



US007720799B2

(12) **United States Patent**  
**Kon et al.**

(10) **Patent No.:** **US 7,720,799 B2**  
(45) **Date of Patent:** **May 18, 2010**

(54) **SYSTEMS AND METHODS FOR EMPLOYING  
AN ORTHOGONAL CORPUS FOR  
DOCUMENT INDEXING**

(75) Inventors: **Henry Kon**, Lexington, MA (US);  
**George Burch**, Boston, MA (US)

(73) Assignee: **IndraWeb, Inc.**, Lexington, MA (US)

(\*) Notice: Subject to any disclaimer, the term of this  
patent is extended or adjusted under 35  
U.S.C. 154(b) by 489 days.

(21) Appl. No.: **11/707,394**

(22) Filed: **Feb. 16, 2007**

(65) **Prior Publication Data**

US 2008/0010311 A1 Jan. 10, 2008

**Related U.S. Application Data**

(63) Continuation of application No. 09/548,796, filed on  
Apr. 13, 2000, now Pat. No. 7,275,061.

(60) Provisional application No. 60/129,103, filed on Apr.  
13, 1999.

(51) **Int. Cl.**  
**G06F 17/00** (2006.01)

(52) **U.S. Cl.** ..... **707/104.1**; 707/101; 707/102;  
707/103 R

(58) **Field of Classification Search** ..... 707/1-10,  
707/100-104.1, 200-206

See application file for complete search history.

(56) **References Cited**

**U.S. PATENT DOCUMENTS**

5,463,773 A 10/1995 Sakakibara et al.

5,568,640 A	10/1996	Nishiyama et al.
5,778,363 A	7/1998	Light
5,812,995 A	9/1998	Sasaki et al.
5,963,965 A	10/1999	Vogel
6,128,613 A	10/2000	Wong et al.
6,185,550 B1	2/2001	Snow et al.
6,363,375 B1	3/2002	Hoshino et al.
6,611,840 B1	8/2003	Baer et al.
2002/0038308 A1	3/2002	Cappi

**OTHER PUBLICATIONS**

Chang et al. Customizable multi-engine search tool with clustering.  
*Computer Networks and ISDN Systems* 29, 1217-1224 (Sep. 1,  
1997).

O'Kane, K. Generating hierarchical document indices from common  
denominators in large document collections. *Information Processing  
and Management* 32, 105-115 (1996).

Sahami, M. et al. SONIA: a service for organizing networked infor-  
mation autonomously. *Proc. of Digital Libraries: Third ACM Con-  
ference on Digital Libraries*, Pittsburgh, PA 200-209 (Jun. 23-26,  
1998).

Chen, H. et al. Internet Categorization and Search: A Self-Organizing  
Approach. *J. Visual Communication and Image Representation* 7,  
88-102 (Mar. 1, 1996).

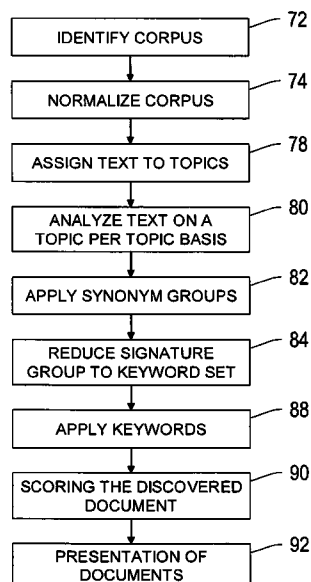
*Primary Examiner*—Isaac M Woo

(74) *Attorney, Agent, or Firm*—Ropes & Gray LLP

(57) **ABSTRACT**

Methods and systems for processing a body of reference  
material to generate a directory for accessing information  
from a database.

**10 Claims, 9 Drawing Sheets**



70

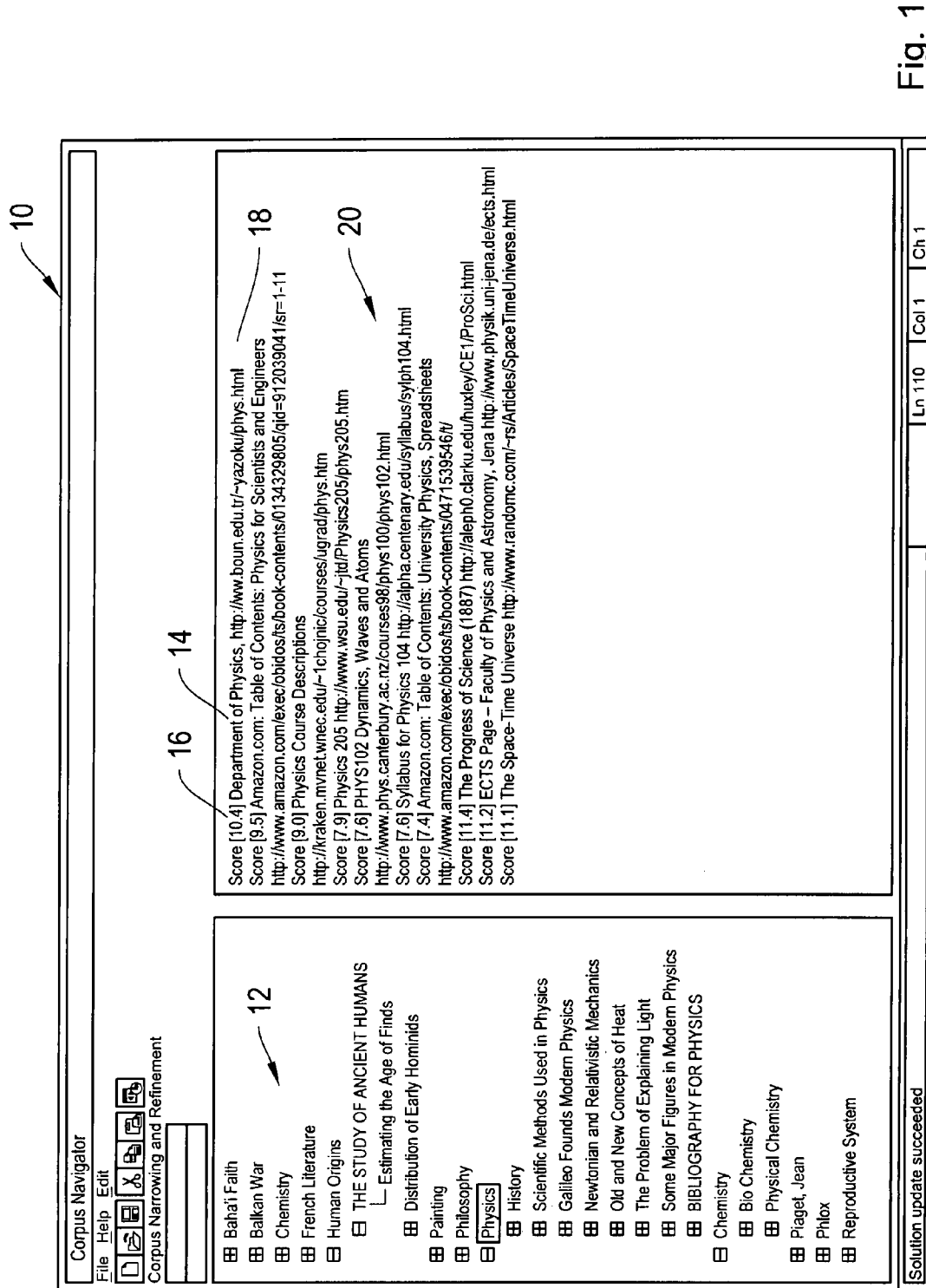
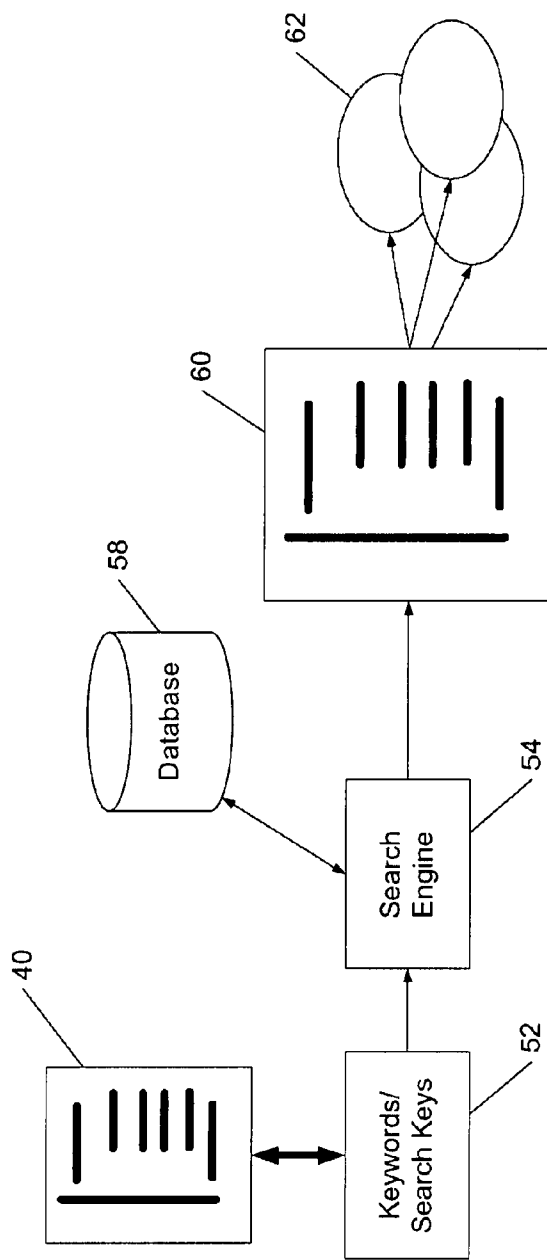
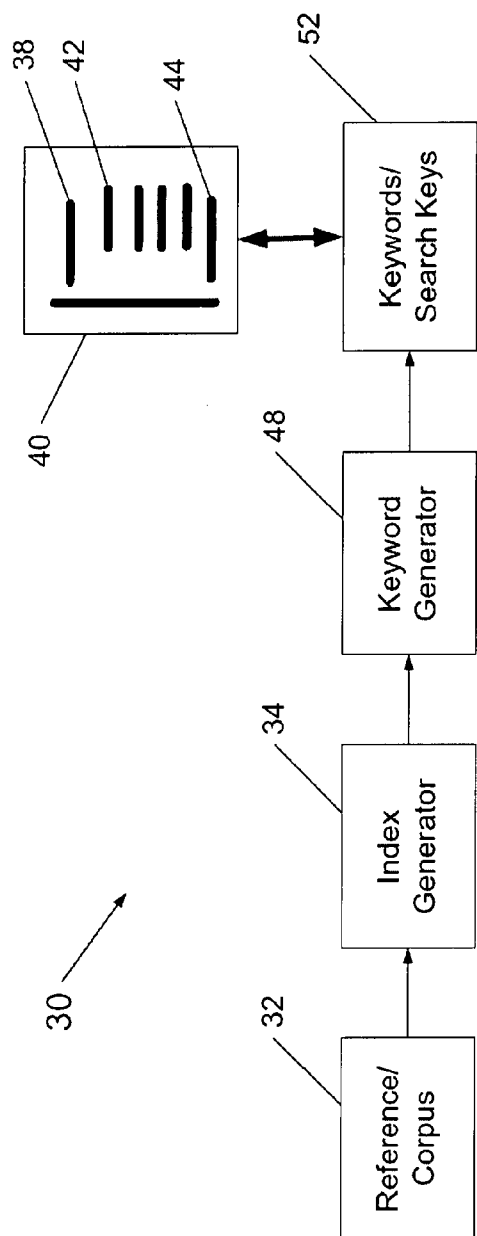


Fig. 1



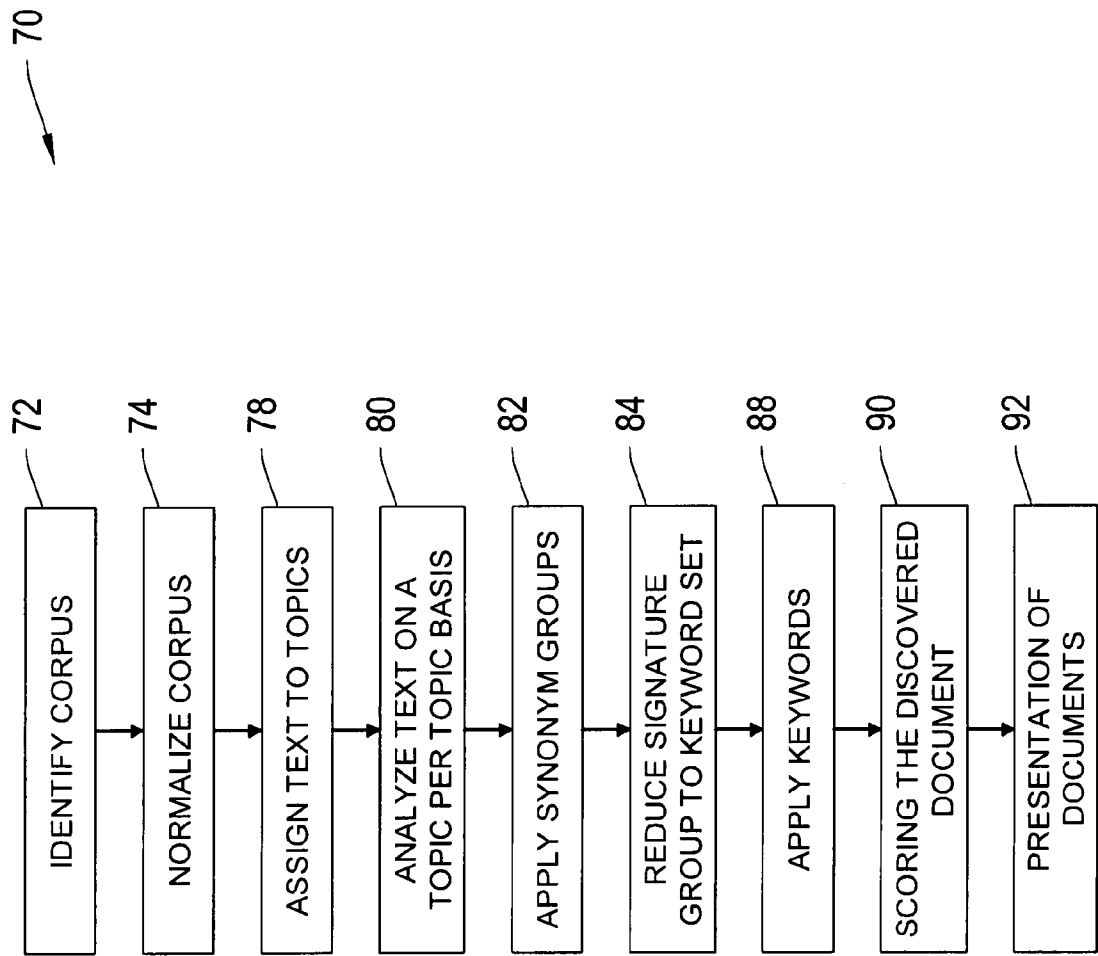


Fig. 3

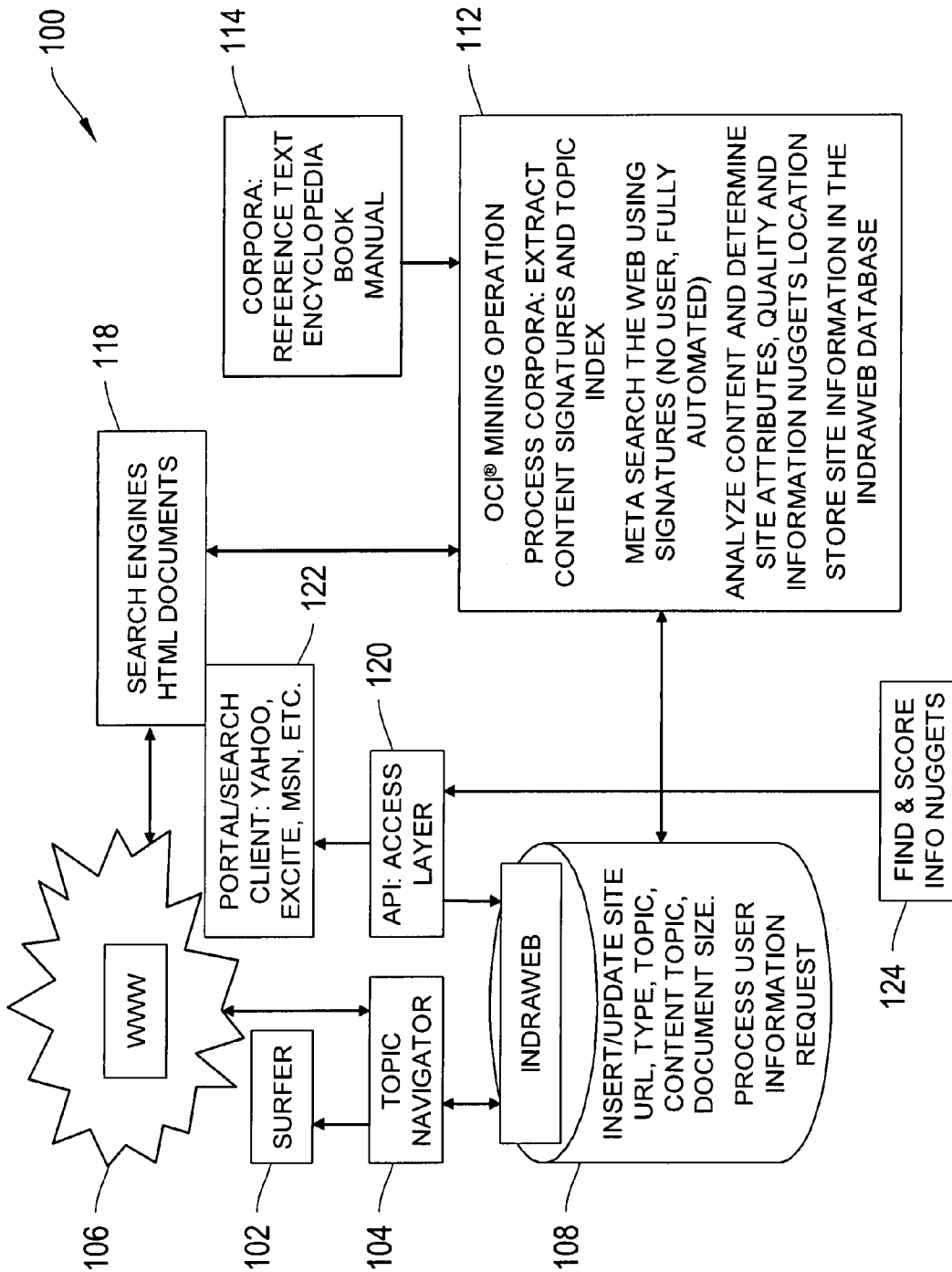


Fig. 4

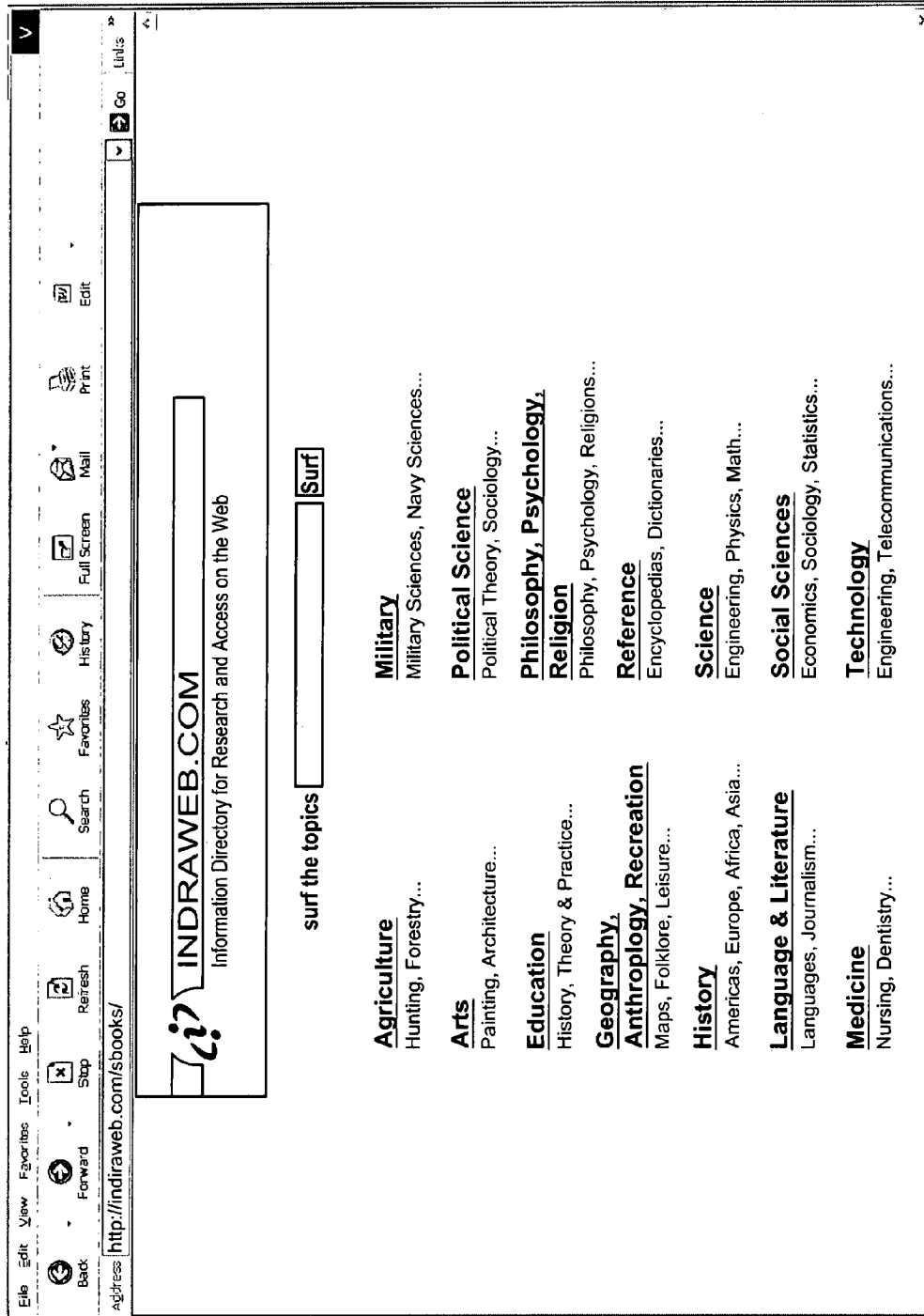


Fig. 5

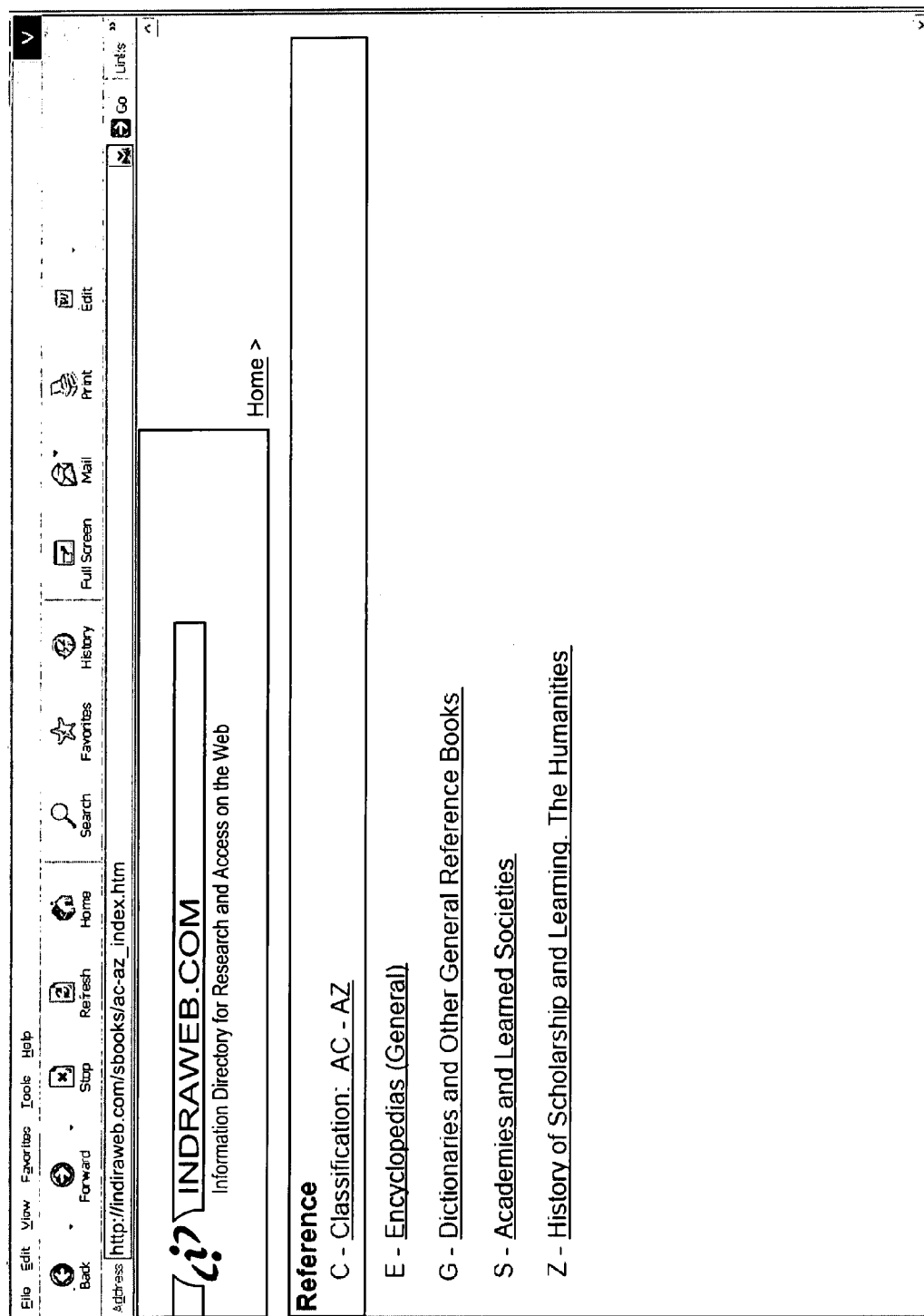


Fig. 6

File Edit View Favorites Tools Help	Back Forward Stop Refresh Home Search Favorites History Full Screen Mail Print Edit	Go Links
Address: <a href="http://indraweb.com/sbooks/ae.htm">http://indraweb.com/sbooks/ae.htm</a>		
Home > Reference >		
<b>E - Encyclopedias (General)</b>		
Omne bonum: a fourteenth-century encyclopedia of universal knowledge : British Library MSS Royal 6 E VI- 6 E VI	1996	AE2.L463 S26
Academic American encyclopedia	1998	AE5.A23
Encyclopedia Britannica	1997	AE5.B
Cambridge paperback encyclopedia	1995	AE5.C36
Cambridge paperback encyclopedia	1993	AE5.C36
Collier's encyclopedia: with bibliography and index	1997	AE5.C683
Encyclopedia Americana	1999	AE5.E333
Encyclopedia of knowledge	1991	AE5.E478
Family encyclopedia	1997	AE5.F36
Funk & Wagnalls new encyclopedia	1996	AE5.F83
Macmillan compact encyclopedia	1991	AE5.M24
Merit students encyclopedia	1993	AE5.M38
New standard encyclopedia	1998	AE5.N64
Oxford illustrated encyclopedia	1993	AE5.O94
Webster's family encyclopedia.	1992	AE5.W35
World Book encyclopedia.	1999	AE5.W55
World Book encyclopedia of people and places.	1998	AE5.W563
		H. Miller Publishers
		Grolier Inc.
		Routledge
		Cambridge University Press
		Cambridge University Press
		Collier's
		Grolier Incorporated
		Grolier
		Oxford University Press
		Funk & Wagnalls
		Macmillan
		Macmillan Educational Co.
		Ferguson Pub. Co.
		Oxford University Press
		Meriam-Webster, Inc.
		World Book
		World Book

Fig. 7



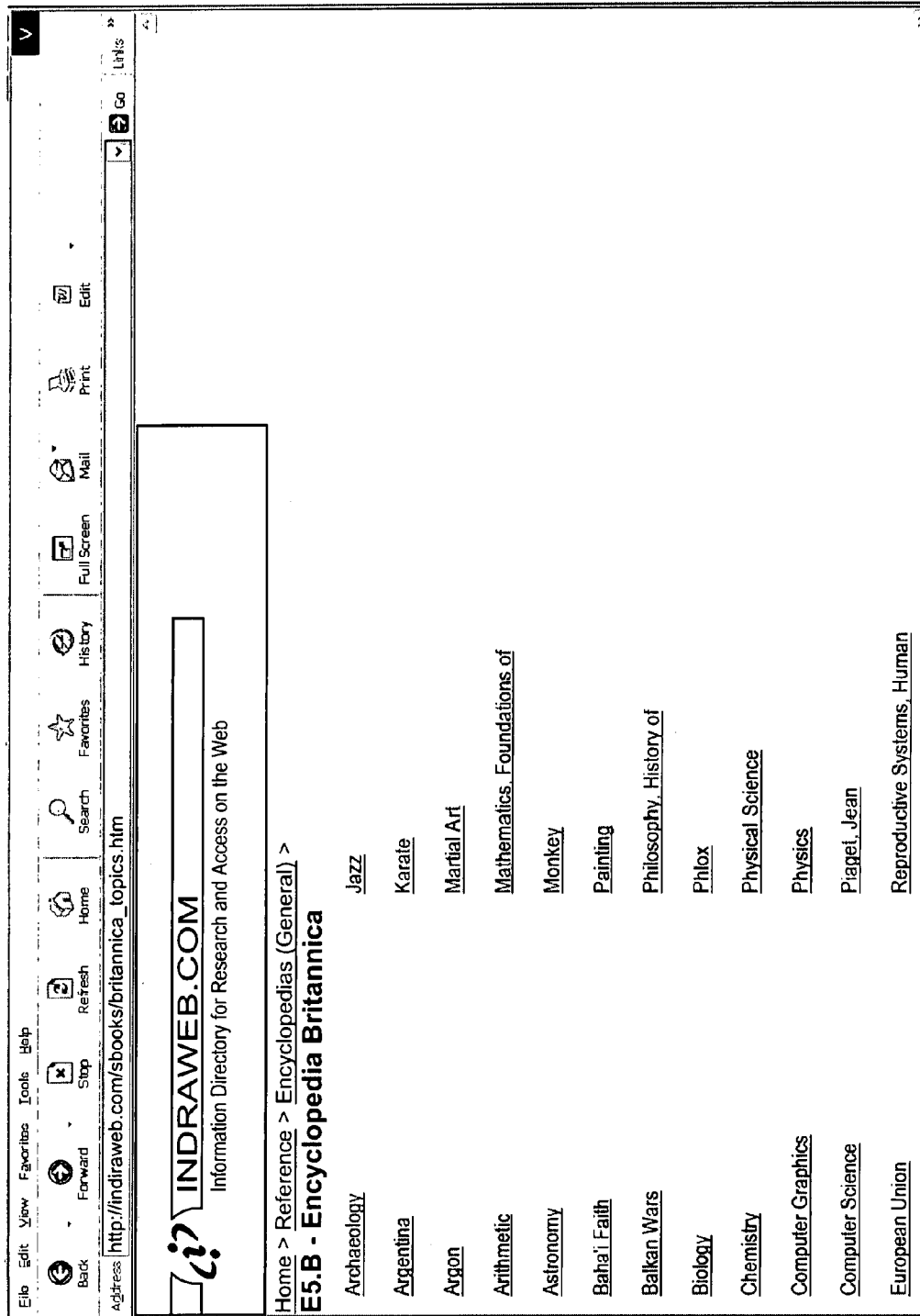


Fig. 8

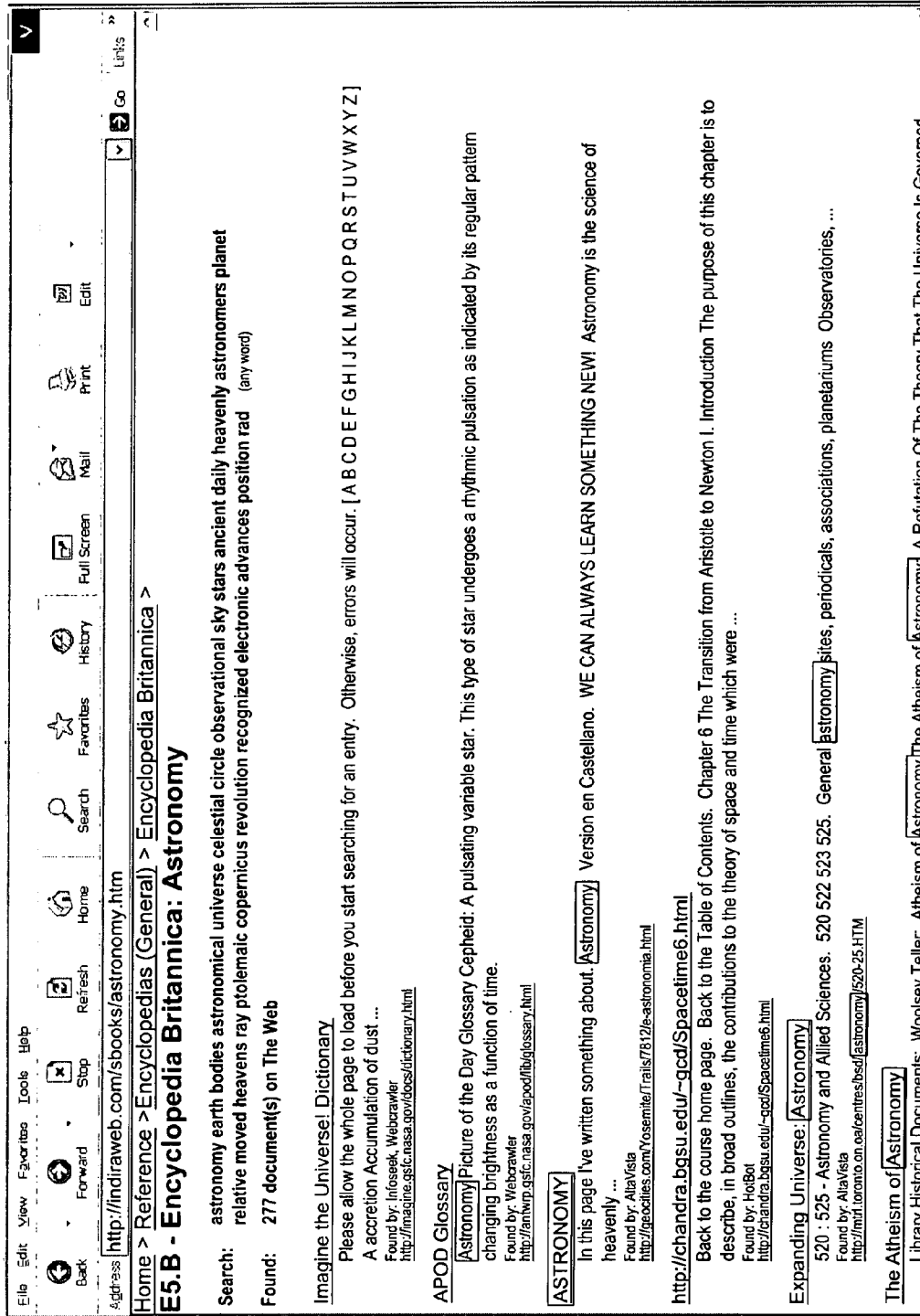


Fig. 9

1

# SYSTEMS AND METHODS FOR EMPLOYING AN ORTHOGONAL CORPUS FOR DOCUMENT INDEXING

## REFERENCE TO RELATED APPLICATIONS

This application is a continuation of U.S. application Ser. No. 09/548,796, filed on Apr. 13, 2000, now U.S. Pat. No. 7,275,061 which claims the benefit of U.S. Provisional Application No. 60/129,103, filed on Apr. 13, 1999. The entire contents of the above referenced applications are incorporated herein by reference.

## FIELD OF THE INVENTION

The invention provides systems and methods for organizing a collection of information, such as a database of documents, and more particularly, to systems and methods for classifying information with respect to a set of independent concepts, and applications of the classified information.

## BACKGROUND

The computer-based storage of information has become the norm for our knowledge-based society, and today huge databases of computer-stored information are available to researchers. For example, through the World Wide Web service, the Internet has become an enormous repository of readily accessible information. However, the sheer volume of information maintained on the World Wide Web becomes problematic when someone wishes to identify those resources, or documents that are maintained on the Internet and that are related to a certain topic. The problem is that a simple search request for information stored on the Internet and related to a certain topic returns an enormous volume of information, most of which is extraneous and irrelevant. This problem arises in part from the difficulty that existing Internet retrieval systems have with prioritizing and categorizing documents stored on the World Wide Web. Today, most systems for organizing content on the web work by searching through a document stored on the web, and associating with the document a set of key words that represent the information contained on the reviewed pages. Accordingly, documents can be organized according to the keywords that appear in the document.

However, lists of keyword, although helpful, lack an identifiable hierarchy of concept and therefore fail to provide a topic-based system for organization and document retrieval. Thus, keyword lists are less powerful tools for document research than topic-based hierarchal systems of document organization, such as tables of content, or topic indexes. Today there is a need for a "table of content" for the World Wide Web. Such a table of content would provide a more powerful metaphor for using the Web as a research tool and would provide a flexible platform on which other research tools could be built. As such, there is a need for improved systems for organizing information stored in a computer database, such as the World Wide Web.

## SUMMARY

The systems and methods described herein provide systems for document indexing and scoring of content on a computer database, such as the World Wide Web. The systems generally include an orthogonal corpus that may comprise a collection of blocks of text, and that may be employed to

2

index and score textual information for applications in retrieving, classifying, or browsing over a set of documents.

An orthogonal corpus, as the term is employed herein, may be understood to encompass, without being limited to, any collection of blocks of text that are outlined or referenced by a table of contents, topic index, chapter heading or other topical indicia where each topic either stands alone or is an identified subpart (subtopic) of another topic, forming a tree of topics and their descendant subtopics. Encyclopedias, text and reference books, periodicals, web sites, dictionaries, thesauri, the library of congress, the Dewey decimal system, and glossaries are examples of, surrogates for, or extenders of orthogonal corpora. A set of topics is understood as orthogonal in the sense that substantially every member topic (e.g., chapter or article) covers a different concept or substantially different concept than any other topic under the same ancestor topic in the tree. A topic in one practice may be assumed to include or not include, its ancestor topics or underlying subtopics.

The orthogonal construction, or decomposition, of a corpus provides for algorithmic identification of keywords in each topic, which distinguish it from its sibling, cousin, ancestor, or descendent topics. Keywords may be employed to numerically score over an underlying pool of documents. Alternatively, if more practical for a given application, such as when working with a large document collection, rather than individually score all documents in the collection against the keywords, then a set of search words may be generated to identify a subset of candidate documents for scoring.

Parameters employed during the scoring process may relate to the identification of keywords and their refinement into a set of search words, with weightings for associated document or sub-document scoring. The scoring, search term, and keyword methods may employ conventional information retrieval techniques including the use of synonyms, stemming, frequency, proximity, stop words, hyponyms, and synonyms.

For purposes of clarity certain terms will now be described, although the understandings set forth are not to be understood as limiting and are only provided for purposes of achieving clarity by way of providing examples. The term "word" as employed herein may be understood to encompass a lexical type found in a common or specialty dictionary of any language. The term "phrase" as employed herein may be understood to encompass any sequence of one or more words. Heretofore for simplicity we use "Word" to mean "Word or Phrase." The term "synonym group" shall be understood to encompass a set of words which may be used as alternates for a given word. Each word in a synonym group has a similar or identical meaning. The term "topic" shall be understood to encompass textual content typically having a title, having corresponding text, concerning a single topic, or covering a set or tree of related subtopics. The term "subtopic" shall be understood to encompass a block of text within a Topic. Typically, a subtopic may be identified by its subtopic header or other outline indicator. In different calculation contexts, subtopics may or may not be included as part of parent topics. The term "word count" may be understood as an integer count of the number of times a word or a word in its synonym group occurs in a given topic or text area, potentially including text in the title and headers and any text elements in that text. The term "word frequency" may be understood to encompass the word count in a text area divided by the number of words total in that text. A word map is a representation of textual content within a text area that is more precise than a word count. A word map may describe a word's relative location in the text, its linguistic type or contexts of use, its prominence indicators

such as use in a title or highlighting fonts. The term orthogonal corpus may be understood to encompass a collection of topically organized information referenced by a table of contents and/or index, where each topic is clearly identified as a SubTopic of a topic or else stands alone. Nodes in the table of contents tree may represent topics. The information may be understood as orthogonal in the sense that a stand-alone topic (e.g., chapter or article) covers a substantially different concept than any other stand-alone topic, and any subtopic expresses a substantially different concept from any other subtopic within the same parent topic. The term document may be understood to encompass formatted textual content with topic beginnings, endings, and marked hierarchy. A document may contain one or more topics and may include subtopics. A corpus may include one or more documents. The relationship between documents and topics is not mandated, though in some embodiments each document to represent one top-level topic along with its subtopics. The term "discovered document" may be understood to encompass a document (or a set of documents such as a web site or portion of a web site) which is being scored. Scoring of a discovered document may be relative to one or more corpus documents or corpus topics. In one practice, scoring measures the degree of topical relevance to the corpus topics. The discovered document will often be a member of a search result set.

More particularly, the systems and methods described herein include methods for processing a body of reference material to generate a directory for accessing information from a database. These methods may comprise processing the body of reference material to identify a hierarchical organization of a plurality of topics. Additionally, the processes may include the step of associating with at least one of the topics a portion of the reference material and processing the assigned portion of reference material to generate a plurality of search keys representative of search strings for selecting information from the database. The process may then apply the search keys to the database to retrieve information from the database and may create an association between the at least one topic and the information retrieved from the database.

In an optional step, the methods described herein may create a graphical interface that is representative of the identified hierarchical organization of a plurality of topics for allowing a user to access information retrieved from the database and having an association with the topic. Accordingly, the user may be provided with a graphical interface that allows the user to activate, typically by clicking with a mouse, a graphical representation of a topic to identify a set of links to content, such as web pages that are associated with the topic selected by the user.

In one practice, processing the body of reference material includes processing a body of reference material that has been selected from the group consisting of an encyclopedia, a dictionary, a text book, a novel, a newspaper, or a website. Processing the material may include identifying a hierarchical organization of a plurality of substantially orthogonal topics. This may include identifying a table of contents for the body of reference material, identifying an index for the reference material, identifying chapter or subchapter headings within the reference table, identifying definition entries within a dictionary, and other similar operations that identify different topics that occur within the reference material.

Optionally, when processing a body of reference material, the process may normalize the identified hierarchical organization of the plurality of topics.

In one practice when processing the assigned text, the process includes a step of generating a word map that is

representative of a statistical analysis of the words contained in the assigned text. Generating the word map may include performing a word count process for determining word frequency of a word within the assigned text and for employing the word frequency for determining the relevance of a word to the associated topic. Processing the assigned text for different topics may also include a step of identifying a set of key words that have an associated measure intra-document orthogonality.

In an optional step, processing the assigned text may include identifying a set of synonyms for extending the search keys. Further, a subset of search keys may be selected that have a predetermined measure of correlation to the topic. The search keys may be applied to the database such as through an Internet search engine, to discover documents that are related to the search keys. Optionally, the Internet search engine may be a meta-search engine.

Once documents have been discovered from the database that are related to the search keys, documents may be further processed to determine their relationship to the topics associated with the search keys.

In the methods described herein creating an association between the at least one topic and the information retrieved from the database may include capturing a location pointer that is associated with the information retrieved from the database. Creating that association may include generating a data structure for the topic which allows storing location pointers that are associated with information retrieved from the database.

In another aspect, the systems and methods described herein include systems for organizing a collection of documents. Such systems may comprise an orthogonal corpus of information that is arranged according to an index of topics, a keyword generator for generating a set of keywords representative of a documents associated with a topic in the index of topics, a scoring system for processing documents within the collection of documents to associate with at least a portion of the documents a score representative of the document's association to a particular topic, and a graphical representation for depicting at least a portion of the index of topics and having respective portions of the graphical representation linked to documents associated with a respective portion of the index of topics.

In a further embodiment the systems of the invention may include systems for extending the content of the document. These systems can include a parser for selecting terms within the document to be extended, an orthogonal corpus of information arranged according to an index of topics, a keyword generator for generating a set of key words representative of a documents association with a topic in the index of topics, and a linking system for processing the documents within a collection of documents, to associate with at least a portion of the document, a score representative of the documents association to a particular topic and for providing the first document with links to the collection of documents for extending the content of that document.

Accordingly, the systems and methods described herein may leverage the electronically stored content of the World Wide Web in an intelligent and meaningful way, to provide a database of content organized under an orthogonal and hierarchical index of topics and subtopics.

5

Other objects of the invention will, in part, be obvious, and, in part, be shown from the following description of the systems and methods shown herein.

#### BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing and other objects and advantages of the invention will be appreciated more fully from the following further description thereof, with reference to the accompanying drawing wherein;

FIG. 1 depicts a screen shot of a portion of a orthogonal corpus and a set of documents and the scores associated with those documents;

FIGS. 2A through 2B depict a dataflow diagram of one process according to the invention for processing a body of reference material for organizing a collection of documents according to a hierarchical arrangement of topics provided by the reference material;

FIG. 3 depicts one flow chart diagram of a process according to the invention;

FIG. 4 depicts one system according to the invention; and

FIGS. 5-9 depict a further practice according to the invention organizing content according to indices generated from a plurality of references.

#### DETAILED DESCRIPTION

To provide an overall understanding of the invention, certain illustrative embodiments will now be described. However, it will be understood by one of ordinary skill in the art that the systems and methods described herein can be adapted and modified to provide systems and methods suitable for other applications and that other additions and modifications can be made to the illustrated embodiments without departing from the scope hereof.

It is a realization of the invention that the indexing and cataloging of content on the Internet, as well as from other stores of information, may be performed by applying a process that employs an orthogonal corpus, or corpora, of information, such as an Encyclopedia. To this end, the processes described herein identify the topics discussed within the corpus. The process also identifies within the corpus a set of keywords that are relevant to the topics presented in the corpus. The keywords associated with a topic may be employed to identify documents stored in another database that are related to the topic. A graphical representation of the index of topics found in the corpus may then be generated, with individual topics operating as links to these related documents. Thus, a user interested in reviewing content in the corpus related to a certain topic, may also activate a link in the graphical representation of the index to access other documents that have been identified as related to the topic of interest to the user.

Turning to FIG. 1, there is depicted a graphical user interface 10 of the type created and employed by systems according to the invention. The graphical user interface 10 represents a topic index 12, a portion of which is shown in this illustration. The topic index 12 may be a graphical representation of the table of contents of an encyclopedia, or other corpus. A user may employ the graphic interface 10 to access information that relates to the different topics listed in the index 12. Additionally, the depicted index 12 includes topics and subtopics, including subtopics of the same ancestor topic. For example, in FIG. 1, the topic Human Origins is the ancestor topic for the subtopics, The Study of Ancient Human and

6

the Distribution of Early Hominids. A topic, or a subtopic, may be understood to include, optionally, its ancestor topics or underlying subtopics.

The graphical representation of the index 12 may include a hypertext link, or other linking mechanism, for each topic or subtopic in the index 12. For example, the user may activate the links, as depicted by the highlighted topic PHYSICS in FIG. 1, to retrieve a group of documents having content that is associated with the selected topic. As further depicted by FIG. 1, the system 10 may provide a display 20 such that for a selected topic or a subtopic, such as the selected topic Physics, a document 18, or a plurality of documents 18, may be presented to the user as documents associated with the topic. In the depicted embodiment, a pointer to the document, such as the title and URL 14 may be presented to the user. Additionally, an associated numerical score 16, that represents that document's association to the topic may also be presented. The development of such scores 16 will be described in more detail hereinafter. Optionally, all the documents associated with a topic may be displayed in a window 20 of the system 10.

Turning now to FIGS. 2A and 2B, dataflow diagrams are presented that illustrate one process for creating a graphical interface, such as the interface 10 of FIG. 1. Specifically, FIGS. 2A and 2B depict a process 30 wherein a corpus, such as an existing published book of reference material, is processed by an orthogonal corpus indexing (OCI) process that extracts content signatures and topic indices from the corpus' content. The depicted process employs the content signatures to generate search strings for search engines to identify content associated with topics described in the corpus. The retrieved or discovered documents may be examined for content relevance and the relevant documents may be associated with topics presented in the orthogonal index of the corpus. Optionally, site attributes such as document type, timeliness, source and other such attributes may also be identified and employed to select relevant websites that may be associated with a topic in the index of the orthogonal corpus.

More specifically FIG. 2A depicts that the process 30 operates on a corpus 32 that may be input to the index generator 34. The index generator 34 may generate an index for the corpus 32 and this index may be provided to the keyword generator 48. The keyword generator 48 may produce a set of key words 52 and may be associated with the index 40. The process 30 continues in FIG. 2B which shows the index 40 in the search keys 52 being applied to a search engine 54. The search engine 54 discovers documents from a database of content, or from a collection of databases of content 58 to thereby create an association between at least one of the topics of the index 40 and the information retrieved from the database 58.

The depicted corpus 32 may be any collection of information and may include, but is not limited to, encyclopedias, text books, dictionaries, thesauruses, atlases, maps, and other reference material. In one embodiment, the corpus 32 may be a published book that may be turned into or stored in an electronic format such as a conventional computer data file of text information. The corpus 32, preferably in an electronic format, may be provided to the index generator 34. The index generator 34 may process the corpus 32 to identify a hierarchical organization of a plurality of topics that appear within the corpus 32. To this end, the index generator 34 may decompose the corpus 32 to create a standard hierarchical topic orientation that is capable of assigning text content to title, headers, topics, subtopics, or any other device that may be employed for representing a section of text related to a topic, meaning, category, or some other similar abstraction.

U.S. Pat. No. 5,963,203 entitled "Automatic index creation for a word processor" issued to Sotomayor, Bernard, describes methods that may be employed by the index generator **34**. For example, Sotomayor describes methods that enable scanning one or more documents to automatically identify key topics and phrases in a document's text, as well as methods to generate an index to those key topics. Similarly, U.S. Pat. No. 5,819,258 entitled "Method and apparatus for automatically generating hierarchical categories from large document collections", by Vaithyanathan, Shivakumar, Travis, Robert, and Prakash, Mayank, further describes techniques that may be employed by the index generator **34** for determining an index for a corpus. Other techniques known in the art may also be employed by the index generator **34** without departing from the invention.

In an alternative practice, the index generator **34** allows an operator to identify the type of corpus **32** being input into the index generator **34**. For example, the index generator **34** may present an interface to the operator that allows the operator to identify whether the corpus being presented comprises and encyclopedia, a dictionary, a textbook, or another known type of reference document. Additionally, the index generator **34** may allow the operator to identify whether the corpus **32** includes a table of contents, an index, chapter heading, or any other representation of the different topics contained within the corpus. In this embodiment, the user may identify, for example, that the corpus **32** comprises an encyclopedia and that the encyclopedia includes a table of contents that is representative of the index of orthogonal topics maintained within the encyclopedia. In this embodiment, the index generator **34** may process the presented corpus **32** to identify the table of contents for the encyclopedia. This table of contents, in one embodiment, may be formatted into an HTML document that presents the table of contents in an organized format that emphasizes the topics, subtopics and other hierarchical structure of the table of contents. In one process the index generator **34** processes the notation for the table of contents, such as the topic numbering employed, to identify which topics are understood as parent topics and which are understood as main topics and which are understood as subtopics. In a further optional embodiment, the index generator **34** may present the generated index with the orthogonal corpus **40** to the operator to allow the operator to edit or amend the generated index for the orthogonal corpus **40**.

As shown in FIG. 2A, once the index generator **34** has processed the corpus **32** the index generator **34** may present the index **40** for the corpus **32** to the keyword generator **34**. The index **40** may comprise a hierarchical representation of the orthogonal topics maintained within the corpus **32**. This hierarchical representation may include primary topics, such as the depicted topic **38** and a plurality of subtopics **42** that are associated with the primary topic **38**.

The keyword generator **48** in one embodiment operates to identify sections of text of the corpus **32** to be associated with the different topics and subtopics of the index **40**. Continuing with the above example, in those practices where the index **40** is generated from the table of contents for the corpus **32**, the keyword generator **48** may identify those pages that contain information associated with a topic presented within the index **40**. For example, the keyword generator **48** may process the table of content for the corpus **32** to identify a page number associated with a topic, such as the topic **40** and may analyze the page associated with topic **40** to identify that portion of the page that may be associated with the topic **40**. In one embodiment, where headings are presented within the corpus **32**, the keyword generator **48** may analyze the page associated with the topic **40** to identify a heading that is

representative of the beginning of the presentation within the corpus **32** of information that is associated with topic **40**. For example, the keyword generator **48** may identify a section of text within the associated page that contains the information associated with topic **40** and that is presented in a type font and font size that is representative of a heading. In a subsequent step the keyword generator **48** may identify the location of the heading for the subsequent topic **44** that indicates the beginning of content related to the new topic. The keyword generator **48** may identify the content that is delimited by the heading **40** and **44** and associate that content as content related to the topic **40**.

Once the portion of the corpus **32** that is to be associated with the topic **40** is identified, the keyword generator **48** may process this assigned portion of text to generate a plurality of search keys, each of which may be representative of a search string for selecting information from a database.

In one embodiment, the system **10** employs the orthogonal construction of the corpus for algorithmic identification of keywords in each topic that distinguish that topic from its sibling, cousin, ancestor, or descendent topics. Accordingly, the systems described herein may create a set of keywords for a topic that identifies a document associated with a topic and that are keywords which may act to distinguish documents associated with one topic, from documents associated with another topic. For example, the system **10** may employ processes that identify keywords that are associated strongly with a particular topic. Techniques for creating keywords will be understood from Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A. (1990), "Indexing by latent semantic analysis." *Journal of the Society for Information Science*, 41(6), 391-407. Additionally, the system **10** may identify other keywords that act to disassociate a document from one or more other topics. These keywords may be employed by the system **10** to numerically score over an underlying pool of documents.

The system **10** may employ scoring methods that may utilize traditional information retrieval techniques including the use of synonyms, stemming, frequency, proximity, stop words, hyponyms, and synonyms. If, as in most large document collections, it is not practical for all documents to be individually scored against the keywords, then a subset of search words is selected to identify candidate documents for scoring. Keyword and search terms are identified based on a numerical method that apportions words among topics. The goal is for the keywords and search terms to identify individual blocks of text as found at the nodes of the orthogonal corpus topic hierarchy. In an ideal sense, the keywords would be partitioned across the hierarchical tree nodes, with each word occurring in only one corpus topic. In addition to word rarity among corpus topics, rarity in the underlying document pool may contribute to a word being identified as a keyword or search term for a given topic. For example, a keyword occurring in only one node and only once on the Web, would be a top candidate as a keyword and search term.

The keyword generator **48** may present as an output, a set of keywords **52** each of which may be associated with a topic or subtopic in the index **40**. As described above, these keywords may be employed to act to distinguish documents associated with one topic from documents associated with another topic. Accordingly, as depicted in FIG. 2B the search keys **52** and associated topics on the index **40** may be presented to the search engine **54** for retrieving information from a database or databases of content **58**. To this end, the process **30** applies the search keys to the database **58** to retrieve information from the database **58**. In one practice as will be described in more detail hereinafter, an optional step in pro-

cess 30 is performed wherein the search keys 52 are processed to identify a subset of search keys that may be employed for generating search queries to one or more search engines, such as Internet search engines, to discover a set of documents 60 which are relevant to the topic of interest. Each of the resulting documents 60 may be examined in a subsequent step to determine the relevance of the content contained within the index. The relevance may be scored, as further described below, for identifying the relevance of that document, and the score may be employed for ordering the sequence in which content is listed as being relevant to a particular document.

Once the discovered documents are scored for relevance, the process 30 may associate portions 62 of the discovered documents to associated topics within the index 38. In a practice wherein the database 58 includes links to URL's for websites, the process 30 may create a web database that contains website information such as URL's, types, dates, topics, contents, size and editor notes that are inserted or updated in the database from time to time. Information about the corpus 32 that has been processed, such as the publisher, the ISBN, and other types of information needed to purchase the book through an online transaction may also be stored. The search engine may then provide a navigation tool that comprises the HTML representation of the index 38 wherein topics and subtopics within the index 38 link to URL's of web content identified as being related to the topic or subtopic selected by the user. Optionally, in certain embodiments, the topics and subtopics may also include links to portions of the corpus 32 that are related to the topic selected by the user. In this way, a user may select a topic presented by the corpus 32 in view of the information presented by the corpus 32 and related information stored on the World Wide Web. In other embodiments, other techniques are employed for semantic processing and for determining a topic that can be associated with a portion of text within the corpus.

The data flow diagram depicted in FIGS. 2A and 2B may be implemented in a data processing process wherein a data processing program processes the corpus and generates an index that links topics in the corpus to information from a data sources, such as the Internet. Turning now to FIG. 3, a flow chart illustration of one such process is depicted. Specifically, FIG. 3 depicts a process 70 for extending a corpus by identifying topics covered by that corpus and employing information stored in the corpus and related to the topics to identify information in a database that is also related to the topic. The process 70 also generates an optional graphical user interface, such as the interface depicted in FIG. 1, that includes links for topics listed in the index, and that may be employed by a user to access the information associated with the listed topics.

The process 70 begins with the act 72 of identifying a corpus that is to be extended, such as by selecting a publication that contains reference material. In step 74, the process 70 transforms, or casts, the corpus into a normal form for processing. In one practice, this involves decomposing the document format of the corpus into a standard hierarchical topic orientation with a mechanism for assigning text content to title, headers, topics, and sub-topics. Optionally, stop words, such as the common words "and", "them", and "within", are identified and removed during normalization.

After normalization, the process 70 proceeds to step 78 wherein the corpus is processed to identify which portions of the corpus relate to which topic. In one practice wherein the corpus includes a table of contents, the process 70 analyzes the document format of the corpus to locate within the text headings associated with the different topics. For example, as described in the above cited publication U.S. Pat. No. 5,963, 203 entitled "Automatic index creation for a word processor",

header information set off by HTML tags may be identified to find indicia of topic entries in the document being processed. However, any technique for processing a document to identify the sections of text related to a topic may be applied, including other techniques for analyzing the mark up form language of the document.

Proceeding to step 80, the process 70 analyzes the topics to identify a signature that may be understood as representative of the semantics of the topic. In one practice, the process 70 creates a word map per topic and subtopic. To this end the process 70 in step 80 may create a summary representation of the words in the text based on the number of, location of, and proximity of words within each topic and sub-topic. Other factors may be employed, or substituted for these. Statistics are maintained on different parts of the document structure such as titles, headings, paragraphs, sentences, and image.

Table I depicts that several topics may be identified within the corpus. For example, Table I depicts that the processed corpus includes the topics Archaeology, Argentina, Arithmetic, Art, and Astronomy.

TABLE I

Topic
Archaeology
Argentina
Arithmetic
Art
Astronomy

The process 70 in one practice may then determine for a given topic, the word count for the words that appear within the portion of text, or other content, associated with the respective topic. This is depicted in Table II, that shows an example of the word count, with stop words removed, for words that appear in the portion of the corpus related to the topic "Astronomy."

TABLE II

Word counts in Topic Astronomy	
Word	Count
actual	1
ad	1
adopted	1
advances	2
ancient	3
application	2
assigning	1
astronomer	2
astronomers	2
astronomical	5
Astronomy	12
astronomy	11
Astrophysicists	1
astrophysics	1

In process 70, after the word count, and other statistics are determined, signatures are generated using orthogonalization. For example, in one practice, given the word counts or word maps for all or a selected subset of topics simultaneously, the process 70 assigns a weight based on word count to each word within each topic or subtopic. Where using word counts the weight may be defined as the count. When using the word map, the weight of a word in a topic or subtopic may be assigned by an intra-document scoring function. Any suitable technique may be employed for performing intra-document scoring. These signatures may be edited or cleaned

11

manually to enhance the topical relevance and precision of the subsequent search and scoring process. Table III depicts an example signature for the topic "Astronomy."

TABLE III

<u>Signature for Topic Astronomy</u>	
Word	Count
Astronomy	23
earth	9
bodies	5
Astronomical	5
universe	4
celestial	3
circle	3
Observational	3
sky	3
Stars	3
Ancient	3
Daily	3
Heavenly	2
Astronomers	2
planet	2
relative	2
moved	2
heavens	2

After determining a signature, the process 70 may perform the optional step, step 82, of applying synonym Groups. In this optional step, the process 70 extends the signatures with synonym groups. To this end, words are replaced by groups of word substitutes having similar or identical meaning. Table IV depicts such an extension.

TABLE IV

<u>Astronomy Signature Post Synonym Reduction</u>	
Word	Count
Astronom	30
earth	9
bodies	5
universe	4
celestial	3
Circle	3
Observational	3
Sky	3
stars	3
ancient	3
daily	3
Heavenly	2
planet	2
relative	2
moved	2
heavens	2

After step 82, the process 70 may proceed to step 84, wherein the process reduces the signature to Keyword sets, optionally tailored for the search. The set of documents to be scored against a topic is preferably identified and manageable in size. The web for example is a large a document set to collect up and score against all web documents. Accordingly, in one practice traditional large scale search engines, such as Lycos and Alta Vista, may be used to identify a set of candidate relevant documents using a keyword set for search. Which subset of the Signature and synonym groups is included in the Keyword set may be determined based on a variety of measures including corpus document word count of the word and general frequency of the word. An example is presented in TABLE V.

12

TABLE V

Word	Count
Astronomy	30
Earth	9
Bodies	5
Universe	4
Celestial	3
Circle	3
Observational	3
Sky	3
Stars	3
Ancient	3
Daily	3

The keyword set may be applied to a search mechanism to pull in multiple discovered documents based on the keyword set. This may occur in step 88. For example, the query Find: Astronomy or Astronomical or Astronomers or earth or bodies or universe or celestial or circle or Observational or sky or stars or ancient or daily; may be generated from the keyword set and applied to the search mechanism to discover documents related to the selected topic.

After the step 88, the process 70 may proceed to step 90 for scoring of the discovered documents. The many discovered documents returned from a search function may be assigned individual scores against the corresponding corpus topics and subtopics. Scoring may be based on multiple tunable metrics and rules including functions over the word count or word map data structures. The score of topical overlap between two documents as a baseline is measured as a dot product of word counts or word frequencies within those documents).

Scoring Example

Word	Count in Astronomy	Count in discovered document	Score Contribution
Astronom	30	2	0.555556
earth	9	1	0.083333
bodies	5		
universe	4	5	0.185185
celestial	3		
circle	3	2	0.055556
Observational	3		
sky	3		
stars	3		
ancient	3	1	0.027778
daily	3	2	0.055556
Heavenly	2		
planet	2		
relative	2		
moved	2	1	0.018519
heavens	2	1	0.018519

After step 90, the process 70 proceeds to optional step 92, wherein the topic hierarchy and set of associated documents may be presented directly through an HTML or graphical user interface, such as the interface depicted in FIG. 1. Alternatively, content may be delivered through software API's (application program interfaces) to allow integration of output content with other content. Content may be navigated by walking the directory tree structure, or by keyword searching over the directory structure trees, corpus content, or discovered document content. Search results may point to topic paths or discovered documents.

FIG. 4 depicts one embodiment of the system 100 according to the invention. Specifically FIG. 4 depicts a functional block diagram that shows a system 100 that allows a surfer



13

102 to access a user interface 104 that couples to a database system 108. The database system 108 further couples to an OCI processor 112 that accesses a database of corpora 114 and a plurality of search engines 118. The database system 108 further couples to an application programming interface access layer 120 and through the API 120 can access a portal/search client 122. Additionally, the API 120 may also couple to a scoring mechanism 124.

More particularly, FIG. 4 depicts that a user 102 such as an Internet user may access a user interface 104, that may be similar to the user interface depicted in FIG. 1. As shown in FIG. 1 the user interface 10 may present to the user 102 a list of topics 112. The user 102 may select a topic from the index 112. As described with reference to FIG. 1, the selection of a link directs the user interface 104 to retrieve information from the database system 108. The database system 108 processes the users request from user 102 for information related to the selected topic.

The database system 108 may be any suitable database system, including the commercially available Microsoft Access database, and can be a local or distributed database system. The design and development of suitable database systems are described in McGovern et al., *A Guide To Sybase and SQL Server*, Addison-Wesley (1993). The database 108 can be supported by any suitable persistent data memory, such as a hard disk drive, RAID system, tape drive system, floppy diskette, or any other suitable system.

As further depicted by FIG. 4 that database system 108 may communicate with the OCI mechanism 112. The OCI mechanism 112 may be, in one embodiment, a computer process capable of implementing a process such as process 70 depicted in FIG. 3. The OCI mechanism can be realized as a software component operating on a conventional data processing system such as a Unix workstation. In that embodiment, the OCI mechanism can be implemented as a C language computer program, or a computer program written in any high level language including C++, Fortran, Java or basic. Techniques for high level programming are known, and set forth in, for example, Stephen G. Kochan, *Programming in C*, Hayden Publishing (1983). Accordingly, the OCI mechanism 112 may be employed by a system administrator to process corpora stored within the database 114. As discussed with reference to FIG. 3, the processed corpora results in a graphical user interface that may be stored within the database mechanism 108 and accessed by the user 102 through the topic navigator 104. Additionally, the OCI mechanism 112 may generate for the processed corpora of database 114 a set of links or pointers to content that corresponds with different listed topics within the index of the processed corpora. The OCI mechanism 112 may also store these associated links within the database system 108.

To this end, the OCI mechanism 112 may couple to one or more search engines 118 that allow the OCI mechanism 112 to retrieve content from a database source. In the depicted embodiment of FIG. 4, the database source that search engines 118 access is the World Wide Web 106. In this embodiment, the user interface 104 also couples to the World Wide Web 106 so that links activated by the user that relate to URL's of content stored on the World Wide Web 106 may be directly accessed by the user 102 through the user interface 104 through the connection between the user interface 104 and the World Wide Web 106. FIG. 4 further depicts that the database 108 communicated with an API layer 120. As shown in FIG. 4 the API layer sits between the portal search client 122 and the database system 108 and also sits between the scoring mechanism 124 and the database system 108. Accordingly, a portal search client such as the Yahoo site may

14

access the database system 108 through the API layer to provide users with access to an index linked to content on the World Wide Web.

Similarly, FIG. 4 depicts the scoring mechanism 124. The scoring mechanism 124 may be a computer process that accesses the database system 108 through the API 120. The scoring mechanism may perform data mining for identifying topics that are to be associated with different websites. In this way, the database system 108 may be employed for categorizing web sites according to their content. Thus, the system 100 depicted in FIG. 4 provides a system for categorizing information stored on the World Wide Web. The system described in FIG. 4 may operate on any suitable computer hardware, such as PC compatible computer systems, Sun workstations, or any other suitable hardware. The list of topics and the associated documents, or links to documents may then be stored in a relational database, or any suitable database with proper indexing for allowing rapid accessing of the data stored therein.

Once the system is operating, the system may be employed to provide a set of tools, such as that may operate as stand alone applications for single users, or that may be tools provided as client/server programs over a network. The tools may be provided as a collection of functions incorporated into an integrated research tool, or may co-exist as individual functions in a separate application.

Further embodiment, the systems and methods described herein may be employed for organizing a plurality of corpora into an indexed format that may be presented as a graphical user interface for a user to allow a user to access information related to the contents of a plurality of corpora. For example, FIGS. 5 through 9 depict the operation of a system that processes a plurality of text, such as reference texts. Accordingly, the system may be employed for the automatic creation of a topically organized book catalog, such as a catalog of reference books, with navigation, search, click through to external documents such as web documents, with information purchasing interfaces also. For example, FIG. 5 depicts a graphical user interface that presents to a user a plurality of topics each having a set of books within the topic. For example, the FIG. 5 depicts a topic reference that includes a set of encyclopedias and dictionaries within that reference. By activating the reference link, the user may be presented with the user interface shown in FIG. 6. FIG. 6, the individual references presented under the reference topic of FIG. 5 are outlined for the user allowing the user to select what type of reference the user would like to view.

For example, the user may select from encyclopedias, dictionaries, academic and learned society publications and other such publications. After making a selection FIG. 6, the user may be presented with the different books under each category. The example presented in FIGS. 5 through 9 shows that upon activating the link for encyclopedias, the user is presented with the different encyclopedias that have been processed by the system according to the invention. Upon selecting a link, such as the link for the Encyclopedia Britannica, the user may be presented with the interface shown in FIG. 8 that lists the different topics covered by the Encyclopedia Britannica. At this level, the process now proceeds as described above, with reference to FIGS. 1 through 4 wherein the individual topics maintained within the Encyclopedia Britannica may be employed for accessing contact, such as web contact particularly associated with the individual topics.

Those skilled in the art will know or be able to ascertain using no more than routine experimentation, many equivalents to the embodiments and practices described herein. For example, the systems and methods described herein may be

15

employed for providing encyclopedia (i.e., corpus) extender. An encyclopedia (as an archetype example of an orthogonal corpus) may be automatically extended by application of the systems and methods described above, to include links into the World Wide Web, or other database, via searching or meta-searching over the Web. The breadth and depth of the corpus enables a high quality, high coverage database of web links, with the web links organized according to the location in the topic hierarchy whose text was used to generate them. Such links may provide geographical maps, histories of topics of interest, access to theses and other types of information. Other applications include web book companions wherein the system processes a book, including a fictional work, a non-fictional work, or a reference book, through this system will allow automated construction of topical web sites as Web Companions to individual books. For example, a book such as *The Hunt for Red October* may be processed by the systems described herein to create links into the Web to documents associated with concepts from the book, such as links to the Navy Submarine division, links to topographic maps of the ocean floor, links to Russian Naval History, and other similar links.

A search engine extension may be provided by accessing the database 108 through the API. Thus a user may do a search on a web search engine, they may want to refine their search or get a second search opinion. Given a broad topic database such as that created in the Encyclopedia Extender application described above, refinement of a user's intended topic is enabled—through keyword-based narrowing, web link browsing, and display of proximal or correlated topics in the corpus topic hierarchy. For e-commerce, the systems described herein book/article browser/seller. Browsing over the topic hierarchy may provide indexes into books or articles for sale.

Additional applications can include a user Interface. The user interface allows users to view Web links through the topic hierarchies defined by the corpus. The topic hierarchy on the left lists the topics as per the corpus. The user may select keywords from the corpus outline, or from provided sample text inside the corpus documents, to better focus and score the topic. Users may augment the search terms or keywords with their own keywords or selected synonyms to more specifically tailor a concept to a need. Searching across the corpus or across the referenced links may include synonyms, stemming, frequency, proximity, stop words, hyponyms, and synonyms.

Additionally, authoring toolkits may be provided that allow publishers, editors, and authors to create corpus extensions and associated applications. For example, the systems and methods described herein may be employed to create development kits that publishers may use to index a book and create a web site that acts as the book companion described above.

It may be noted that human oversight or auditing of the document scoring and database may be done in order to augment the purely automated document selection. This may be done on a sampling basis for quality control. Different levels of sensitivity to content or product price points may be implied by different levels of human quality control. Moreover, it will be noted that the system described above has been done so with reference to documents stored on the Web. However, it will be understood by those of ordinary skill in the art, that the Web is being used here as a metaphor for any electronic document archive, and the systems and methods described herein are not limited to the Web.

Accordingly, it will be understood that the invention is not to be limited to the embodiments disclosed herein, that other

16

applications, such as information mining may be practiced with the systems and methods described herein, and that the invention is to be understood by the following claims which are to be interpreted as broadly as allowed under the law.

What is claimed is:

1. A method for extending a body of textual reference material to generate a directory for accessing documents from a document collection that is initially unconnected with said body of textual reference material, comprising

processing the body of textual reference material into a plurality of text portions, each text portion being associated with a single topic from a plurality of topics,

processing said plurality of text portions, generating keywords for each, assigning a weight to each keyword in a text portion, associating a keyword with a corresponding text portion if the weight of said keyword in said corresponding text portion is equal to or greater than a weight of said keyword in the text portions other than the corresponding text portion, or is equal to or greater than a predetermined threshold value, and forming first keyword-weight pairs of said associated keywords, and

applying the associated keywords to at least one document from the initially unconnected document collection, and forming second keyword-weight pairs associated with the at least one document, forming a numeric score between the first and second keyword-weight pairs, and associating based on said score the at least one document from the initially unconnected document collection and the single topic from the plurality of topics.

2. A method according to claim 1 comprising creating a graphical interface representative of an identified orthogonal organization of topics for allowing a user to access information retrieved from the document collection and having an association with a topic.

3. A method according to claim 2, wherein the orthogonal organization of topics includes a hierarchical organization of substantially orthogonal topics.

4. A method according to claim 1, wherein the body of textual reference material is selected from the group consisting of an encyclopedia, a dictionary, a text book, a novel, a newspaper, a web site, and a glossary.

5. A method according to claim 1, wherein processing the body of textual reference material includes identifying a table of contents and chapter headings for the body of reference material.

6. A method according to claim 1, wherein processing the body of textual reference material includes identifying a keyword index for the body of reference material.

7. A method according to claim 1, wherein processing the body of textual reference material includes identifying definition entries in a dictionary or glossary.

8. A method according to claim 1, wherein processing the body of textual reference material includes normalizing the identified orthogonal organization of topics.

9. A method according to claim 1, wherein processing the plurality of text portions includes generating a word map representative of a statistical analysis of words contained in at least one text portion.

10. A computer readable medium having computer readable program codes embodied therein for causing a computer to extend a body of textual reference material to generate a directory for accessing documents from a document collection that is initially unconnected with said body of textual reference material, the computer readable medium program codes performing functions comprising:

17

processing the body of textual reference material into a plurality of text portions, each text portion being associated with a single topic from a plurality of topics, processing said plurality of text portions, generating keywords for each, assigning a weight to each keyword in a text portion, associating a keyword with a corresponding text portion if the weight of said keyword in said corresponding text portion is equal to or greater than a weight of said keyword in the text portions other than the corresponding text portion, or is equal to or greater than a predetermined threshold value, and forming first keyword-weight pairs of said associated keywords, and

18

applying the associated keywords to at least one document from the initially unconnected document collection, and forming second keyword-weight pairs associated with the at least one document, forming a numeric score between the first and second keyword-weight pairs, and associating based on said score the at least one document from the initially unconnected document collection and the single topic from the plurality of topics.

\* \* \* \* \*