 will recognize that the web site for the Rialto Theatre is sought. In this event, the subject evaluation function 220 will either direct the user to use one of the well known simple web search engines such as Google or Yahoo, or will directly invoke the web site of Rialto Theatre, or will directly invoke one of the simple search engines named above. This is achieved by an automatic phrase recognition techniques (see Kellely, F., Smeaton, A.F. 1997. *Automatic Phrase Recognition and Extraction from Text*. Proceedings of the 19th Annual BCS-IRSG Colloquium on IR Research) using the rule that when precisely two perfect phrases comprise the subject 220 and one of the phrases is a proper geographical name (e.g. “New York City”) or a proper name (“Rialto Theatre”) and one of the phrases is an adjective + noun phrase (“show times” or “Italian restaurants”), the simple web search engine should be invoked. More sophisticated rules can easily be defined to cover most circumstances.

The third input mode illustrated in Figure 1 is more fully illustrated in Figure 2A of patent application Serial No. 11/273,568 wherein the input to the correlation function 110 is a user question, and the user question shall be composed of an incomplete question in canonical form and, in addition, one or more key words, wherein the key words complete the question [comparable to the well known paradigm of “fill in the blanks”]. Alternatively, the incomplete question will be explicitly selected by the user. In one embodiment, the incomplete question will be explicitly selected by the user from a list or menu of supported canonical form questions. In another, the list or menu of incomplete supported canonical form questions will be “static” – that is, the list will not vary at each invocation. Alternatively, the list or menu of incomplete supported canonical form questions will be “dynamic” – that is, the list varies at each invocation. Referring to Figure 1, the dynamic list or menu of incomplete supported canonical form questions will

be generated at each invocation by means of a software function, the canonical form question generation function 250, a software program component, written in a computer programming language (e.g. Java, a product of Sun Microsystems, Inc.). Alternatively, the incomplete question will be implicit, the question being selected by a software program component, the canonical form question generation function 250. Or, the incomplete implicit question that will be selected by the canonical form question generation function 250 will be “static” – that is, it will not vary at each invocation.

In a currently preferred embodiment, the static implicit selected question is “What are the connections between [keyword 1] and [keyword 2]?” Alternatively, the static implicit selected question is “What are the connections between [keyword 1] and [keyword 2] in the context of [keyword 3] and/or [keyword 4] and/or [keyword 5]?” Or, the incomplete implicit question that will be selected by the canonical form question generation function 250 will be “dynamic” – that is, it will vary at each invocation.

In one embodiment, the digital information object 230 will be provided by a user. The digital information object 230 will include, but not be limited to the forms:

- (i) text (plain text) files.
- (ii) Rich Text Format (RTF) (a standard developed by Microsoft, Inc.). An alternative method is to first obtain clean text from RTF by the intermediate use of a RTF-to-text conversion utility (e.g. RTF-Parser-1.09, a product of Pete Sergeant).
- (iii) Extended Markup Language (XML) (a project of the World Wide Web Consortium) files.
- (iv) any dialect of markup language files, including, but not limited to: HyperText Markup Language (HTML) and Extensible HyperText Markup Language (XHTML™) (projects of the World Wide Web Consortium), RuleML (a project of the RuleML Initiative), Standard Generalized Markup Language (SGML) (an international standard), and Extensible Stylesheet Language (XSL) (a project of the World Wide Web Consortium).
- (v) Portable Document Format (PDF) (a proprietary format of Adobe, Inc.) files (by means of the intermediate use of a PDF-to-text conversion utility).

- (vi) MS WORD files e.g. DOC files used to store documents by MS WORD (a word processing software product of Microsoft, Inc.) This embodiment programmatically utilizes a MS Word-to-text parser (e.g. the Apache POI project, a product of Apache.org). The POI project API also permits programmatically invoked text extraction from Microsoft Excel spreadsheet files (XLS). An MS Word file can also be processed by a NLP as a plain text file containing special characters, although XLS files cannot.
- (vii) event-information capture log files, including, but not limited to: transaction logs, telephone call records, employee timesheets, and computer system event logs.
- (viii) web pages
- (ix) blog pages
- (x) a relational database row.
- (xi) a relational database view.
- (xii) a relational database table.
- (xiii) a relational database answer set (i.e. the set of rows resulting from a relational algebra operation).

The topic of the digital information object 230 will be determined by a software function, the topic detection function 240, a software program component. Examples of such topic detection software have been well described in the literature (see Chen, K. 1995. *Topic Identification in Discourse*. Morgan Kaufman). The topic detection function 240 will be implemented with software adapters 235 that handle each form of digital information object 230. Such software adapters 235 are well known (for an example, see <http://www-306.ibm.com/software/integration/wbiadapters/framework>). The output of the topic detection function will be keywords and/or phrases which will then be submitted to the correlation function 110.

Figure 8 is a flow chart of the search engine process initiated by the knowledge correlation function 110 upon inputs as described in Figure 1, and continuing through to presentation of results to a user in accordance with one aspect of the invention. The correlation function 110 places relevant data structure objects 830, triples 835 and associated objects 837 into an answer space 885. The significance of the objects in the

answer space 885 is determined by a significance calculation function 840 which sets up data for the ranking function 845 to rank by significance. Output is then displayed to the user. In the event the correlation function creates any kind of directed acyclic graph, the graph can be displayed to the user after being organized for layout by the hierarchical layout function 850.

The present invention is dependent upon the success of the correlation function 110. The following text from patent application Serial No. 11/273,568 summarizes the correlation function 110 used in the present invention. Note that all references to figures in the following quote apply only to figures from patent application Serial No. 11/278,568:

"In an example embodiment of the present invention as represented in Figure 1A, a user enters at least one term via using a GUI interface. Figure 2A is a screen capture of the GUI component intended to accept user input. Significant fields in the interface are "X Term", "Y Term" and "Tangents". As described more hereinafter, the user's entry of between one and five terms or phrases has a significant effect on the behavior of the present invention. In a preferred embodiment as shown in Figure 2A, the user is required to provide at least two input terms or phrases. Referring to FIGURE 1A, the user input 100, "GOLD" is captured as a searchable term or phrase 110, by being entered into the "X Term" data entry field of FIGURE 2A. The user input 100 "INFLATION" is captured as a searchable term or phrase 110 by being entered into the "Y Term" data entry field of FIGURE 2A. Once initiated by the user, a search 120 is undertaken to identify actual and potential sources for information about the term or phrase of interest. Each actual and potential source is tested for relevancy 125 to the term or phrase of interest. Among the sources searched are computer file systems, the Internet, Relational Databases, email repositories, instances of taxonomy, and instances of ontology. Those sources found relevant are called resources 128. The search 120 for relevant resources 128 is called "Discovery". The information from each resource 128 is decomposed 130 into digital information objects 138 called nodes. Referring to FIGURE 1C, nodes 180A and 180B are data structures which contain and convey meaning. Each node is self contained. A node requires nothing else to convey meaning. Referring once again to FIGURE 1A, nodes 180A, 180B from resources 128 that are

successfully decomposed 130 are placed into a node pool 140. The node pool 140 is a logical structure for data access and retrieval. The capture and decomposition of resources 128 into nodes 180A, 180B is called "Acquisition". A correlation 155 is then constructed using the nodes 180A, 180B in the node pool 140, called member nodes. Referring to FIGURE 1B, the correlation is started from one of the nodes in the node pool that explicitly contains the term or phrase of interest. Such a node is called a term-node. When used as the first node in a correlation, the term-node is called the origin 152 (source). The correlation is constructed in the form of a chain (path) of nodes. The path begins at the origin node 152 (synonymously referred to as path root). The path is extended by searching among node members 151 of the node pool 140 for a member node 151 that can be associated with the origin node 152. If such a node (qualified member 151H) is found, that qualified member node is chained to the origin node 152, and designated as the current terminus of the path. The path is further extended by means of the iterative association with and successive chaining of qualified member nodes of the node pool to the successively designated current terminus of the path until the qualified member node associated with and added to the current terminus of the path is deemed the final terminus node (destination node 159), or until there are no further qualified member nodes in the node pool. The association and chaining of the destination node 159 as the final terminus of the path is called a success outcome (goal state), in which case the path is thereafter referred to as a correlation 155, and such correlation 155 is preserved. The condition of there being no further qualified member nodes in the node pool, and therefore no acceptable destination node, is deemed a failure outcome (exhaustion), and the path is discarded, and is not referred to as a correlation. A completed correlation 155 associates the origin node 152 with each of the other nodes in the correlation, and in particular with the destination node 159 of the correlation. The name for this process is "Correlation". The correlation 155 thereby forms a knowledge bridge that spans and ties together information from all sources identified in the search. The knowledge bridge is discovered knowledge."

Because the present invention depends upon the success of the correlation function 535 as described above, the relevancy tests (Serial No. 11/278,568 Figure 1A Item 125) applied to potential sources are of interest. The Discovery phase of the

correlation function 110 as described above utilizes relevancy tests (Serial No. 11/278,568 Figure 1A Item 125) to identify resources (Serial No. 11/278,568 Figure 1A Item 128) for subsequent Acquisition. These relevancy tests (Serial No. 11/278,568 Figure 1A Item 125) are analogous to and intersect with the relevancy approaches described in the Related Art. Note that the fact that a resource (Serial No. 11/278,568 Figure 1A Item 128) is deemed sufficiently relevant to warrant Acquisition by the correlation function 110 does not imply or guarantee that the resource (Serial No. 11/278,568 Figure 1A Item 128) will be found to contribute in a significant way to the answer space 800. The relevancy tests (Serial No. 11/278,568 Figure 1A Item 125) enumerated in Serial No. 11/278,568 are listed following. In one embodiment of the present invention, all the relevancy tests (Serial No. 11/278,568 Figure 1A Item 125) enumerated in Serial No. 11/278,568 as well as all relevancy approaches described in the Related Art will be utilized to select resources (Serial No. 11/278,568 Figure 1A Item 128) for Acquisition by the correlation function 110.

Tests for relevancy in accordance with Serial No. 11/278,568 can include, but are not limited to:

- (i) *that the potential source contains a match to the singular or plural form of the term or phrase of interest.*
- (ii) *that the potential source contains a match to a synonym of the term or phrase of interest.*
- (iii) *that the potential source contains a match to a word related to the term or phrase of interest (related as might be supplied by a thesaurus).*
- (iv) *that the potential source contains a match to a word related to the term or phrase of interest where the relation between the content of a potential source and the term or phrase of interest is established by an authoritative reference source.*
- (v) *use of a thesaurus such as Merriam-Webster's Thesaurus (a product of Merriam-Webster, Inc) to determine if any content of a potential source located during a search is a synonym of or related to the term or phrase of interest.*
- (vi) *that the potential source contains a match to a word appearing in a definition in an authoritative reference of one of the terms and/or phrases of interest.*

- (vii) *use of a dictionary such as Merriam-Webster's Dictionary (a product of Merriam-Webster, Inc) to determine if any content of a potential source located during a search appears in the dictionary definition of, and is therefore related to, the term or phrase of interest.*
- (viii) *that the potential source contains a match to a word appearing in a discussion about the term or phrase of interest in an authoritative reference source.*
- (ix) *use of an encyclopedia such as the Encyclopedia Britannica (a product of Encyclopedia Britannica, Inc) to determine if any content of a potential source located during a search appears in the encyclopedia discussion of the term or phrase of interest, and is therefore related to the term or phrase of interest.*
- (x) *that a term contained in the potential source has a parent, child or sibling relation to the term or phrase of interest.*
- (xi) *use of a taxonomy to determine that a term contained in the potential source has a parent, child or sibling relation to the term or phrase of interest. In this embodiment, the vertex containing the term or phrase of interest is located in the taxonomy. This is the vertex of interest. For each word located in the contents of the potential source, the parent, siblings and children vertices of the taxonomy are searched by tracing the relations (links) from the vertex of interest to parent, sibling, and children vertices of the vertex of interest. If any of the parent, sibling or children vertices contain the word from the content of the potential source, a match is declared, and the source is considered an actual source of information about the term or phrase of interest. In this embodiment, a software function, called a graph traversal function, is used to locate and examine the parent, sibling, and child vertices of term or phrase of interest.*
- (xii) *that the term or phrase of interest is of degree (length) one semantic distance from a term contained in the potential source.*
- (xiii) *that the term or phrase of interest is of degree (length) two semantic distance from a term contained in the potential source.*
- (xiv) *use of an ontology to determine that a degree (length) one semantic distance separates the source from the term or phrase of interest. In this embodiment, the vertex containing the term or phrase of interest is located in the ontology. This is*

the vertex of interest. For each word located in the contents of the potential source, the ontology is searched by tracing the relations (links) from the vertex of interest to all adjacent vertices. If any of the adjacent vertices contain the word from the content of the potential source, a match is declared, and the source is considered an actual source of information about the term or phrase of interest.

- (xv) *uses an ontology to determine that a degree (length) two semantic distance separates the source from the term or phrase of interest. In this embodiment, the vertex containing the term or phrase of interest is located in the ontology. This is the vertex of interest. For each word located in the contents of the potential source, the relevancy test for semantic degree one is performed. If this fails, the ontology is searched by tracing the relations (links) from the vertices adjacent to the vertex of interest to all respective adjacent vertices. Such vertices are semantic degree two from the vertex of interest. If any of the semantic degree two vertices contain the word from the content of the potential source, a match is declared, and the source is considered an actual source of information about the term or phrase of interest.*
- (xvi) *uses a universal ontology such as the CYC Ontology (a product of Cycorp, Inc) to determine the degree (length) of semantic distance from one of the terms and/or phrases of interest to any content of a potential source located during a search.*
- (xvii) *uses a specialized ontology such as the Gene Ontology (a project of the Gene Ontology Consortium) to determine the degree (length) of semantic distance from one of the terms and/or phrases of interest to any content of a potential source located during a search.*
uses an ontology and for the test, the ontology is accessed and navigated using an Ontology Language (e.g. Web Ontology Language)(OWL) (a project of the World Wide Web Consortium).

Computers that may be searched in this way include individual personal computers, individual computers on a network, network server computers, network ontology server computers, network taxonomy server computers, network database server computers, network email server computers, network file server computers. Network ontology servers are special typically high performance computers which are dedicated to

the task of supporting semantic search functions for a large group of users. Network taxonomy servers are special typically high performance computers which are dedicated to the task of supporting taxonomic search functions for a large group of users. Network database servers are special typically high performance computers which are dedicated to the task of supporting database functions for a large group of users. Network email servers are special typically high performance computers which are dedicated to the task of supporting email functions for a large group of users. Network file servers are special typically high performance computers which are dedicated to the task of supporting file persistence and retrieval functions for a large group of users. The computer network has a minimum of two network nodes and the maximum number of network nodes is infinity. The computer file system has a minimum of two files and the maximum number of files is infinity.

Upon successful completion of the correlation function 110, an answer space 800 will exist. As described in said Serial No. 11/273,568, and illustrated in Figure 8 of this application, the answer space 800 is composed of correlations (Serial No. 11/278,568 Figure 1B Item 155). The correlations (Serial No. 11/278,568 Figure 1B Item 155) are in turn composed of nodes Figure 5 (Serial No. 11/278,568 Figure 1C Items 180A and 180B). The successful correlations Figure 4 (Serial No. 11/278,568 Figure 1B Item 155) produced by the correlation function 110 are together modeled as a directed graph (also called a digraph) of correlations in one preferred embodiment. Alternatively, the successful correlations Figure 4 (Serial No. 11/278,568 Figure 1B Item 155) produced by the correlation function 110 are together modeled as a quiver of paths of successful correlations. Successful correlations Figure 4 (Serial No. 11/278,568 Figure 1B Item 155) produced by the correlation function 110 are together called, with respect to correlation, the answer space 800. Where the correlation function 110 constructs a quiver of paths where each path in the quiver of paths is a successful correlation, all successful correlations share as a starting point the origin node (Serial No. 11/278,568 Figure 1B Item 152), and all possible correlations (Serial No. 11/278,568 Figure 1B Item 155) from the origin node (Serial No. 11/278,568 Figure 1B Item 152) are constructed. All correlations (Serial No. 11/278,568 Figure 1B Item 155) (paths) that start from the same origin term-node (Serial No. 11/278,568 Figure 1B Item 152) and terminate with the

same target term-node (Serial No. 11/278,568 Figure 1B Item 159) or the same set of related target term-nodes (Serial No. 11/278,568 Figure 1B Item 159) comprise a correlation set.

In a currently preferred embodiment, the answer space 800 is stored in a computer digital memory, or stored on a computer digital storage media (e.g. a hard drive). Such digital memory and digital storage devices are well known. The answer space 800 transiently resides or is persisted on a computing device, a computer network-connected device, or a personal computing device. Well known computing devices include, but are not limited to super computers, mainframe computers, enterprise-class computers, servers, file servers, blade servers, web servers, departmental servers, and database servers. Well known computer network-connected devices include, but are not limited to internet gateway devices, data storage devices, home internet appliances, set-top boxes, and in-vehicle computing platforms. Well known personal computing devices include, but are not limited to, desktop personal computers, laptop personal computers, personal digital assistants (PDAs), advanced display cellular phones, advanced display pagers, and advanced display text messaging devices. The answer space 800 contains or associates a minimum of two nodes (Serial No. 11/278,568 Figure 1C Items 180A and 180B) and the maximum number of nodes (Serial No. 11/278,568 Figure 1C Items 180A and 180B) is infinity.

Because the nodes (Serial No. 11/278,568 Figure 1C Items 180A and 180B) are the products of a decomposition function (Serial No. 11/278,568 Figure 1B Item 130) applied against the resources (Serial No. 11/278,568 Figure 1A Item 128) identified by the Discovery phase of the correlation function 110, the nodes (Serial No. 11/278,568 Figure 1C Items 180A and 180B) are strongly associated with the resources (Serial No. 11/278,568 Figure 1A Item 128) from which the nodes (Serial No. 11/278,568 Figure 1C Items 180A and 180B) were derived. Such resources (Serial No. 11/278,568 Figure 1A Item 128) are here called contributing resources. Further, the answer space 800 is strongly associated with a user query (manifested as input subjects 200, digital information objects 230, or questions 250) because a successful correlation (Serial No. 11/278,568 Figure 1B Item 155) is an existential proof (existential quantification) that the user query can be satisfied from the contents of corpus. The present invention is based

upon the fact that the strong association of the user query to the answer space 800 is transitive to the resources (Serial No. 11/278,568 Figure 1A Items 128) which contributed nodes (Serial No. 11/278,568 Figure 1C Items 180A and 180B) to the answer space, thereby enabling the present invention of a knowledge correlation search engine to deliver highly accurate links of resources (Serial No. 11/278,568 Figure 1A Items 128) which are relevant to the user query.

A requirement of the present invention is that the resources (Serial No. 11/278,568 Figure 1A Item 128) which contributed nodes (Serial No. 11/278,568 Figure 1C Items 180A and 180B) to the answer space 185 must be identified (i.e what are the contributing resources 000?). As can be seen in Serial No. 11/278,568 Figure 1C, Item 180B, a member of node Serial No. 11/278,568 Figure 1C Item 180B is the Sequence (source) (Serial No. 11/278,568 Figure 1C Item 188). The sequence (Serial No. 11/278,568 Figure 1C Item 188) contains the URI of the resource (Serial No. 11/278,568 Figure 1A Item 128) from which the node (Serial No. 11/278,568 Figure 1C Item 180B) was derived (the contributing resource 128 for that node Serial No. 11/278,568 Figure 1C Item 188). Therefore, the present invention can identify contributing resources 128 which are relevant to the user query by simply enumerating the URIs of all resources (Serial No. 11/278,568 Figure 1A Item 128) found in all nodes (Serial No. 11/278,568 Figure 1C Item 188) in the answer space 185.

In an improved, but still rudimentary embodiment, each correlation (Serial No. 11/278,568 Figure 1B Item 155) can be examined, and the frequency of occurrence of a contributing resource 128 in the correlation (Serial No. 11/278,568 Figure 1B Item 155) can be captured in a histogram. The cumulative counts for the occurrence of all contributing resources 128 can then be sorted. The URIs for all contributing resources 000 can then be presented to the user in order of descending frequency of occurrence. For this embodiment and referring to Figure 2, the examination of the correlations (Serial No. 11/278,568 Figure 1B Item 155), capture of frequency of occurrence of contributing resources 128, and the placement of the captured frequency of occurrence of contributing resources 128 into a histogram is performed by a significance calculation function 540. The sorting of the cumulative counts for the occurrence of all contributing resources 128

is performed by a ranking function 545, and the presentation to the user of the sorted results is performed by a hierarchical layout function 550.

In another rudimentary example, the significance calculation function 842 is a statistical function that is based upon the number of unique nodes (Serial No. 11/278,568 Figure 1C Item 180B) contributed to the answer space 885 by each contributing resource 128. In this embodiment, any correlations (Serial No. 11/278,568 Figure 1B Item 155) in the answer space 885 are not considered. The significance calculation function 842 first lists the unique nodes (Serial No. 11/278,568 Figure 1C Item 180B) in the answer space 885, with one entry in the list for each node (Serial No. 11/278,568 Figure 1C Item 180B). Then, the frequency of reference to each contributing resource 128 is counted. Using standard and well-known statistical criteria and methods to measure statistical significance, the k threshold to be used by the ranking function 845 is established, and the most significant contributing resources 128 can be identified and presented to the user.

For another example, the significance calculation function 842 correlates the simple occurrence frequency to the simple contribution frequency value, resulting in a rudimentary significance score. If a scatter plot were used to display this data, the significant resources 128 with highest occurrence frequency and the highest contribution frequency would place farthest to the right and closest to the top. Again, as for all the varied embodiments of the significance calculation function 842 described more hereinafter, standard and well known statistical significance measures are utilized to provide appropriate k threshold information for the ranking function 845. Other statistical techniques that may be utilized by the significance calculation function 842 - as needed - include, but are not limited to: linear (the well known Pearson r) correlation between the frequency of occurrence and simple contribution; non-linear correlations of the plot data; nonparametric statistical approaches such the Kendall coefficient of concordance, computation of the geometric mean for data which have logarithmic relation to each other, and other well known techniques to measure the relationship between the variables.

In one embodiment, a node significance score can be computed by using measures such as the ratio, frequency of occurrence over number of nodes (Serial No. 11/278,568 Figure 1C Item 180B) contributed by that specific node's (Serial No.

11/278,568 Figure 1C Item 180B) contributing resource 128, or the ratio, frequency of occurrence over the average number of nodes (Serial No. 11/278,568 Figure 1C Item 180B) contributed by all contributing resources 128. To improve the speed of the significance calculation function 842, node significance scores can be normalized (0,1) or (-1,1), with the possibility thereby to rapidly determine if a given contributing resource 128 was significant or not significant to the answer space.

In another, the significance calculation function 842 is a link analysis function, the link analysis function 842 taking the correlation (Serial No. 11/278,568 Figure 1B Item 155) as input. This exploits the differences between the correlation (Serial No. 11/278,568 Figure 1B Item 155) created by the correlation function 110 compared to a web graph. The significance calculation function 842 as link analysis function establishes a link popularity score on each of node (Serial No. 11/278,568 Figure 1C Item 180B) in the answer space 128. The link popularity score is determined by means of the number of in-degree links to each node (Serial No. 11/278,568 Figure 1C Item 180B) in the answer space 885. The popularity score values of all nodes (Serial No. 11/278,568 Figure 1C Item 180B) contributed by a contributing resource 128 are then summed. In this embodiment, the aggregate popularity scores of all nodes (Serial No. 11/278,568 Figure 1C Item 180B) contributed by a contributing resource 128 are transit to the contributing resource 128 itself.

In one embodiment, the significance calculation function 842 as link analysis function establishes an importance score on each of the nodes (Serial No. 11/278,568 Figure 1C Item 180B). The importance score is determined by means of the well known Kleinberg Hubs and Authorities algorithm. Hub or Authority scores for all nodes (Serial No. 11/278,568 Figure 1C Item 180B) contributed by a contributing resource 128 are then summed. In this embodiment, the aggregate Hub and Authority scores of all nodes (Serial No. 11/278,568 Figure 1C Item 180B) contributed by a contributing resource 128 are transit to the contributing resource 128. In an embodiment, the importance score is determined by means of the well-known 2nd version of the PageRank algorithm. PageRank scores for all nodes (Serial No. 11/278,568 Figure 1C Item 180B) contributed by a contributing resource 128 are then summed. In this embodiment, the aggregate

PageRank scores of all nodes (Serial No. 11/278,568 Figure 1C Item 180B) contributed by a contributing resource 128 are transit to the contributing resource 128.

The results of resource significance calculation function 842 will be ranked by means of a software function, the ranking function 845, a software program component. In an embodiment, the ranking function 845 implements a simple descending sort, with the contributing resource 128 given the highest value by the significance calculation function 842 awarded the number one rank by the ranking function 845, and the ordinal rank of the other contributing resources 128 being assigned based upon their relative position in the sorted list of significance values. When the significance calculation function 842 is a statistical function that is based upon the number of discrete nodes (Serial No. 11/278,568 Figure 1C Item 180B) contributed to the answer space 885 by each contributing resource 128, and when the ranking function 845 implements a simple descending sort, the ranking function is called rank by contribution. When the significance calculation function 842 is a statistical function that will calculate the sum of the relevance scores for all nodes (Serial No. 11/278,568 Figure 1C Item 180B) contributed to the answer space 885 by each contributing resource 128, and when the ranking function 845 implements a simple descending sort, the ranking function is called rank by relevance. When the significance calculation function 842 is a statistical function that will calculate the sum of the popularity scores, Hub and Authority scores, or PageRank scores for all nodes (Serial No. 11/278,568 Figure 1C Item 180B) contributed to the answer space 885 by each contributing resource 128, and when the ranking function 845 implements a simple descending sort, the ranking function is called rank by significance.

In a currently preferred embodiment, at least two categories of contributing resources 128 contribute nodes (Serial No. 11/278,568 Figure 1C Item 180B) to the answer space 885. The two categories of contributing resources are here designated topical resources, and reference resources. Topical resources provide nodes (Serial No. 11/278,568 Figure 1C Item 180B) with explicit reference to a topic, for example the triple GLOBAL WARMING-AFFECTS-GLACIERS. Reference resources provide nodes (Serial No. 11/278,568 Figure 1C Item 180B) which anchor the foundations in knowledge that support topical resource nodes (Serial No. 11/278,568 Figure 1C Item

180B), for example the triple GLOBAL WARMING-IS-CLIMATE CHANGE, or GLOBAL WARMING-FROM-EMISSIONS.

In an embodiment, a Support Vector Machine (SVM) is created to classify and rank contributing resources. Depending upon the characteristics and number of contributing resource 125 classifications, variations of SVM are utilized, including SVM-RFE (SVM-Recursive Feature Elimination), and R-SVM (Reduced-SVM). In a currently preferred embodiment, where the classification of contributing resources 128 into topical resources and reference resources is required, the application of SVM procedures to classify and rank the contributing resources 128 is essentially identical the SVM procedure used as a diagnostic classifier to identify healthy tissue samples from cancer tissue samples.

In a currently preferred embodiment, the correlation (Serial No. 11/278,568 Figure 1B Item 155) constructed or discovered by the correlation function 110 can be displayed to a user. This display is called a presentation. In a currently preferred embodiment, the presentation of the answer space 128 will be implemented using a hierarchical layout 890. In a currently preferred embodiment, the hierarchical layout 890 will be created using a software function, the hierarchical layout function 850, a software program component. The hierarchical layout function 850 assigns the nodes of graphs on different layers in such a way that most edges in the graph flow in the same direction and the number of intersecting edges are minimized. In a currently preferred embodiment, hierarchical layout function 850 uses the Sugiyama-layout algorithm.

While various embodiments of the present invention have been illustrated herein in detail, it should be apparent that modifications and adaptations to those embodiments may occur to those skilled in the art without departing from the scope of the present invention as set forth in the following claims.

What is claimed is:

1. A search engine comprising:
 - a. a search process using correlation, and
 - b. an input evaluation function for extracting inputs for the search process from user supplied input and providing the inputs for the search to the search process.
2. The search engine of claim 1 in which the input evaluation function comprises a subject evaluation function for extracting subject information from keywords, phrases, sentences, concepts, compound, complex or orthogonal inputs or a simple web query and passing the subject information to the search process.
3. The search engine of claim 2 in which the subject evaluation function for keywords and phrases comprises a pass through function.
4. The search engine of claim 2 in which the subject evaluation function for sentences comprises a natural language parser.
5. The search engine of claim 2 in which the subject evaluation function for concepts comprises subject, object and optionally context information.
6. The search engine of claim 2 in which the subject evaluation function for compound, complex or orthogonal inputs comprises a clause recognition function and a natural language parser.
7. The search engine of claim 2 in which the subject evaluation function for a simple web query comprises a phrase recognition process.
8. The search engine of claim 1 in which the input comprises a digital information object and the input evaluation function comprises a topic detection module adapter and a topic detection module.
9. The search engine of claim 8 further comprising a natural language parser receiving the output of the topic detection module.
10. The search engine of claim 1 in which the input comprises a question generation function for defining a query.
11. The search engine of claim 10 in which the question generation function creates a static menu of questions.

12. The search engine of claim 10 in which the question generation function creates a dynamic menu of questions.
13. The search engine of claim 10 in which the query is presented in a canonical form.
14. The search engine of claim 1 in which the output of the search process is based on an answer space.
15. The search engine of claim 14 in which the output of the search process is strongly associated with resources identified from the answer space.
16. The search engine of claim 14 in which the output is strongly associated with resources used to create the answer space.
17. The search engine of claim 14 in which the output is transitively associated with the user input through resources used to create the answer space.
18. The search engine of claim 14 in which relevancy of the resources to the user input is guaranteed by the existence of an answer space.
19. The search engine of claim 14, which does not use a similarity measure to establish relevancy.
20. The search engine of claim 1 in which the output of the search process is applied to a post search process for determining the order of presentation to a user.

1/12

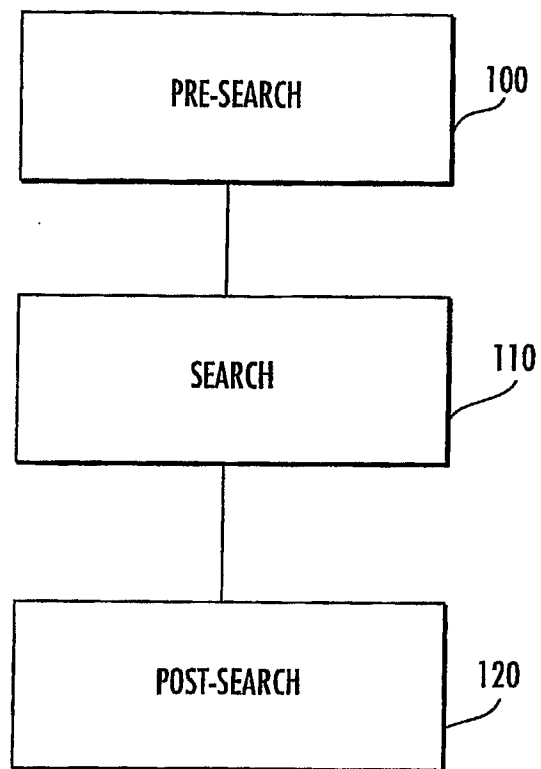


FIG. 1

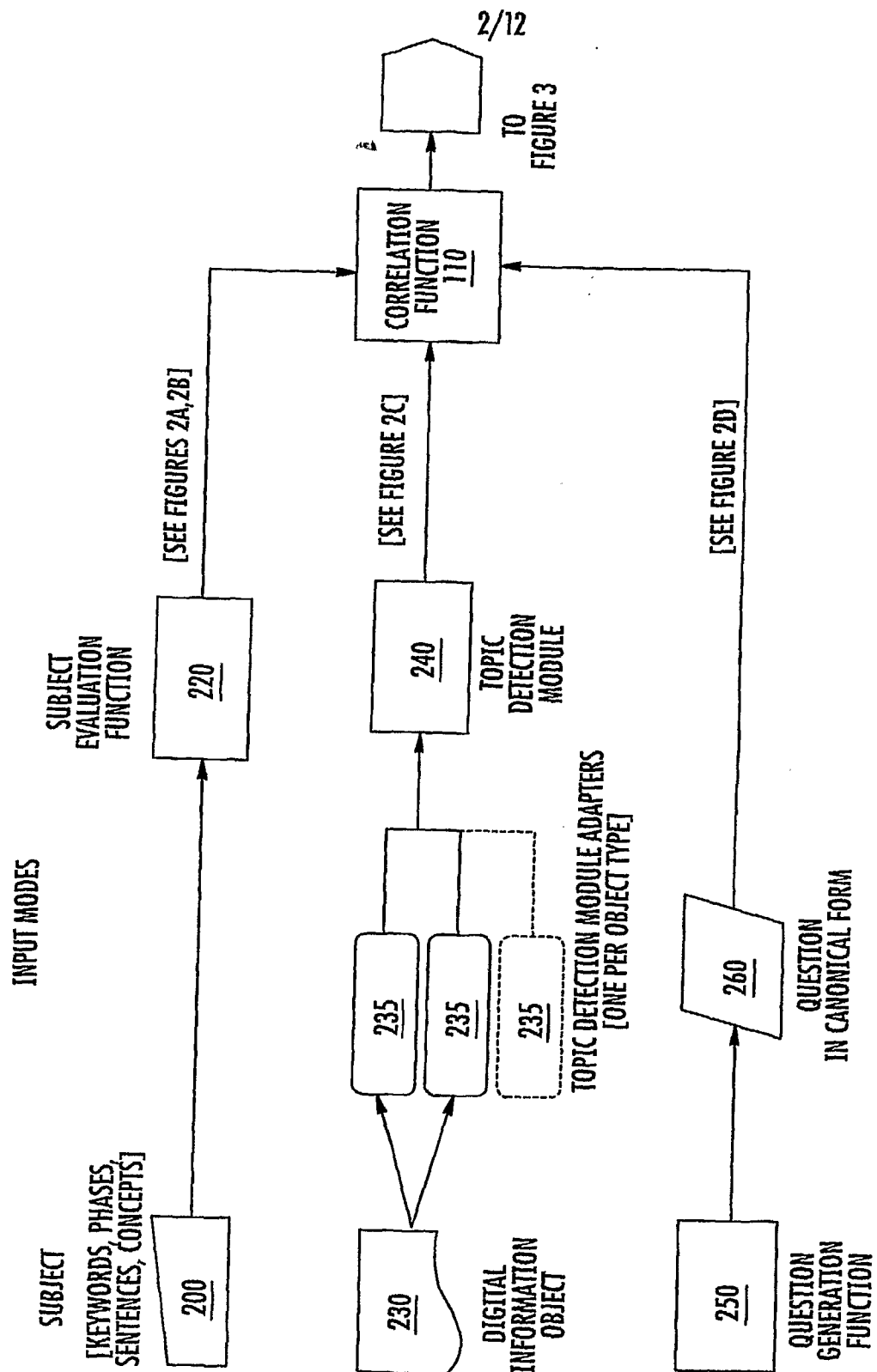


FIG. 2

3/12

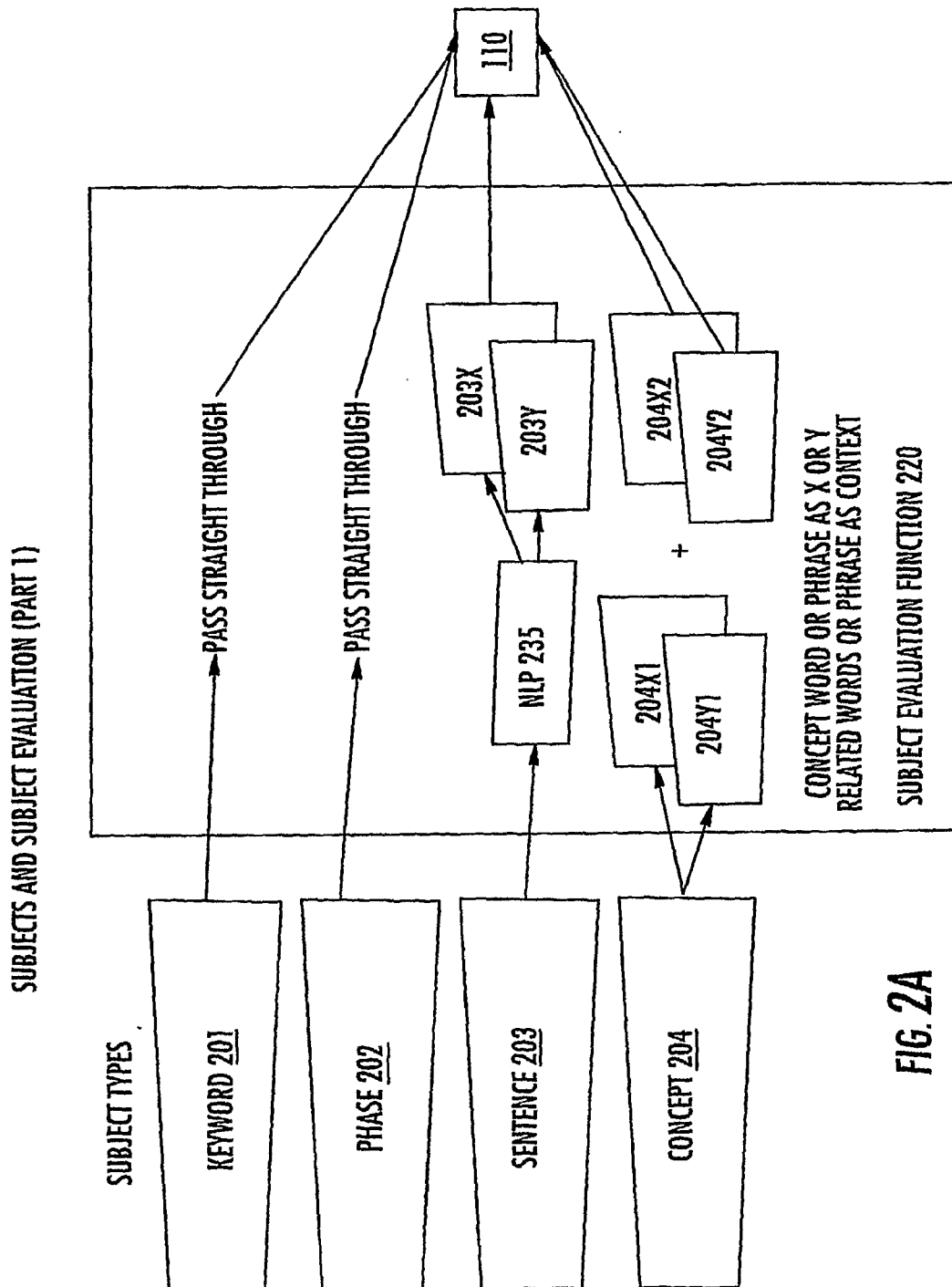


FIG. 2A

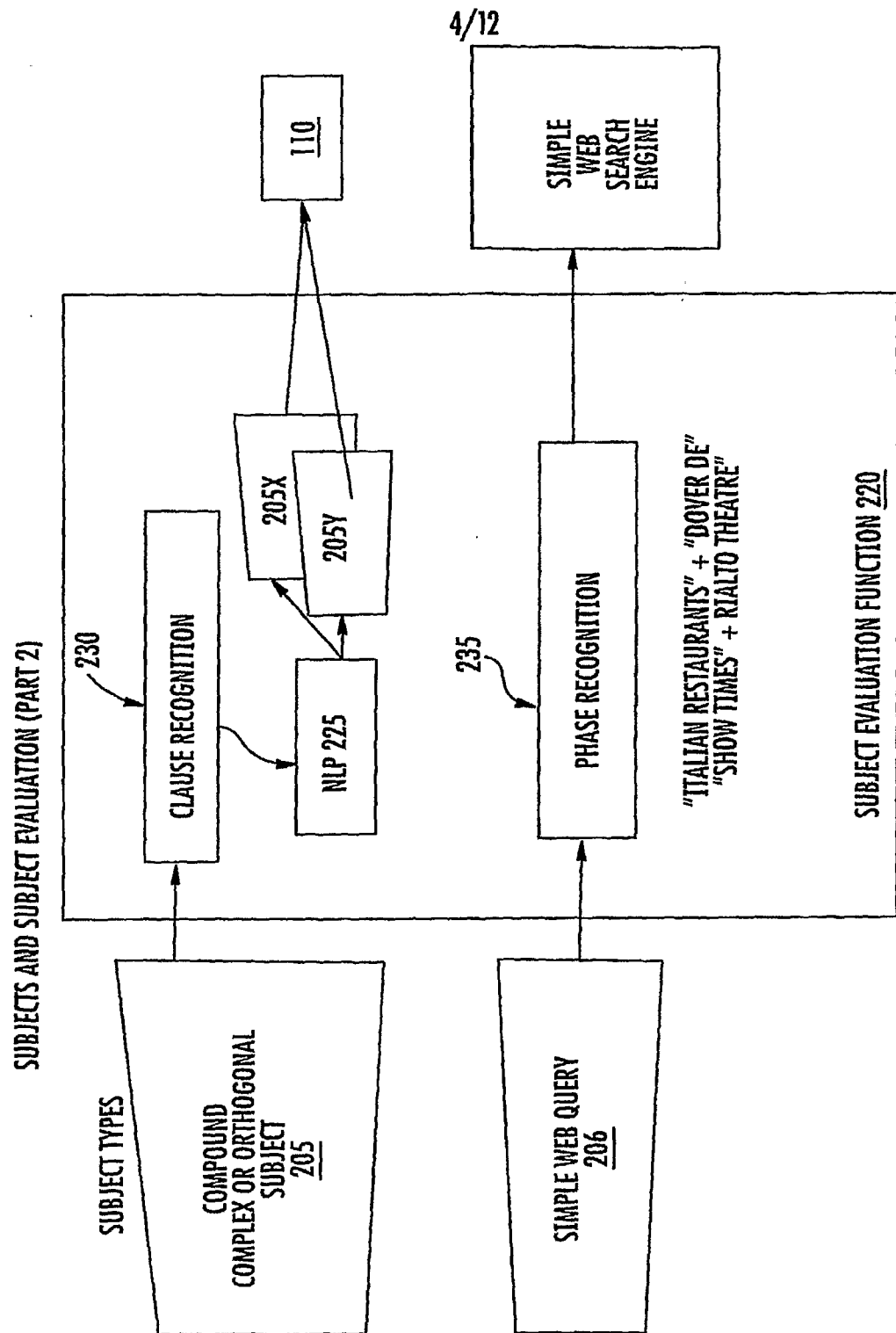


FIG. 2B

5/12

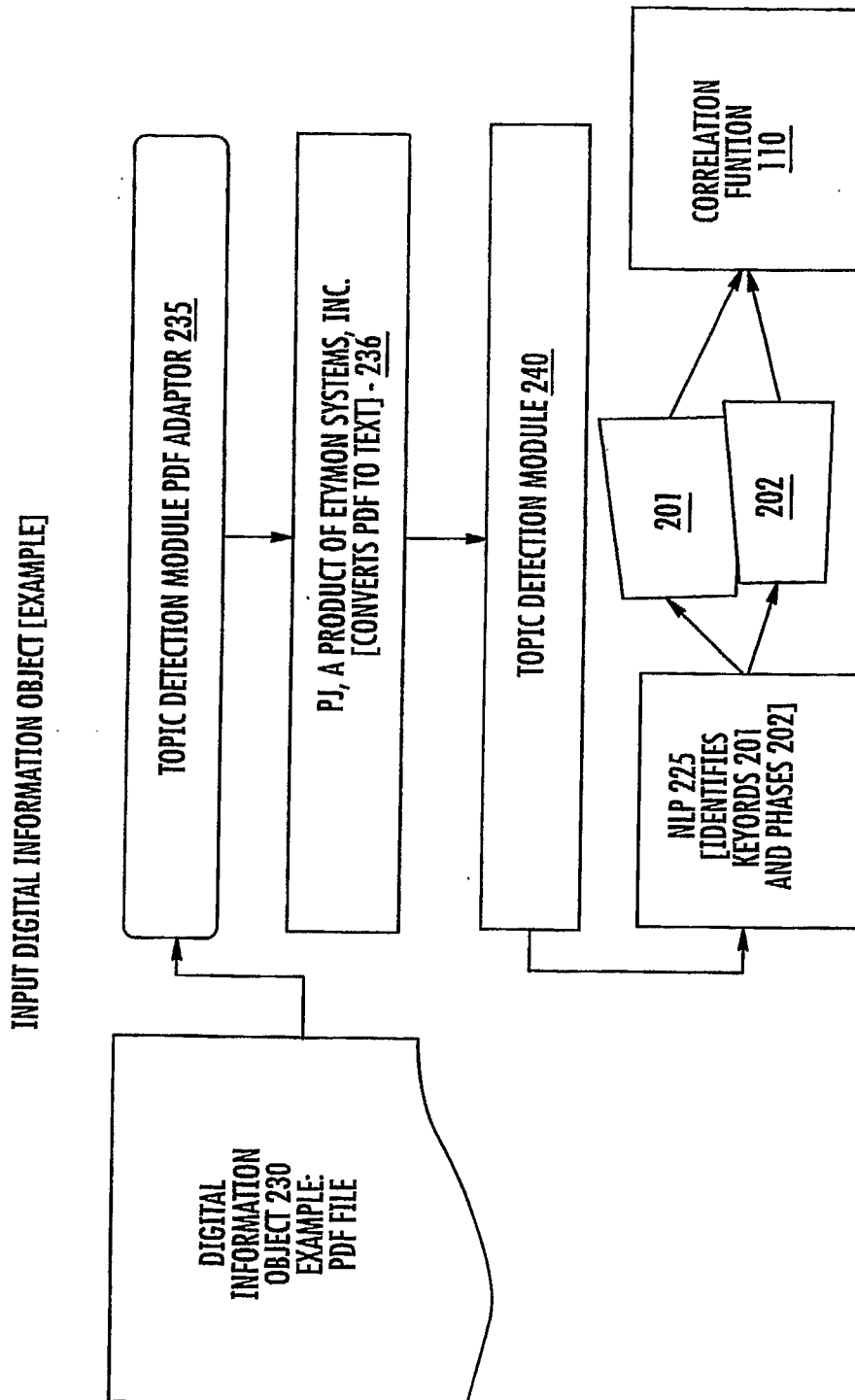


FIG. 2C

6/12

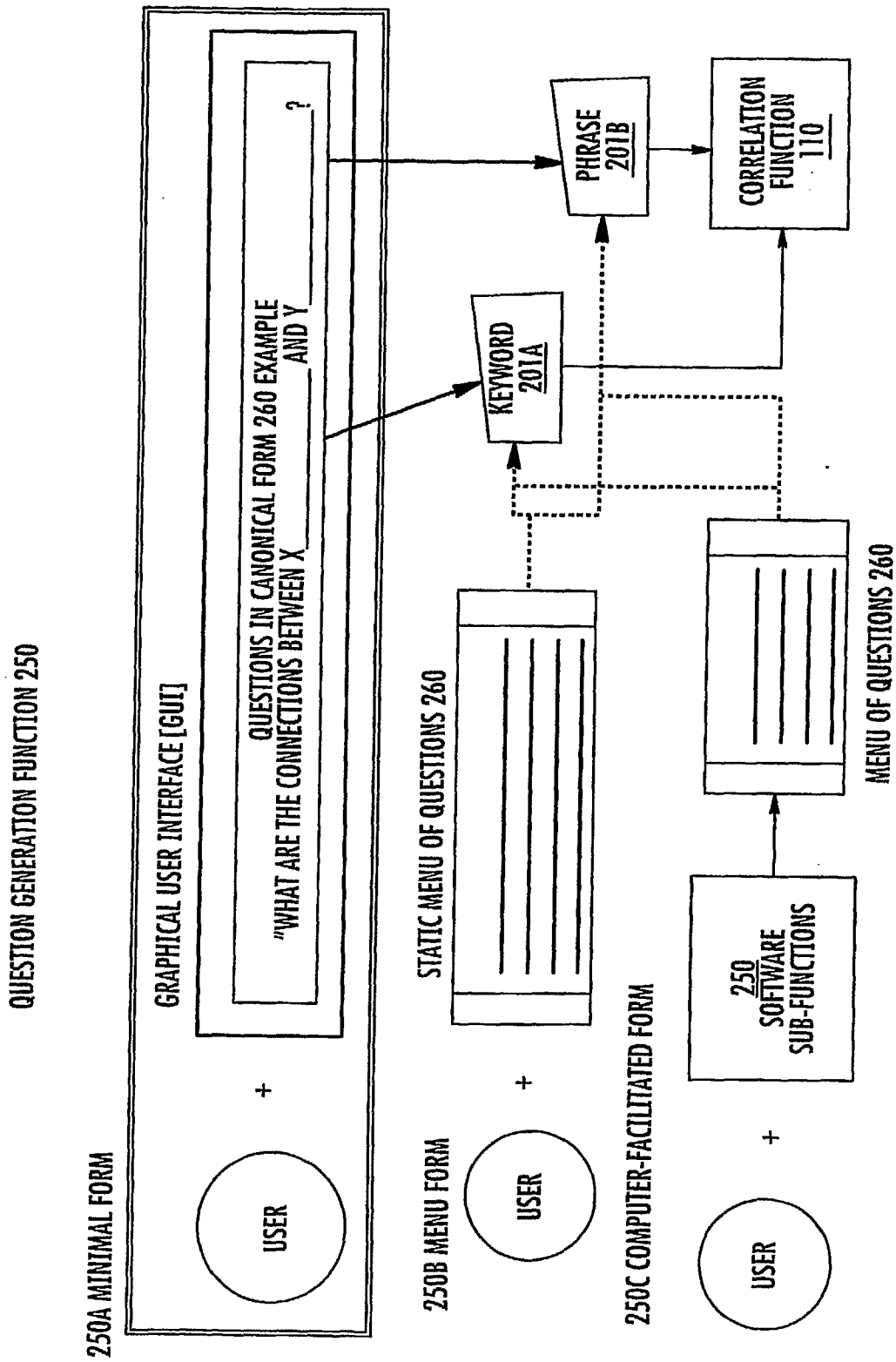
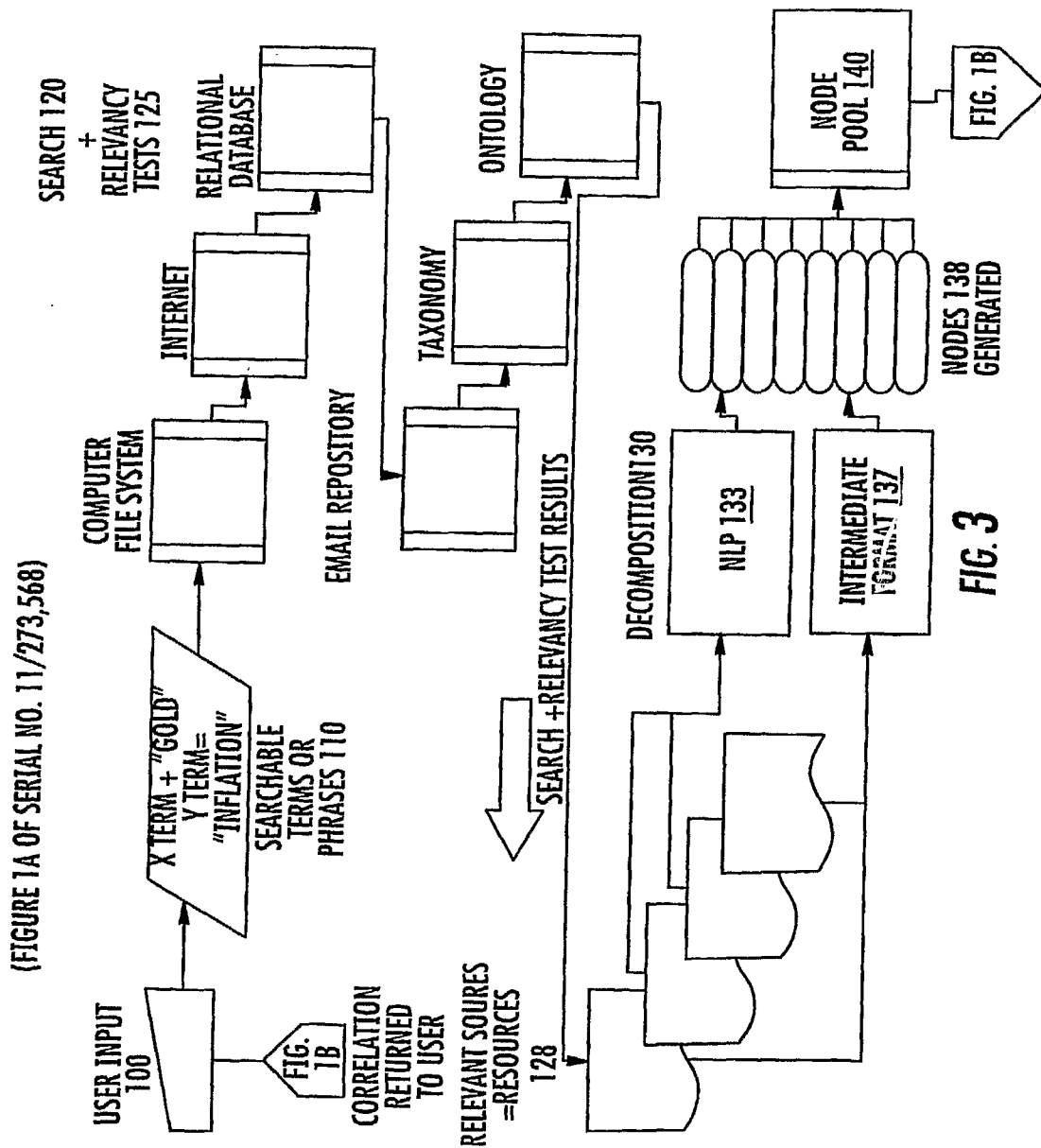
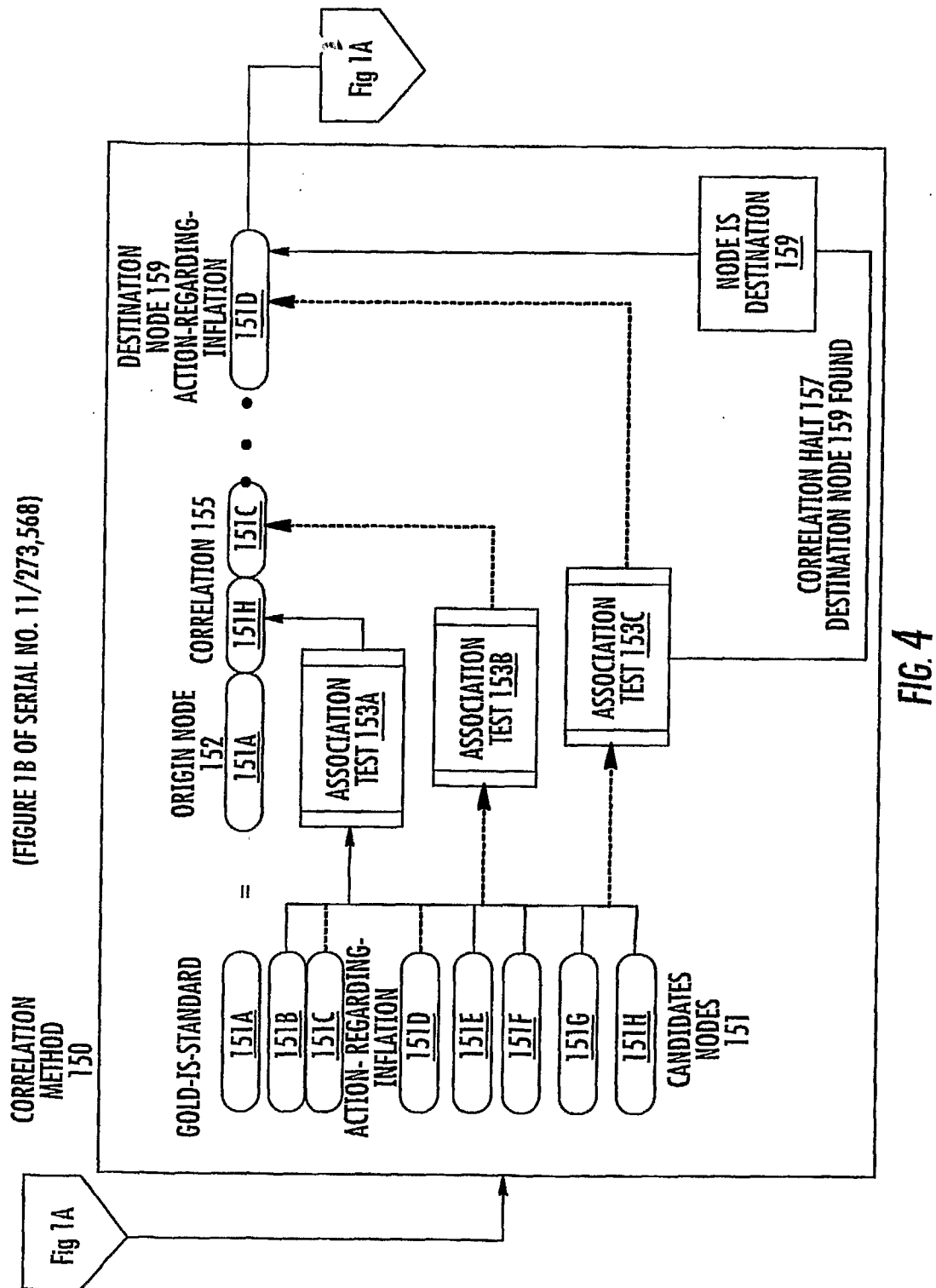


FIG. 2D

7/12





(FIGURE 1C OF SERIAL NO. 11/273,568)

NODE IMPLEMENTATIONS

NODE 180A

SUBJECT 182 (FIRST RELATUM)	BOND 184 (RELATION)	ATTRIBUTE 186 (SECOND RELATUM)
--------------------------------	------------------------	-----------------------------------

EX:

"GOLD"

"IS"

"STANDARD"

NODE 180B

SUBJECT 182 (FIRST RELATUM)	BOND 184 (RELATION)	ATTRIBUTE 186 (SECOND RELATUM)	SEQUENCE 188 (SOURCE)
--------------------------------	------------------------	-----------------------------------	--------------------------

EX:

"ACTION"

"REGARDING"

"INFLATION"

"WWW.WMHODGES/
HOME.ATT.NET/
INFLATION.HTM"

FIG. 5

10/12

(FIGURE 2A OF SERIAL NO. 11/273,568)

EXXY - STEP 1: ASK THE QUESTION

FILE VIEW OPTION HELP

FIGURE 2A

KNOWLEDGE DISCOVERY VEHICLE

EXXY

WHAT ARE THE XY CORRELATIONS?

X TERM: [] Y TERM: []

INVOICE TANGENTS (OPTIONAL)

[] [] [] []

HOW MANY ANSWER SETS?

MIN [] MAX []

SAVE CLEAR GET THE ANSWER TOOLS EX

STATUS READY

FIG. 6

11/12

(FIGURE 2E OF SERIAL NO. 11/273,568)

EXY - STEP 3: RANKXY REPORT			FIGURE 2E	
FILE	OPTIONS	HELP		
TOTAL SITE COUNT + 71 FOR GOLD INFLATION >>>			RANKXY REPORT:	
RANK	CONTRIBUTING SITE	USE COUNT		
1	http://www.abeland.org/inflation.htm	9572		
2	http://www.wmhodges/home.att.net/inflation.htm	7629		
3	http://www.goldcopd.com	5084		
4	http://en.wikipedia.org/wiki/inflation	4925		
5	http://airnow.gov/	2888		
6	http://en.wikipedia.org/wiki/terrorism	2567		
7	http://www.darksky.org/	2061		
8	http://crterrorism.org/home/	1515		
9	http://oregonstate.edu/depot/pol_ati/fag/sahr.htm	1514		
10	http://www.terror.gen.tr/english/	1471		
11	http://en.wikipedia.org/wiki/pollution	899		
12	http://www.lifesci.sussex.ac.uk/home/john_gribbin/cosmo.htm	847		
13	http://www.gold-eagle.com/	811		
14	http://www.umich.edu/~ge265/society/waterpollution.htm	631		
15	http://www.state.gov/s/ct/	336		
16	http://en.wikipedia.org/wiki/gold	275		
17	http://www.bankofcanada.ca/en/rates/inflation_cale.html	220		
18	http://www.mint.org/	177		
19	http://www.worldbank.org/	140		

FIG. 7

12/12

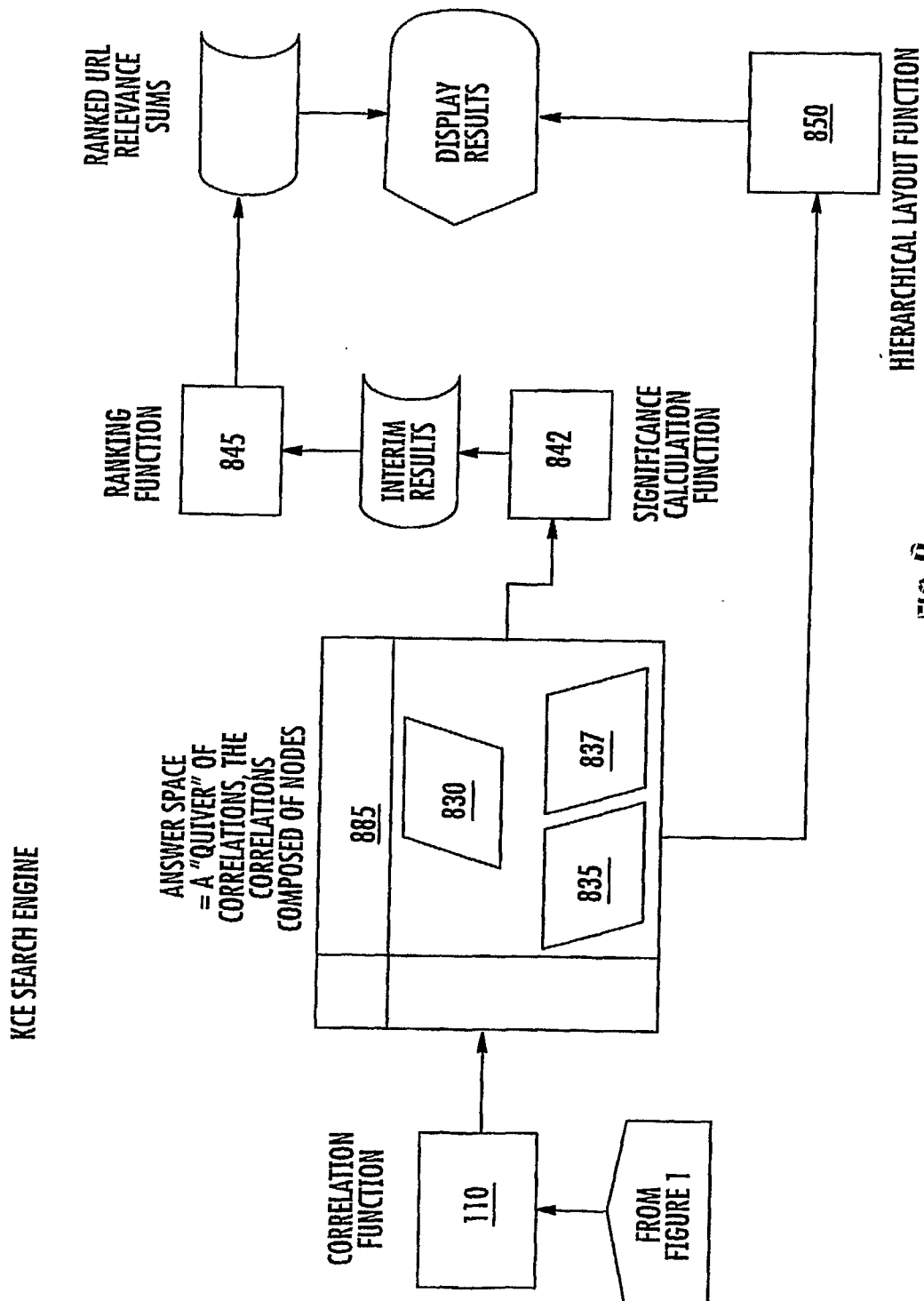


FIG. 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 06/25101

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) G06F 17/30 (2007.01)

USPC 707/2

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC(8) G06F 17/30 (2007.01)

USPC 707/2

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

USPC 707/1, 6, 100

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

USPTO WEST (PGPB, USPT, EPAB, JPAB); DIALOG PRO; Google Scholar

Search Terms Used: correlation, phrase, clause, search engine, keyword, input evaluation

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US 2002/0059220 A1 (LITTLE) 16 May 2002 (16.05.2002), entire document	1-9, 14-20 ----- 10-13
Y	US 2002/0152202 A1 (PERRO et al.) 17 October 2002 (17.10.2002), para [0050]-[0057] and Fig. 4	10-13

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

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

07 February 2007 (07.02.2007)

Date of mailing of the international search report

19 MAR 2007

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents

P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-3201

Authorized officer:

Lee W. Young

PCT Helpdesk: 571-272-4300

PCT OSP: 571-272-7774