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

(19) **World Intellectual Property Organization**
International Bureau

(43) **International Publication Date**
**8 September 2006 (08.09.2006)**

**PCT**

(10) **International Publication Number**
**WO 2006/093789 A2**

(51) **International Patent Classification:** **Not classified**

(21) **International Application Number:** PCT/US2006/006429

(22) **International Filing Date:** 24 February 2006 (24.02.2006)

(25) **Filing Language:** English

(26) **Publication Language:** English

(30) **Priority Data:**

| | | |
|---|---|---|
| 60/656,088 | 25 February 2005 (25.02.2005) | US |
| 11/232,898 | 23 September 2005 (23.09.2005) | US |

(71) **Applicant** *(for all designated States except US)*: **APPLIED LINGUISTICS, LLC** [US/US]; 33920 US Highway 19 North, Suite 269, Palm Harbor, Florida 34684 (US).

(72) **Inventor** *(for all designated States except US)*: **SCOTT, Bernard**; 369 Moorings Cove Drive, Tarpon Springs, Florida 34689 (US).

(72) **Inventors: TIMOFEYEV, Maksim**; Comerston Apartments, 644, Haven Place, Tarpon Springs, FL 34689 (US). **SPEERS, d'ARMOND**; 2026, Gold Dust Court, Trinity, FL 34655 (US).

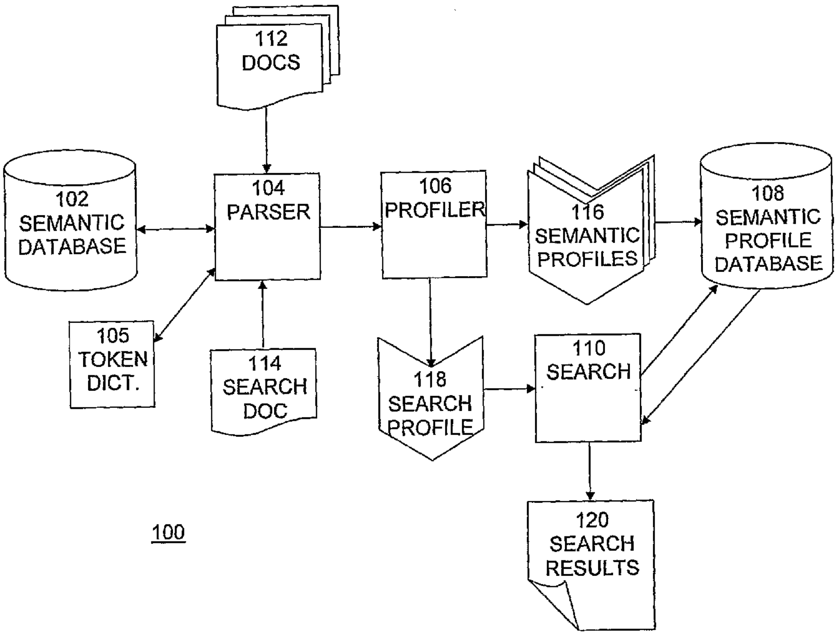
(74) **Agents: PENNINGTON, Edward, A.** et al.; Swidler Berlin Shereff Friedman LLP, 3000 K Street, N.W., Suite 300, Washington, D.C. 20007-5116 (US).

(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, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, 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, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, 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).

*[Continued on next page]*

(54) **Title:** SEMANTIC DOCUMENT PROFILING



(57) **Abstract:** A method of semantic profiling of documents comprises receiving a document to be profiled, the document comprising a plurality of terms, for each of at least a portion of the plurality of terms in the document determining a part of speech and a grammatical function of the term, obtaining senses of the term, selecting a sense as a most likely meaning of the term, and calculating an information value of the term, and generating a semantic profile of the document comprising at least some of the calculated information values.

**Published:**

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

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

Patent Application for:

# SEMANTIC DOCUMENT PROFILING

Inventors:

Bernard SCOTT
Maksim TIMOFEYEV
d'Armond SPEERS

Applied Linguistics, L.L.C.

# SEMANTIC DOCUMENT PROFILING

## BACKGROUND OF THE INVENTION

### 1. Field of the Invention

**[0001]** The present invention relates to information technology and database management, and more particularly to natural language processing of documents, search queries, and concept-based matching of searches to documents.

### 2. Description of the Related Art

**[0002]** Information technology, the Internet and the Information Age have created vast libraries of information, both formal and informal, such as the compendium of websites accessible on the Internet. While representing vast investments of tremendous potential value, the usefulness of such data depends on its accessibility, which depends upon the ease with which a particularly relevant document can be located, and the ease with which relevant information within a document can be found. Consequently, locating relevant information among and within large volumes of natural language documents (referred to often as text data) is an important problem.

**[0003]** Current commercial text retrieval systems generally focus on the use of keywords to search for information. Such systems typically use a combination of keywords supplied by the user to identify documents containing those same terms. In general, documents identified by means of such a search are ranked on the basis of keyword occurrence and frequency, a procedure that often requires users to plough through numerous documents to find relevance.

**[0004]** A second problem with current search systems concerns the requirement that the user specify the precise search terms (i.e., key words). However, problems may arise because key words are often ambiguous (i.e., the terms have multiple meanings and different terms may be used for the same subject matter), so documents may be retrieved that are not relevant to the intended search. Further, relevant documents may exist that do not include the same terms as those used in the query. In order to improve the results of such a search, a user may have to revise the query more than once, manually rewording the query based solely on the documents retrieved. The time spent on formatting queries and scanning documents can be quite burdensome and, when accomplished on a pay-for-use commercial database, may be quite expensive to obtain satisfactory results.

**[0005]** The prior art search methods suffer from inability or weak ability to identify and correlate concepts (as opposed to key words) within documents and a query, and thus are unable to reliably identify related documents when there are few, if any, common terms between them. Similarly, search methods based purely on term lookup are unable to rank documents based upon their conceptual relatedness, which would be highly desirable to a user researching a particular idea or concept.

**[0006]** Accordingly, there is a need for a true natural language processor that can index documents based upon concepts, that can correlate documents based upon their conceptual relatedness, that can serve as a search engine or component of same that will identify documents matching or relating to the concepts contained within a search query consisting of a document or portion of a document, that can focus user attention on portions of a returned document that are most relevant to the user's query, and finally, that can summarize the contents of returned documents when so requested.

[0007] Other problems with the prior art not described above can also be overcome using the teachings of the present invention, as would be readily apparent to one of ordinary skill in the art after reading this disclosure.

## SUMMARY OF THE INVENTION

[0008] The present invention provides true natural language processing by profiling documents based upon concepts contained in those documents. The present invention provides the capability to correlate documents based upon their conceptual relatedness, to perform searching of documents using a search query consisting of a document or portion of a document, to highlight portions of a returned document that are most relevant to the user's query, to summarize the contents of returned documents.

[0009] In one embodiment of the present invention, a method of semantic profiling of documents comprises receiving a document to be profiled, the document comprising a plurality of terms, for each of at least a portion of the plurality of terms in the document determining a part of speech and a grammatical function of the term, obtaining senses of the term, selecting a sense as a most likely meaning of the term, and calculating an information value of the term, and generating a semantic profile of the document comprising at least some of the calculated information values.

[0010] In one aspect of the present invention, a sense may be selected as a most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document. A sense may be selected as a most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document by determining at what level of a semantic taxonomy hierarchy the senses of each term are related to the senses of the other terms and assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to senses of the other terms. The lower the level at which each sense of the term is related to senses of the other terms, the higher the weighting assigned to the relation may be. The determination of taxonomic relationships of the senses of the term with senses of other terms in the document may be done using a moving window based on the term.

**[0011]** In one aspect of the present invention, a sense may be selected as a most likely meaning of the term based on statistics of co-occurrence of senses at a selected taxonomic level. A sense may be selected as a most likely meaning of the term based on a most probable meaning of each sense based on statistics of occurrences of meaning. A sense may be selected as a most likely meaning of the term by selecting as the most likely meaning of the term either the sense determined to be the most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document, or the sense determined to be the most likely meaning of the term based on statistics of co-occurrence of senses at a selected taxonomic level, or the sense determined to be the most likely meaning of the term based on the most probable meaning of each sense based on statistics of occurrences of meaning.

**[0012]** In one aspect of the present invention, the information value may be calculated by obtaining information relating to the selected sense of the term, weighting the information relating to the selected sense of the term, and calculating the information value based on the weighted information relating to the selected sense of the term.

**[0013]** In one aspect of the present invention, the information relating to the selected sense of the term may comprise semantic and syntactic information relating to the term and to the selected sense of the term. The information relating to the selected sense of the term may be at least one of information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, and information identifying a polysemy count of the term. The information relating to the selected sense of the term may further comprise a level of a semantic taxonomy hierarchy at which the senses of each term are related to selected senses of other terms in the document. The information relating to the selected sense of the term may be weighted by applying a weighting to the at least one of information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, and information identifying a polysemy count of the term and assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to a sense of at least one other term.

**[0014]** In one aspect of the present invention, the information relating to the selected sense of the term may comprise information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, information identifying a polysemy count of the term, and a level of a semantic taxonomy hierarchy the senses of each term are related. The information relating to the selected sense of the term may be weighted by applying a weighting to the information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, and information identifying a polysemy count of the term and assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to a sense of at least one other term. The information value may be calculated according to:

$$PICW = wPOS\frac{(wPCT + wSSC)}{2}\sum_{i=N}^{N} wGF,$$

$$ICW = PICW \times TW, \text{ and}$$

$$InfoVal = \frac{\sum_{i=N}^{N} ICW_i}{\sum ICW};$$

wherein wPOS is the weighted information identifying a part of speech of the term, wGF is the weighted information identifying a grammatical function of the term, wSSC is the weighted information identifying a semantico-syntactic class of the term, wPCT is the weighted information identifying a polysemy count of the term, and TW is the taxonomic weight.

**[0015]** In one embodiment of the present invention, a method for performing document-based searching comprises receiving a search document, the search document comprising a plurality of terms, for each of at least a portion of the plurality of terms in the search document determining a part of speech and a grammatical function of the term, obtaining senses of the term, selecting a sense as a most likely meaning of the term, and calculating an information value of the term, and generating a semantic profile of the search document comprising at least some of the calculated information values, and accessing a database comprising a plurality of semantic profiles of documents to retrieve documents having semantic profiles that are similar to the semantic profile of the search documents, each semantic profile in the database comprising a plurality of information values of terms included in the document.

**[0016]** In one aspect of the present invention, a sense may be selected as a most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document. A sense may be as a most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document by determining at what level of a semantic taxonomy hierarchy the senses of each term are related to the senses of the other terms and assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to senses of the other terms. The lower the lower the level at which each sense of the term is related to senses of the other terms, the higher the weighting assigned to the relation may be. The determination of taxonomic relationships of the senses of the term with senses of other terms in the document may be done using a moving window based on the term. A sense may be selected as a most likely meaning of the term based on statistics of co-occurrence of senses at a selected taxonomic level. A sense may be selected as a most likely meaning of the term based on a most probable meaning of each sense based on statistics of occurrences of meaning. A sense may be selected as a most likely meaning of the term by selecting as the most likely meaning of the term either the sense determined to be the most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document, or the sense determined to be the most likely meaning of the term based on statistics of co-occurrence of senses at a selected taxonomic level, or the sense determined to be the most likely meaning of the term based on the most probable meaning of each sense based on statistics of occurrences of meaning.

[0017] In one aspect of the present invention, the information value is calculated by obtaining information relating to the selected sense of the term, weighting the information relating to the selected sense of the term, and calculating the information value based on the weighted information relating to the selected sense of the term. The information relating to the selected sense of the term may comprise semantic and syntactic information relating to the term and to the selected sense of the term. The information relating to the selected sense of the term may comprise at least one of information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, and information identifying a polysemy count of the term.

[0018] In one aspect of the present invention, the information relating to the selected sense of the term may further comprise a level of a semantic taxonomy hierarchy at which the senses of each term are related to selected senses of other terms in the document. The information relating to the selected sense of the term may be weighted by applying a weighting to the at least one of information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, and information identifying a polysemy count of the term, and assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to a sense of at least one other term.

**[0019]** In one aspect of the present invention, the information relating to the selected sense of the term may comprise information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, information identifying a polysemy count of the term, and a level of a semantic taxonomy hierarchy the senses of each term are related. The information relating to the selected sense of the term may be weighted by applying a weighting to the information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, and information identifying a polysemy count of the term, and assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to a sense of at least one other term. The information value may be calculated according to:

$$PICW = wPOS\frac{(wPCT + wSSC)}{2}\sum_{i=N}^{N} wGF\,,$$

$$ICW = PICW \times TW\text{, and}$$

$$InfoVal = \frac{\sum_{i=N}^{N} ICW_i}{\sum ICW}\,;$$

wherein wPOS is the weighted information identifying a part of speech of the term, wGF is the weighted information identifying a grammatical function of the term, wSSC is the weighted information identifying a semantico-syntactic class of the term, wPCT is the weighted information identifying a polysemy count of the term, and TW is the taxonomic weight.

**[0020]** In one aspect of the present invention, the method may further comprise ranking the retrieved documents based on similarity to the search document. The retrieved documents may be ranked by comparing semantic profiles of the retrieved documents to the semantic profile of the search document.

**[0021]** In one aspect of the present invention, the method may further comprise generating a summary of each of at least some of the retrieved documents. The summary of each of at least some of the retrieved documents may be generated by calculating an information value for each sentence in the document, deleting sentences having an information value below a threshold value, deleting non-main clauses having an information value below a threshold value, normalizing to declarative form each remaining sentence and clause, deleting noun phrases and verb phrases having an information value below a threshold value from each remaining sentence and clause, selecting at least a portion of the remaining sentences and clauses as kernel phrases, and replacing terms present in the kernel phrases with terms relating to similar concepts selected from a taxonomic hierarchy to form the summary of the document. Terms present in the kernel phrases may be replaced with terms relating to similar concepts selected from a taxonomic hierarchy by identifying a subject, verb, and object terms of each kernel phrase, determining intersections of the subject, verb, and object terms at a level of the taxonomic hierarchy, obtaining concept labels of the level of the taxonomic hierarchy at which the intersections, and combining the obtained concept labels to form sentences that summarize the kernel phrases.

[0022] In one embodiment of the present invention, a method of summarizing a textual document comprises calculating an information value for each sentence of the document, deleting from consideration for a summary sentences having an information value below a first threshold value to form retained sentences, deleting from the retained sentences non-main clauses having information values below a second threshold value to form retained clauses, normalizing the retained clauses to declarative form, deleting modifiers having information values below a third threshold value from the normalized retained clauses to from kernel phrases, selecting at least a portion of the kernel phrases, and replacing at least portions of the kernel phrases with terms relating to similar concepts selected from a taxonomic hierarchy.

[0023] In one aspect of the present invention, the information value for a sentence may be calculated by calculating an information value for at least some terms in the sentence, summing the calculated information values for the terms in the sentence and dividing the sum by a number of the terms for which the information values were summed. The information value for a term may be calculated by determining a part of speech and a grammatical function of the term, obtaining senses of the term, selecting a sense as a most likely meaning of the term, and calculating an information value of the term.

[0024] In one aspect of the present invention, a sense may be selected as a most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document. A sense may be selected as a most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document by determining at what level of a semantic taxonomy hierarchy the senses of each term are related to the senses of the other terms and assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to senses of the other terms. The lower the level at which each sense of the term is related to senses of the other terms, the higher the weighting assigned to the relation may be. The determination of taxonomic relationships of the senses of the term with senses of other terms in the document may be done using a moving window based on the term.

**[0025]** In one aspect of the present invention, a sense may be selected as a most likely meaning of the term based on statistics of co-occurrence of senses at a selected taxonomic level. A sense may be selected as a most likely meaning of the term based on a most probable meaning of each sense based on statistics of occurrences of meaning. A sense may be selected as a most likely meaning of the term by selecting as the most likely meaning of the term either the sense determined to be the most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document, or the sense determined to be the most likely meaning of the term based on statistics of co-occurrence of senses at a selected taxonomic level, or the sense determined to be the most likely meaning of the term based on the most probable meaning of each sense based on statistics of occurrences of meaning.

**[0026]** In one aspect of the present invention, the information value may be calculated by obtaining information relating to the selected sense of the term, weighting the information relating to the selected sense of the term, and calculating the information value based on the weighted information relating to the selected sense of the term. The information relating to the selected sense of the term may comprise semantic and syntactic information relating to the term and to the selected sense of the term. The information relating to the selected sense of the term may comprise at least one of information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, and information identifying a polysemy count of the term. The information relating to the selected sense of the term may further comprise a level of a semantic taxonomy hierarchy at which the senses of each term are related to selected senses of other terms in the document. The information relating to the selected sense of the term may be weighted by applying a weighting to the at least one of information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, and information identifying a polysemy count of the term and assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to a sense of at least one other term. The information relating to the selected sense of the term may comprises information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, information identifying a polysemy count of the term, and a level of a semantic taxonomy hierarchy the senses of each term are related. The information relating to the selected sense of the term may be weighted by applying a weighting to the information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, and information identifying a polysemy count of the term and assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to a sense of at least one other term. The information value may be calculated according to:

$$PICW = wPOS\frac{(wPCT + wSSC)}{2}\sum_{i=N}^{N} wGF,$$

$$ICW = PICW \times TW, \text{ and}$$

$$InfoVal = \frac{\sum_{i=N}^{N} ICW_i}{\sum ICW};$$

wherein wPOS is the weighted information identifying a part of speech of the term, wGF is the weighted information identifying a grammatical function of the term, wSSC is the weighted information identifying a semantico-syntactic class of the term, wPCT is the weighted information identifying a polysemy count of the term, and TW is the taxonomic weight.

**[0027]** In one aspect of the present invention, the information value for a clause may be calculated by calculating an information value for at least some terms in the clause, summing the calculated information values for the terms in the clause, and dividing the sum by a number of the terms for which the information values were summed. The information value for a modifier may be calculated by calculating an information value for at least some terms in the modifier, summing the calculated information values for the terms in the modifier, and dividing the sum by a number of the terms for which the information values were summed. The kernel phrases may be selected by selecting as the selected kernel phrases that portion of the kernel phrases having higher information values than kernel phrases that are not selected. The portion of kernel phrases that are selected may be a fraction or percentage of the kernel phrases. The portion of kernel phrases that are selected may be determined so as to form a summary of a selected size. The kernel phrases may be replaced with terms relating to similar concepts selected from a taxonomic hierarchy by identifying subject, verb, and object terms for each kernel phrase, determining intersections of the identified subject, verb, and object terms at a level of the taxonomic hierarchy, combining concept labels of the level of the taxonomic hierarchy at which the identified subject, verb, and object terms intersect to form sentences that summarize each kernel phrase, and replacing the kernel phrases with the summary sentences.

**[0028]**

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0029]** Further features and advantages of the invention can be ascertained from the following detailed description that is provided in connection with the drawings described below:

**[0030]** Fig. 1 is an exemplary block diagram of a system in which the present invention may be implemented.

**[0031]** Fig. 2 is an exemplary flow diagram of a process of operation of a parser and a profiler, shown in Fig. 1.

**[0032]** Fig. 3 is an exemplary flow diagram of a process of part of speech/grammatical function analysis shown in Fig. 2.

**[0033]** Fig. 4 is an exemplary flow diagram of a process of word sense and feature analysis shown in Fig. 2.

**[0034]** Fig. 5 is an exemplary flow diagram of a process of word sense disambiguation shown in Fig. 2.

**[0035]** Fig. 6 is an exemplary flow diagram of a process of information value generation shown in Fig. 2.

**[0036]** Fig. 7 is an exemplary flow diagram of a process of searching based on a search document.

**[0037]** Fig. 8 is illustrates an example of a structure of a semantic database shown in Fig. 1.

**[0038]** Fig. 9 is an exemplary format of a data structure generated by the parser shown in Fig. 1.

**[0039]** Fig. 10 illustrates an example of the determination and weighting of relationships among senses of terms.

**[0040]** Fig. 11 illustrates an example of a decision table used for word sense disambiguation.

**[0041]** Fig. 12 illustrates an example of a decision table used for word sense disambiguation.

**[0042]** Fig. 13 is an exemplary flow diagram of a process of word sense disambiguation using a co-occurrence matrix.

**[0043]** Fig. 14 illustrates an example of a co-occurrence matrix.

**[0044]** Fig. 15 is an exemplary format of a semantic profile.

**[0045]** Fig. 16 illustrates an example of the parsing of a sentence.

**[0046]** Fig. 17 is an exemplary format of a data structure.

**[0047]** Fig. 18 illustrates an example of weighting parameters that may be set for the profiling process.

**[0048]** Fig. 19 illustrates an example of statistics relating to an input sentence.

**[0049]** Fig. 20 illustrates an example of data retrieved from the semantic database and input to the profiler.

**[0050]** Fig. 21 illustrates an example of output from a process of determination of the taxonomic relationships among the senses of the words in the exemplary sentence.

**[0051]** Fig. 22 illustrates an example of information included in a semantic profile generated by the profiler.

**[0052]** Fig. 23 illustrates an example of similarity ranking of two documents.

**[0053]** Fig. 24 illustrates an example of similarity ranking of a plurality of documents.

**[0054]** Fig. 25 illustrates an example of a process of text summarization.

**[0055]** Fig. 26 is an exemplary block diagram of a computer system, in which the present invention may be implemented.

## DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

**[0056]** The present invention provides true natural language processing by profiling documents based upon concepts contained in those documents. The present invention provides the capability to correlate documents based upon their conceptual relatedness, to perform searching of documents using a search query consisting of a document or portion of a document, to highlight portions of a returned document that are most relevant to the user's query, to summarize the contents of returned documents.

**[0057]** An exemplary block diagram of a system 100 in which the present invention may be implemented is shown in Fig. 1. System 100 includes semantic database 102, parser 104, profiler 106, semantic profile database 108, and search process 110. Semantic database 102 includes a database of words and phrases and associated meanings associated with those words and phrases. Semantic database 102 provides the capability to look up words, word forms, and word senses and obtain one or more meanings that are associated with the words, word forms, and word senses. Parser 104 uses semantic database 102 in order to divide language into components that can be analyzed by profiler 106. For example, parsing a sentence would involve dividing it into words and phrases and identifying the type of each component (e.g., verb, adjective, or noun). The language processed by parser 104 is included in documents, such as documents 112 and search document 114. Parser 104 additionally analyses senses and features of the language components.

**[0058]** Profiler 106 analyzes the language components and generates semantic profiles 116 that represent the meanings of the language of the documents being processed. Semantic profile database 108 stores the generated semantic profiles 116, so that they can be queried. In order to generate semantic profile database 108, language extracted from a corpus of documents 112 is parsed and profiled and the semantic profiles 116 are stored in semantic profile database 108. In order to perform a search based on a search document 114, the search document 114 is parsed and profiled and the resulting search profile 118 is used by search process 110 to generate queries to search semantic profile database 108. Semantic profile database 108 returns results of those queries to search process 110, which performs additional processing, such as ranking of the results, selection of the results, highlighting of the results, summarization of the results, etc. and forms search results 120, which may be returned to the initiator of the search.

**[0059]** The documents input into system 100, such as documents 112 and search document 114, may be any type of electronic document that includes text or that may be converted to text, or any type of physical document that has been converted to an electronic form that includes text or that may be converted to text. For example, the documents may include documents that are mainly text, such as text or word processor documents, documents from which text may be extracted, such as Hypertext Markup Language (HTML) or eXtensible Markup Language (XML) documents, image documents that include images of text that may be extracted using optical character recognition (OCR) processes, documents that may include mixtures of text and images, such as Portable Document Format (PDF) documents, etc., or any type or format of document from which text may be extracted or that may be converted to text. The present invention contemplates the use of any and all types and formats of such documents.

**[0060]** A process of operation of parser 104 and profiler 106, shown in Fig. 1, is shown in Fig. 2. Parser 104 receives input language and performs part of speech (POS) / grammatical function (GF) / base form analysis 202 on the language. POS/GF analysis involves classifying language components based on the role that the language component plays in a grammatical structure. For example, for language components such as words and phrases (terms), each term is classified based on the role that the term plays in a sentence. Parts of speech are also known as lexical categories and include open word classes, which constantly acquire new members, and closed word classes, which acquire new members infrequently if at all.

**[0061]** Parts of speech are dependent upon the language being used. For example, in traditional English grammar there are eight parts of speech: noun, verb, adjective, adverb, pronoun, preposition, conjunction, and interjection. The present invention contemplates application to any and all languages, as well as to any and all parts of speech found in any such language.

**[0062]** Grammatical function (GF) is the syntactic role that a particular term plays within a sentence. For example, a term may be (or be part of) the object, subject, predicate or complement of a clause, and that clause may be a main clause, dependent clause or relative clause, or be a modifier to the subject, object predicate or complement in those clauses.

**[0063]** The words and phrases that may be processed by parser 104 are stored in a token dictionary 105. The input language is tokenized, that is broken up into analyzable pieces, based on the entries in token dictionary 105. Once the POS/GF for a token has been determined, the actual token may be replaced by its base form, as any variation in meaning of the actual token from the base form is captured by the POS/GF information.

**[0064]** Parser 204 also performs a lookup of the senses and features of words and phrases. Word senses relate to the different meanings that a particular word (or phrase) may have. The senses may depend upon the context of the term in a sentence.

**[0065]** A flow diagram of a process of POS/GF analysis 202, shown in Fig. 2, is shown in Fig. 3. It is best viewed in conjunction with Fig. 1 and with Fig 16. POS/GF analysis involves classifying language components based on the role that the language component plays in a grammatical structure. POS/GF analysis process 202 begins with step 302, in which full text strings are passed to parser 104. The text strings are extracted from documents 112 or search document 114, which, as described above, may be text documents, word processor documents, text documents derived from optical character recognition of image documents, text documents derived from voice recognition of audio files, etc. Typically, the text strings are sentences, but may be any other division of the document, such as paragraphs, pages, etc. An example 1600 of the parsing of a sentence is shown in Fig. 16. In the example shown in Fig. 16, the sentence to be parsed 1602 is "My associate constructed that sailboat of teak."

**[0066]** In step 302, parser 104 tokenizes a text string. A token is a primitive block of a structured text that is a useful part of the structured text. Typically, in text documents, each token includes a single word of the text, multi-word phrases, and punctuation. However, tokens may include any other portions of the text, such as numbers, acronyms, symbols, etc. Tokenization is performed using a token dictionary 105, shown in Fig. 1. Token dictionary 105 may be a standard dictionary of words, but preferably, token dictionary 105 is a custom dictionary that includes a plurality of words and phrases, which may be queried for a match by parser 104. For each fragment of the text, parser 104 accesses token dictionary 105 to determine if the fragment of text matches a term stored in the dictionary. If there is a match, that fragment of text is recognized as a token. In the example shown in Fig. 16, matched tokens are shown in column 1604 and unmatched tokens are shown in column 1606. In addition, matches of so-called stop words are shown in column 1608. Stop words are words that are not processed by the parser and profiler and are not included in the semantic profile of the text. Words are typically considered stop words because they likely add little or no information to the text being processed, due to the words themselves, the commonness of usage, etc.

**[0067]** In step 306, parser 104 determines the part of speech (POS) of the token and the grammatical function (GF) of the token, and tags the token with this information. POS/GF analysis involves classifying language components based on the role that the language component plays in a grammatical structure. For example, for language components such as words and phrases, each term is classified based on the role that the term plays in a sentence. Parts of speech are also known as lexical categories and include open word classes, which constantly acquire new members, and closed word classes, which acquire new members infrequently if at all. For some terms, there may be only one possible POS/GF, while for other terms, there may be multiple possible POS/GFs. For each token processed by parser 104, the most likely POS/GF for the token is determined based on the context in which each token is used in the sentence, phrase, or clause being parsed. If a single POS/GF cannot be determined, two or more POS/GFs may be determined to be likely to be related to the token. The token is then tagged with the determined POS/GF information. In the example shown in Fig. 16, POS tags are shown in column 1610 and GF tags are shown in column 1612. As shown in the example, the word "associate" is tagged as being a noun (N) (part of speech) and as a subject (grammatical function). Likewise, constructed is shown as being a verb (V) (POS) and as the main verb (GF).

[0068] In step 308, further processing of the tagged tokens is performed. For example, such processing may include identification of clauses, etc. In addition, the actual token may be replaced by the base form of the term represented by the token. In the example shown in Fig. 16, the base forms of words are shown in column 1514. As shown in the example, the word "constructed" is replaced by its base form "construct".

[0069] In step 310, a data structure including each token that has been identified and the POS/GF information with which it has been tagged is generated. An example of a format 1700 of this data structure is shown in Fig. 17.

[0070] A flow diagram of a process of word sense and feature analysis 204, shown in Fig. 2, is shown in Fig. 4. In step 402, each word (base form) or phrase in the data structure generated in step 310 of Fig. 3 is processed. In step 404, semantic database 102 is accessed to lookup the sense of each term by its tagged part of speech. Turning briefly to Fig. 8, the structure of semantic database 102 is shown. Semantic database 102 is organized using each base form 802 of each term, and the POS/GF information 804B, 804B associated with each base form 802 as the keys to access the database. Associated with each POS/GF are one or more senses or meanings 806A-D that the term may have for the POS/GF.

[0071] Associated with each sense 806A-D is additional information 808A-D relating to that sense. Included in the additional information, such as 808A, is the semcode 809A, which is a numeric code that identifies the semantic meaning of each sense of each term for each part of speech. Semcodes are described in greater detail below. Additional information 808A-D may also include Semantico-Syntactic Class (SSC), such as 810A, polysemy count (PCT), such as 812A, and (MPM), such as 816A.

The SSC, such as 810A, provides information identifying the position of the term in a semantic and syntactic taxonomy, such as the Semantico-Syntactic Abstract Language, which is described in "The Logos Model: An Historical Perspective," Machine Translation 18:1-72, 2003, by Bernard Scott, which is hereby incorporated by reference in its entirety. The SSC is described in greater detail below. It is to be noted that this semantic and syntactic taxonomy is merely an example, and the present invention contemplates use with any semantic and syntactic taxonomy.

**[0072]** The PCT, such as 812A, provides information assigning a relative information value based on the number of different meanings that each term has. Terms with more than one part of speech are called syntactically ambiguous and terms with more than one meaning (semcode) within a part of speech are called polysemous. The present invention makes use of the observation that terms with few meanings generally carry more information in a sentence or paragraph than terms with many meanings. Terms with one or just a few different meanings have a low polysemy count and therefore likely have a high inherent information value, since such terms communicate just one or few meanings and are hence relatively unambiguous. The PCT is described in greater detail below. It is to be noted that this polysemy count is merely an example, and the present invention contemplates use with any meaning based measure of relative information value.

**[0073]** The most probable meaning MPM, such as 816A, provides information identifying the sense of each word that is the most likely meaning of the term based on statistics of occurrences of meaning, which is determined based on standard or custom-generated references of such statistics.

[0074] In step 404, semantic database 102 is accessed using the term and POS/GF tag associated with the term to retrieve the senses 806A-D and additional information 808A-D associated with that term and POS/GF. In order to access semantic database 102, each term in the data structure generated in step 310 is matched (or attempted to be matched) to one or more entries in semantic database 102. While in many cases, a simple matching is sufficient to retrieve the appropriate entries, in some cases, more sophisticated matching is desirable. For example, single words and noun phrases may be matched by simple matching, but for verb phrases, it may be desirable to perform extended searching over multiple tokens, in order to match verb phrases.

[0075] In step 406, fallback processing in the case that the term is not found is performed. If the base form of the term is not found at all in semantic database 102, then that term is indicated as being unfound. However, if the base form of the term is found, but the part of speech with which the term has been tagged is not found, then additional processing is performed. For example, if a term is tagged as being an adjective, but an adjective POS/GF is not found for that term, then parser 104 looks for a noun POS/GF for the term. If a noun POS/GF for the term is found, the senses associated with the noun POS/GF for the term are used. If a noun POS/GF for the term is not found, then the term is indicated as being unfound.

[0076] In step 408, for each term that is found in semantic database 102, the information 808A-D associated with each sense of each term is retrieved.

[0077] In step 410, a data structure is generated that includes each semcode that has been retrieved, the position of word that the semcode represents in the text being processed, and the retrieved associated additional information. An exemplary format of such a data structure 900 is shown in Fig 9. For example, data structure 900 may include semcodes, such as 902A and 902B, position information, such as 903A and 903B, POS/GF information, such as 904A and 904B, and additional information, such as 908A and 908B, which includes SSC, such as 910A and 910B, PCT, such as 912A and 912B, MPM, such as 914A and 914B.

**[0078]** Returning to Fig. 4, it is seen that in step 410, a data structure, such as that shown in Fig. 9, is generated.

**[0079]** A process of word sense disambiguation 206, shown in Fig. 2, which is performed by profiler 106, is shown in Fig. 5. It is best viewed in conjunction with Fig. 1. Word sense disambiguation (WSD) is the problem of determining in which sense a word having a number of distinct senses is used in a given sentence, phrase, or clause. One problem with word sense disambiguation is deciding what the senses are. In some cases, at least some senses are obviously different. In other cases, however, the different senses can be closely related (one meaning being a metaphorical or metonymic extension of another), and there division of words into senses becomes much more difficult. Due to this difficulty, the present invention uses a number of analyses to select the senses that will be used for further analysis. Preferably, the present invention uses three measures of likelihood that a particular sense represents the intended meaning of each term in a document:

- Relationships of senses within the document itself, weighted based on the taxonomic level at which each relationship occurs;
- Statistics of co-occurrence of senses at a selected taxonomic level within a large corpus of text; and

- The most probable meaning of each sense based on statistics of occurrences of meaning.

[0080] Referring to Fig. 1, profiler 106 receives the data structure that was generated by parser 104. Profiler 106 then processes the data structure for each semcode (which represents a term (semcode)) in the data structure. Returning to Fig. 5, in step 502, profiler 106 determines the senses (meanings) of each term (semcode) that is the most likely intended sense based on a determination of taxonomic relationships among the different senses. In particular, profiler 106 determines the taxonomic relationships by determining at what level of a semantic taxonomy hierarchy the different senses of each term (semcode) are related. As described in greater detail below, semantic taxonomies are typically organized in a number of levels, which indicate a degree of association among various meanings. Typically, meanings that are related at some levels of the taxonomy are more closely related, while meanings that are related at other levels of the taxonomy are less closely related. For example, a suitable taxonomic structure may comprise seven levels, levels L0 – L6. Level 0 (L0) contains the literal word. Level L1 (L1) contains a small grouping of terms that are closely related semantically, e.g., synonyms, near synonyms, and semantically related words of different parts of speech derived from the same root, such as, e.g., the verb "remove" and the noun "removal." Each level above L1 represents an increasingly larger grouping of terms according to an increasingly more abstract, common, concept, or group of related concepts, i.e., the higher the level, the more terms are included in the group, and the more abstract the subject matter of the group. Thus, taxonomic levels convey, at various levels of abstraction, information regarding the concepts conveyed by terms and their semantic relatedness to other terms. Thus, profiler 106 determines at what level of a semantic taxonomy the different senses of each word are related, if at all. The Semantico-Syntactic Class (SSC) associated with each sense of a term (semcode) indicates the semantic taxonomy groupings that the term (semcode) belongs to and is used to determine the relationships among the different senses of each word.

**[0081]** It is to be noted that the processing performed by profiler 106 preferably uses the semcodes that are included in the data structure, rather than the actual term represented by the semcode. The use of the semcodes provides a quick and efficient way to determine relationships among words and phrases and greatly increases the performance of profiler 106. Thus, when the profiler is described as determining relationships among terms, or senses of terms, the profiler is processing the semcodes representing those terms, or senses of terms, rather than the terms themselves.

**[0082]** Profiler 106 assigns a taxonomic weight (TW) to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term (semcode) is related to senses of other terms in a document. Preferably, the lower the level at which a sense of a term (semcode) is related to the senses of other terms, the higher the weighting assigned to the relation. This is because terms related at lower levels of the semantic taxonomy are more closely related semantically, thus, assigning a higher weight to such a relation indicates greater semantic closeness. However, the present invention does not exclude a different weighting scheme than that described above. Preferably, the weight values for each taxonomic level are settable, such as with user-level or system-level parameters, which facilitates modification of the weighting scheme.

**[0083]** Preferably, the determination of relationships among senses of terms is done using a moving window of terms based on a selected term for which the most likely meaning is being determined. For example, in order to determine a likely meaning for a selected term, the TW may be based upon taxonomic relationships between one sense of the selected term and all senses of all other terms within the window. This determination may be repeated for each sense of the selected term, in order to determine which sense of the selected term is the most likely meaning. Another term is then selected and the window is moved appropriately. The size of the window is preferably settable, such as with a user-level or system-level parameter. In addition, multi-part windows may be used. For example, the TW may be increased for taxonomic relationships that occur within an inner window (such as two open-class words to the left and right of the selected word) of the overall window. Preferably, the weight values for each window part are settable, such as with user-level or system-level parameters, which facilitates modification of the window scheme. Likewise, it is preferred that only nouns, verbs, and adjectives are used in the determination of relationships among senses of terms. However, the present invention does not exclude the use of other parts of speech in the determination of relationships among senses of terms.

**[0084]** An example of the determination and weighting of relationships among senses of terms is shown in Fig. 10. As shown in Fig. 10, as an example, there are two words in a particular window of terms, words W1 and W2. W1 has two senses, S1 and S2, and W2 has three senses, S3, S4, and S5. The relationships of the senses of these words are determined at a number of levels, from L1, the level that includes the senses themselves, up through levels of decreasing semantic closeness and increasing conceptual generality. The level at which senses are related are weighted based on the semantic closeness represented by the level. Preferably, the levels of greatest semantic closeness are given higher weight than levels of lower semantic closeness. For example, L1 is given the highest weight and L4 is given the lowest weight. In this example, S2 is related to other senses at level L2, so S2 is assigned a weight of 1.5, while S3 is related to other senses at level L4 and is assigned a weight of .5.

**[0085]** The output of step 502 is an assignment of weights of taxonomic relationships for each sense of each term (semcode) for each position of a moving window (or windows) of terms in the document. This output may be processed to select a single sense of each term (semcode) that is the most likely to be the intended meaning. For example, the sense having the highest TW may be identified and compared to the sense with the next highest TW. If the highest TW is more than some threshold amount higher than the next highest TW, the sense having the highest TW may be selected as the most likely meaning based on the TW. If the highest TW is less than some threshold amount higher than the next highest TW, the sense having the highest TW may be selected as possibly being the most likely meaning based on the TW.

**[0086]** In step 504, profiler 106 determines the senses (meanings) of each term (semcode) that is the most likely intended sense based on use of a co-occurrence matrix that is generated based on analysis of a large corpus of text. The co-occurrences are analyzed based on a selected taxonomic level. Preferably, the analysis is performed so as to include co-occurrences of open class words and to exclude stop words and closed class words. The co-occurrences of each sense of each included term (semcode) in the corpus is determined at the selected taxonomic level relative to each other included term (semcode) in the corpus. Statistics or counts of the co-occurrences are generated and used to populate the co-occurrence matrix. It is seen that any corpus of documents may be used for generation of a co-occurrence matrix. Likewise, any taxonomic level or any number of taxonomic levels may be used. The present invention contemplates any and all such selections of corpus and selections of taxonomic levels.

**[0087]** Turning briefly to Fig. 13, the process of step 504 is shown. It is best viewed in conjunction with Fig. 14, which is an example of a co-occurrence matrix. The processing of step 504 begins with step 1302, in which the co-occurrence statistics for the term (semcode) being processed is retrieved. An example of a matrix 1400 of co-occurrence statistics is shown in Fig. 14. In step 1304, starting with terms (semcodes) that are least ambiguous, the co-occurrence statistics are summed for each sense of each term (semcode). Once a word sense has been resolved, other senses of that word are effectively eliminated from this computation, thus paring down the number of co-occurrences that will contribute to word sense resolution. In effect, reduced sense ambiguity is made use of as the remaining unresolved senses are addressed.

**[0088]** In step 1306, the sense that has the highest co-occurrence sum is selected. This is consistent with the principle used in resolving by taxonomic association, i.e., senses with highest overall co-occurrence statistics are like senses that have the highest relevance weight based on taxonomic association.

**[0089]** Returning to Fig. 5, in step 506, the most probable meaning of each sense based on statistics of occurrences of meaning is used to select the most likely senses of each term (semcode) in the document being processed. Each sense of each term (semcode) is looked-up in a table of frequency of occurrence, and the sense having the greatest frequency of occurrence is selected as the most probable meaning of the term (semcode) and is indicated as such. Such frequency tables are available as standard reference works, or they may be custom-generated as desired. The present invention contemplates the use of any and all such frequency tables.

[0090] In step 508, the results of steps 502, 504, and 506 are used to select a meaning for each term (semcode) in the document being processed. Preferably, a decision table is used, although any other form of decision logic or decision process is contemplated by the present invention. Examples of decision tables that may be used are shown in Figs. 11 and 12. In the example shown in Fig. 11, an unnormalized decision table 1100 is shown. Decision table 1100 includes column 1102, which includes the senses of the word being disambiguated, column 1104, which includes the assigned taxonomic weights for each sense from step 502, column 1106, which includes the assigned co-occurrence matrix weights from step 504, and column 1108, which includes the assigned indication of the most probable meaning from step 506. In the example shown in Fig. 11, the taxonomic weight indicates that sense S2 is the most likely meaning, the co-occurrence matrix indicates that sense S1 is the most likely meaning, and the most probable meaning indicates that sense S3 is the most likely meaning. Since, in this instance, there is no clear most likely meaning, the sense indicated as the most probable meaning is selected as the meaning of the word being disambiguated.

[0091] In the example shown in Fig. 12, a normalized decision table is shown. In normalized table 1200, the raw weights assigned in steps 502 and 504 are normalized to indicate a most likely meaning selected by those steps. Decision table 1200 includes column 1202, which includes the senses of the word being disambiguated, column 1204, which includes the assigned indication of the most likely meaning based on the taxonomic weights for each sense from step 502, column 1206, which includes the assigned indication of the most likely meaning based on the co-occurrence matrix weights from step 504, and column 1208, which includes the assigned indication of the most probable meaning from step 506. In the example shown in Fig. 12, the columns are not weighted equally, but are assigned different weights as desired to increase the accuracy of the disambiguation process. In this example, the taxonomic weight is assigned 45%, the co-occurrence matrix is assigned 35%, and the most probable meaning is assigned 20%. The taxonomic weight indicates that sense S2 is the most likely meaning, the co-occurrence matrix indicates that sense S1 is the most likely meaning, and the most probable meaning indicates that sense S3 is the most likely meaning. Due to the weighting of the columns, the sense indicated by the taxonomic weight, sense S2, is selected as the meaning of the word being disambiguated.

**[0092]** In step 510, a data structure is generated that includes the semcode of the most likely meaning of each term that was included in the data structure output from step 410 of Fig. 4. This data structure may have a format similar to that shown in Fig. 9, but includes the semcode of the most likely meaning of each term, rather than all semcodes for all terms in the input document.

**[0093]** A process of information value generation 206, shown in Fig. 2, which is performed by profiler 106, is shown in Fig. 6. It is best viewed in conjunction with Fig. 1. Process 206 generates an information value for the each meaning (semcode) that is included in the data structure generated in step 510. Process 206 begins with step 602, in which weights are applied to the information associated with each semcode. For example, the SSC 910A, PCT 912A, POS/GF 904A, etc., may be weighted to form weighted values, such as wSSC, wPCT, wPOS, wGF, etc.

**[0094]** In step 603, intermediate values are calculated for each term (semcode) using the weighted values determined in step 602. For example, intermediate values for each semcode may be calculated as follows:

$$PICW = wPOS\frac{(wPCT + wSSC)}{2}\sum_{i=N}^{N} wGF$$

$$ICW = PICW \times TW$$

where wPOS is the part of speech weight for the semcode obtained from semantic database 102, wPCT is the PCT weight for the semcode obtained from semantic database 102, wSSC the SSC weight for the semcode obtained from semantic database 102, wGF is the grammatical function weight for the semcode obtained from semantic database 102, and TW is the taxonomic weight for the semcode determined in step 502.

**[0095]** In step 604, an information value is calculated based on the intermediate values calculated in step 603. For example, an information value (InfoVal or IV) for each semcode may be calculated as follows:

$$InfoVal = \frac{\sum_{i=N}^{N} ICW_i}{\sum ICW}$$

**[0096]** The information value is a quantitative measure of the amount of information conveyed by each term (semcode) relative to the total information content of a document. Although suitable information values may be calculated as described above, the present invention is not limited to these calculations. Rather, the present invention contemplates any calculations that provide a suitable information value.

**[0097]** A process of semantic profile generation 210, shown in Fig. 2, which is performed by profiler 106, is shown in Fig. 6. It is best viewed in conjunction with Fig. 1. Semantic profile generation involves generating a semantic profile 116 for each document 112 that is processed by system 100. Semantic profile generation process 210 begins with step 606, in which the terms (semcodes) that are to be included in the semantic profile are selected. Typically, the semcodes are selected based on the information value associated with each semcode. For example, those semcodes having an information value greater than or equal to a threshold value may be selected for inclusion in the semantic profile, while those semcodes having an information value less than the threshold value are excluded from the semantic profile.

**[0098]** In step 608, the semantic profiles 116 themselves are generated. Preferably, a semantic profile is a vector of information values for the semcodes that were selected for inclusion in the semantic profile. An example of a format 1500 of a semantic profile is shown in Fig. 15. Each semantic profile having format 1500 includes a plurality of semcodes, such as 1502A, 1502B, ..., 1502N, each with an associated information value, such as 1504A, 1504B, ..., 1504N. It is to be noted that format 1500 is merely an example of a suitable semantic profile format, and that the present invention contemplates any semantic profile format.

**[0099]** In step 610, semantic profiles 116 are stored in semantic profile database 108, and indexed to provide efficient searching performance. Semantic profile database may be any standard or proprietary database that allows storage and searching for data having a format such as that of profiles 116.

**[0100]** An example of the processing performed by profiler 106, and in particular, the processing performed in step 206, shown in Fig. 5, and in steps 208 and 210, shown in Fig. 6, is shown in Figs. 18-22. In this example, the document being processed by profiler 106 includes the sentence shown in the example of Fig. 16. Referring to Fig. 18, an example of weighting parameters 1800 that may be set for the profiling process is shown. For example, parameters 1800 include POS weights 1802, Polysemy (PCT) weights 1804, SSC weights 1806, GF weights 1808, and winner or decision weights 1810.

**[0101]** Referring to Fig. 19, some statistics relating to the input sentence are shown. Referring to Fig. 20, an example of data retrieved from semantic database 102 and input to profiler 106 in data structure 900, shown in Fig. 9, is shown. Typically, the format shown is not used as the format of the data structure, but it may be, if desired. The data shown in this example includes the word 2002 to be analyzed, the type or POS 2004 of the word, the semcode 2006 of the sense of the word, broken into the taxonomic levels of the semcode, the GF 2008 of the word, the SSC 2010 of the word, and the location of the word in the document being analyzed by word number 2012 and sentence number 2014.

**[0102]** Referring to Fig. 21, an example of output from a process of determination of the taxonomic relationships among the senses of the words in the exemplary sentence, performed in step 506 of Fig. 5, is shown. For example, some senses of the words related at taxonomic level L6, some relate at taxonomic level L4, and some relate at taxonomic level L3. Also shown are some results of computations of some intermediate values, such as PICW, performed in step 602 of Fig. 6. Finally, referring to Fig. 22, an example of information 2200 included in a semantic profile 116 generated by profiler 106 is shown. Typically, the format shown is not used as the format of the semantic profile, but it may be, if desired. Semantic profile 2200 includes semcodes 2202 and corresponding information values 2204 computed in step 208 of Fig. 6. In addition, information indicating an importance of the sense of the word, as well as the word itself, is shown.

**[0103]** A process 110 of searching based on a search document, shown in Fig. 1, is shown in Fig. 7. It is best viewed in conjunction with Fig. 1. Search process 110 begins with step 702, in which a search document 114 is input and profiled. Search document 110 is parsed by parser 104 and profiled by profiler 106 to generate a search profile 118. The parsing and profiling operations performed by parser 104 and by profiler 106, respectively, are similar to the operations shown in Figs. 2-6, and need not be described again. The major difference from the operations shown in Figs. 2-6 is that search profile 118 is not stored in semantic profile database 108, but rather, in step 704, is used as the basis of a search query to search semantic profile database 108.

**[0104]** In step 706, the results 120 of the search performed in step 704 are output and may be presented to the user, or they may be further processed before presentation to the user or to other applications. For example, in step 708, the documents included in search results 120 may be ranked based on their similarity to search document 114. This may be done, for example, by comparing the semantic profile 118 of search document 114 with the profiles of the documents included in search results 120. Additional processing of search results 120 may include generating summaries 710 of the documents included in search results 120 or generating highlighting 712 of the documents included in search results 120.

[0105] An example of similarity ranking performed in step 708 of Fig. 7 is shown in Fig. 23. In this example, the document being compared includes the sentence used in the example shown in Fig. 16. This document is ranked compared to a document including a sentence having a similar meaning, but no words in common. This may be done, for example, by comparing the semantic profiles of the two documents. As shown in the example, words having similar meanings are matched and ranked by information value, and a total matched value indicating the similarity of the sentences documents is calculated. Referring to Fig. 24, a number of documents are ranked in order of similarity using a similar process.

[0106] An example of a process of text summarization performed in step 710 of Fig. 7 is shown in Fig. 25. Text summarization may be performed on some or all of the search results output in step 706 of Fig. 7, or text summarization may be performed on any text document, regardless of its source. Process 710 begins with step 2502, in which the information value (IV) for each sentence is calculated. The IV used for text summarization may be calculated as shown above in steps 603 and 604, or the IV used for text summarization may be calculated using other calculations. The present invention contemplates any calculations that provide a suitable IV for text summarization.

[0107] The IV for each sentence is calculated by calculating the IV for each word or phrase in the sentence (preferably excluding stop words), then summing the individual word or phrase IVs and dividing by the number of words or phrase for which IVs were summed. In step 2504, sentences having IV below a threshold value are deleted from the consideration for the summary. Such sentences with low IVs add relatively little information to the document. In step 2506, for the retained sentences, non-main clauses that have low IVs (IVs below a threshold) are deleted from the summary. For example, dependent clauses, relative clauses, parentheticals, etc. with low IVs add relatively little information to the document. In step 2508, the retained clauses are normalized to declarative form. For example, the clause "X is modified by Y" is transformed to the declarative form "Y modifies X".

[0108] In step 2510, modifiers of noun phrases (NPs) and verb phrases (VPs) that have low IVs (IVs below a threshold) are deleted. For example, in the VP "intend to purchase", the modifier "intend to" has a low IV. Upon deletion of the modifier, the phrase becomes simply "purchase". As another example, in the NP "piece of cake" the modifier " piece of " has a low IV. Upon deletion of the modifier, the phrase becomes simply "cake". Those phrases remaining at this point are termed "kernel phrases". In step 2512, all or a portion of the kernel phrases are selected based on an abstraction parameter that controls the quantity of results output by the summarization process. For example, the abstraction parameter may specify a percentage of the kernel phrases to select, in which case the specified percentage of kernel phrases having the highest IVs will be selected. As another example, the abstraction parameter may specify the size of the summary relative to the size of the original document. In this case, the quantity of kernel phrases (having the highest IVs) needed to produce the specified summary size will be selected from among the kernel phrases. The abstraction parameter may be set, for example, by operator input, according to the type of document being analyzed, or by means of a learning system.

[0109] In step 2514, the terms present in the kernel phrases are replaced by terms relating to similar concepts selected from the taxonomic hierarchy. In particular, the subject, verb, and object of each kernel phrase are identified and intersections of these terms at a level of the taxonomic hierarchy is determined. The concept labels of the level of the taxonomic hierarchy at which the intersections was determined may then be combined to form sentences that summarize the kernel phrases. For example, subject, verb, and object terms in kernel phrases may be analyzed to determine their intersections at level 3 (L3) of the taxonomic hierarchy. Summary sentences may then be generated that include the labels of the L3 categories at which the intersections occurred. An example of this is shown in the table below:

**Table 1**

| TYPE | SUBJECT | VERB | OBJECT |
|---|---|---|---|
| kernel | administration | is not for | tax hikes |
| kernel | treasurer | argues against | raising taxes |
| kernel | President | will veto | tax bill |
| | | | |
| label | **Government** | **opposes** | **tax increase** |

[0110] In this example, the intersection of the subject, verb, and object terms of each kernel phrase is determined and the labels of the L3 categories at which the intersections occurred is presented. For example, "administration", "treasurer", and "President" are all terms that intersect at L3 in a category labeled "Government", "is not for", "argues against", and "will veto" are all terms that intersect at L3 in a category labeled "opposes", and "tax hikes", "raising taxes", and "tax bill" are all terms that intersect at L3 in a category labeled "tax increase". Thus, the sentence, "Government opposes tax increase" forms the summary for the kernel phrases shown.

[0111] An exemplary block diagram of a computer system 2600, in which the present invention may be implemented, is shown in Fig. 26. System 2600 is typically a programmed general-purpose computer system, such as a personal computer, workstation, server system, and minicomputer or mainframe computer. System 2600 includes one or more processors (CPUs) 2602A-2602N, input/output circuitry 2604, network adapter 2606, and memory 2608. CPUs 2602A-2602N execute program instructions in order to carry out the functions of the present invention. Typically, CPUs 2602A-2602N are one or more microprocessors, such as an INTEL PENTIUM® processor. Fig. 26 illustrates an embodiment in which System 2600 is implemented as a single multi-processor computer system, in which multiple processors 2602A-2602N share system resources, such as memory 2608, input/output circuitry 2604, and network adapter 2606. However, the present invention also contemplates embodiments in which System 2600 is implemented as a plurality of networked computer systems, which may be single-processor computer systems, multi-processor computer systems, or a mix thereof.

[0112] Input/output circuitry 2604 provides the capability to input data to, or output data from, database/System 2600. For example, input/output circuitry may include input devices, such as keyboards, mice, touchpads, trackballs, scanners, etc., output devices, such as video adapters, monitors, printers, etc., and input/output devices, such as, modems, etc. Network adapter 2606 interfaces database/System 2600 with Internet/intranet 2610. Internet/intranet 2610 may include one or more standard local area network (LAN) or wide area network (WAN), such as Ethernet, Token Ring, the Internet, or a private or proprietary LAN/WAN.

[0113] Memory 2608 stores program instructions that are executed by, and data that are used and processed by, CPU 2602 to perform the functions of system 2600. Memory 2608 may include electronic memory devices, such as random-access memory (RAM), read-only memory (ROM), programmable read-only memory (PROM), electrically erasable programmable read-only memory (EEPROM), flash memory, etc., and electro-mechanical memory, such as magnetic disk drives, tape drives, optical disk drives, etc., which may use an integrated drive electronics (IDE) interface, or a variation or enhancement thereof, such as enhanced IDE (EIDE) or ultra direct memory access (UDMA), or a small computer system interface (SCSI) based interface, or a variation or enhancement thereof, such as fast-SCSI, wide-SCSI, fast and wide-SCSI, etc, or a fiber channel-arbitrated loop (FC-AL) interface.

[0114] The contents of memory 2608 varies depending upon the function that system 2600 is programmed to perform. However, one of skill in the art would recognize that these functions, along with the memory contents related to those functions, may be included on one system, or may be distributed among a plurality of systems, based on well-known engineering considerations. The present invention contemplates any and all such arrangements.

[0115] In the example shown in Fig. 26, memory 2608 includes semantic database 102, semantic profile database 108, token dictionary 105, parser routines 2612, profiler routines 2614, search routines 2616, ranking routines 2618, summarization routines 2620, highlight routines 2622, and operating system 2624. Semantic database 102 stores information about terms included in the documents being analyzed and provides the capability to look up words, word forms, and word senses and obtain one or more meanings that are associated with the words, word forms, and word senses. Semantic profile database 108 stores the generated semantic profiles 116, so that they can be queried. Token dictionary 105 stores the words and phrases that may be processed by parser 104. Parser routines 2612 implement the functionality of parser 104 and, in particular, the processes of steps 202 and 204, shown in Fig. 2. Profiler routines 2614 implement the functionality of parser 106 and, in particular, the processes of steps 206, 208, and 210, shown in Fig. 2. Search routines 2616 implement the functionality of search process 110, shown in Fig. 7. Ranking routines 2618 implement the functionality of ranking step 708, shown in Fig. 7. Summarization routines 2620 implement the functionality of summarization step 710, shown in Fig. 7, and in particular, the process steps shown in Fig. 25. Highlight routines 2622 implement the functionality of highlight step 712, shown in Fig. 7. Operating system 2628 provides overall system functionality.

**[0116]** As shown in Fig. 26, the present invention contemplates implementation on a system or systems that provide multi-processor, multi-tasking, multi-process, and/or multi-thread computing, as well as implementation on systems that provide only single processor, single thread computing. Multi-processor computing involves performing computing using more than one processor. Multi-tasking computing involves performing computing using more than one operating system task. A task is an operating system concept that refers to the combination of a program being executed and bookkeeping information used by the operating system. Whenever a program is executed, the operating system creates a new task for it. The task is like an envelope for the program in that it identifies the program with a task number and attaches other bookkeeping information to it. Many operating systems, including UNIX®, OS/2®, and Windows®, are capable of running many tasks at the same time and are called multitasking operating systems. Multi-tasking is the ability of an operating system to execute more than one executable at the same time. Each executable is running in its own address space, meaning that the executables have no way to share any of their memory. This has advantages, because it is impossible for any program to damage the execution of any of the other programs running on the system. However, the programs have no way to exchange any information except through the operating system (or by reading files stored on the file system). Multi-process computing is similar to multi-tasking computing, as the terms task and process are often used interchangeably, although some operating systems make a distinction between the two.

[0117] The present invention involves identifying each semantic concept of a document as represented by a semcode, and calculating a weighted information value for each semcode within the document based upon a variety of factors. These factors include part-of-speech (POS), semantico-syntactic class (SSC), polysemy count (PCT), usage statistics (FREQ), grammatical function within the sentence (GF), and taxonomic association with other terms in the text (TW). These factors are described in greater detail below.

[0118] **A Taxonomy of Semantic Codes – Semcodes**

[0119] The present invention utilizes a structured hierarchy of concepts and word meanings to identify the concept conveyed by each term in a sentence. Any structured taxonomy that correlates concepts with terms in a hierarchy that groups concepts in levels of relatedness may be used. The oldest and most well known taxonomy is Roget's Thesaurus. A taxonomy like Roget's Thesaurus can be viewed as a series of concept levels ranging from the broadest abstract categories down to the lowest generic category appropriate to individual terms or small sets of terms. For example, Roget's Thesaurus comprises seven levels, which may be summarized as levels L0 (the literal term itself) through L6 (the most abstract grouping).

[0120] The present invention employs a similar taxonomic hierarchy. In one embodiment, this hierarchy comprises seven levels, levels L0 – L6. Level 0 (L0) contains the literal word. Level 1 (L1) contains a small grouping of terms that are closely related semantically, e.g., synonyms, near synonyms, and (unlike the Roget taxonomy) semantically related words of different parts of speech derived from the same root, such as, the verb "remove" and the noun "removal." Each level above L1 represents an increasingly larger grouping of terms according to an increasingly more abstract, common concept or group of related concepts, i.e., the higher the level, the more terms are included in the group, and the more abstract the subject matter of the group. Thus, taxonomic levels convey, at various levels of abstraction, information regarding the concepts conveyed by terms and their semantic relatedness to other terms.

[0121] By assigning a number to each semantic grouping of terms in each taxonomic level, a unique code can thereby be assigned for each meaning of each term populating the taxonomy. At the highest, most abstract level, many terms will share this grouping code. At the lowest, least abstract level, as few as one term may have the grouping code. In effect, when individual codes at all levels of the taxonomy are concatenated, the resultant code summarizes semantic information about each meaning of each term at all the levels of abstraction provided by the taxonomy. This concatenated semantic code is referred to as the "semcode" For example, a taxonomy may comprise 11 codes at level 6, up to 10 codes for any L5 within a given L6, up to 14 codes for any L4 within a given L5, up to 23 codes for any L3 within a given L4, up to 74 codes for any L2 within a given L3, and up to 413 codes for any L1 within a given L2. Thus, in this example, each meaning of each term populating the taxonomy can be uniquely identified by a numeric semcode of the format A.B.C.D.E.F, where: A, representing L6, may be 1 to 11; B, representing L5, may be 1-10; C, representing L4, may be 1 to 14; D, representing L3, may be 001 to 23; E, representing L2, may be 1 to 74; and F, representing L1, may be 1 to 413. Since semcodes uniquely identify each of the particular meanings of a term, semcodes are used to represent the term during analysis to facilitate determination of term meaning and relevance within a document. This distribution of semcode elements in this example is illustrated below.

**Table 2**

| | | |
|---|---|---|
| A | (L6) = 1 to 11 | (total number of L6 codes is 11) |
| B | (L5) = 1 to 10 | (total number of L6+L5 codes is 43) |
| C | (L4) = 1 to 14 | (total number of L6+L5+L4 codes is 183) |
| D | (L3) = 1 to 23 | (total number of L6+L5+L4+L3 codes is 1043) |
| E | (L2) = 1 to 74 | (total number of L6+L5+L4+L3+L2 codes is 11,618) |
| F | (L1) = 1 to 413 | (total number of L6+L5+L4+L3+L2+L1 codes is 50,852) |

Total number of semcodes in the exemplary taxonomy: 263,358

[0122] An example of a suitable taxonomy for the present invention is based upon the 1977 Edition of Roget's International Thesaurus. This taxonomy has been substantially modified to apply the same semantic categories to multiple parts of speech, i.e., nouns, verbs and adverbs are assigned the same value (Roget's Thesaurus separates verbs, nouns and adjectives into different, part-of-speech-specific semantic categories). Certain parts of speech have no semantic significance and in this example have been removed from the taxonomy, e.g., all adverbs, prepositions, conjunctions, and articles (e.g., "quickly," "of," "but," and "the"). These words may be useful for parsing a sentence but are otherwise treated as semantically null "stop words." Additionally, new terms, properly encoded, may be added to the current population of terms, and new concepts may be added to the taxonomy.

[0123] The Roget taxonomy may also be further modified to include a supplemental, semantico-syntactic ontology, as in the present example. Words and phrases populating the taxonomy may also be assigned lexical features to which parametric weighting factors may be assigned by the processor preparatory to analysis. These features and modifications are more fully described below.

[0124] The top-level concepts in a taxonomy suitable for use by the present invention may not provide very useful categories for classifying documents. For example, a exemplary taxonomy may employ top level (i.e., L6) concepts of: Relations, Space, Physics, Matter, Volition, Affections, etc. This level may be useful to provide a thematic categorization of documents. However, one aspect of the present invention is to determine whether a document is more similar in terms of the included concepts to another document than to a third document, and to measure the degree of similarity. This similarity measurement can be accomplished regardless of the topic or themes of the documents, and can be performed on any number of documents.

[0125] While stop words need not be included in the taxonomy, such words in a document or query may be useful to help resolve ambiguities of adjacent terms. Stop words can be used to invoke a rule for disambiguating one or more terms to the left or right. For example, the word "the" means that the next word (i.e., the word to the right) cannot be a verb, thus its presence in a sentence can be used to help disambiguate a subsequent term that may be either a verb or a noun.

[0126] Substituting or using semcodes in place of the terms represented by the semcodes facilitates the semantic analysis of documents by a computer. For example, the broad, general relatedness of two terms can be quickly observed by comparing their semcodes starting with the first digit, representing L6 of the taxonomy. The further to the right in the semcode in which there is a match, the more closely related the terms are semantically. Thus, the semcode digits convey semantic closeness information with no further need to look up other values in a table. Semcodes provide a single, unified means whereby the semantic distance between all terms and concepts can be readily identified and measured. This ability to measure semantic distance between concepts is important to inter-document similarity measures, as well as to the process of resolving meanings of terms and calculating their relative importance to a text.

[0127] An example of a taxonomic analysis of three principal meanings of the word "sound" is provided below. As the tables below illustrates, the noun "sound" may be noise (Table 2), a geographic feature (Table 3) or a medical instrument (Table 4) (This example does not include other meanings, such as the verb "sound" as in to measure depth, or the adjective "sound" as in "of sound mind").

Table 3

| | |
|---|---|
| L6 | SENSATION (5) |
| L5 | HEARING (6) |
| L4 | Sound (2) |
| L3 | SOUND (450) |
| L2 | sound (1) |
| L1 | 1 |
| Word Type | Noun |
| Word | **sound** |
| | |

Table 4

| | |
|---|---|
| L6 | MATTER (4) |
| L5 | INORGANIC MATTER (2) |
| L4 | Liquids (3) |
| L3 | INLET, GULF (399) |
| L2 | inlet (1) |
| L1 | 7 |
| Word Type | Noun |
| Word | **sound** |

Table 5

| | |
|---|---|
| L6 | VOLITION (7) |
| L5 | CONDITIONS (2) |
| L4 | Health (3) |
| L3 | THERAPY (689) |
| L2 | medical and surgical instrument (36) |
| L1 | 99 |
| Word Type | Noun |
| Word | **sound** |

**[0128]** This example illustrates an important advantage of using semcodes provided by a proper taxonomy for analyzing the meaning of terms, instead of the literal terms themselves—terms with different meanings reflecting different concepts will have completely different (i.e., non-overlapping) semcodes. Thus, the semcode identifies the particular concept intended by the term in the sentence, while the term itself may be ambiguous. This encoding of semantic information within an identifier (i.e., semcode) for the various meanings of a term thus facilitates resolution of the term's appropriate meaning in the sentence, and sorting, indexing and comparing concepts within a document, even when the terms themselves standing alone are ambiguous. Referring to the example of "sound" above, it can be seen that without semantic context it is not possible for a computer to determine whether the noun "sound" refers to noise, a body of water, a medical instrument or any of the other meanings of "sound." Semcodes provide this context. For example, in the phrase "hear sounds," one meaning (semcode) of "hear" and one meaning (semcode) of "sounds" share common taxonomic codes from L6 to L4, a circumstance not shared by other meanings of "sound." Thus, by using the context of other semcodes in the document, it becomes possible for the processor to resolve which of the meanings (semcodes) of "sound" applies in the given sentence.

[0129] While a unique semcode corresponds to each term's meaning in the taxonomy and can be used as a replacement for a term, the entire semcode need not be used for purposes of analyzing documents and queries to determine semantic closeness. Comparing semcode digits at higher taxonomic levels allows for richer, broader comparisons of concepts (and documents) than is possible using a keyword-plus-synonym-based search. Referring to the foregoing example of "sound" in the sense of a medical instrument (semcode 7.2.3.689.36.99), another word such as "radiology," sharing just L6-L3 of that code (i.e., 7.2.3.689) would allow both terms to be recognized as relating to healing, and a term sharing down to L2, such as "scalpel," would allow two such terms to be recognized as both referring to medical and surgical instruments. Depending upon the type of analysis or comparison that is desired, semcodes may be considered at any taxonomic level, from L1 to L6. The higher the level at which semcodes are compared, the higher the level of abstraction at which a comparison or analysis is conducted. In this regard, L6 is generally considered too broad to be useful by itself..

[0130] All semcodes, along with their associated terms, senses, and additional information, are stored in semantic database 102. As one skilled in the computer arts will appreciate, semantic database 102 may utilize any of a variety of database configurations, including for example, a table of stored meanings keyed to terms, or an indexed data file wherein the index associated with each term points to a file containing all semcodes for that term. An example of a data structure for semantic database 102 is illustrated in Fig. 8. However, the present invention contemplates the use of any database architecture or data structure.

[0131] **Weighting Factors**

[0132] The present invention employs a series of weighting factors for (1) disambiguating parts of speech and word senses in a text; (2) determining the relative importance of resolved meanings (i.e., semcodes); (3) preparing a conceptual profile for the document; (4) highlighting most relevant passages within a document; and (5) producing a summarization. The calculation of information value and relevance based upon weighting factors for the senses of a term is an important element of the present invention. These weights enable a computer to determine the sense of a term, determine the information value of these senses, and determine which term senses (i.e., particular semcode) best characterize the semantic import of a text.

[0133] Weighting factors are of two kinds; (1) those relating to features of a term (term) and its possible meaning (semcode) independent of context, and (2) those relating to features of a term or meaning (semcode) which are a function of context. The former feature type include weights for part-of-speech, polysemy, semantico-syntactic class, general term and sense frequency statistics, and sense co-occurrence statistics The latter feature type includes weights for grammatical function, taxonomic association, and intra-document frequency. These weighting factors are described more fully below. Weighting factor values may alternatively vary nonlinearly, such as logarithmically or exponentially. In various embodiments, the weighting factors may be adjusted, for example, parametrically, through learning, by the operator or other means.

**[0134] Part-of-Speech Weights**

[0135] Semantic database 102 may include a data field or flag, such as 804A, to indicate the part or parts of speech associated with each sense of a word (semcode). Ambiguous words may have multiple part-of-speech tags. For example, the word "sound" may have tags for noun (in the sense of noise), for verb (in the sense of measuring depth) and as adjective (in the sense of sound health) In this regard, however, a particular semcode may be associated with more than one part of speech, because the taxonomy may use the same semcode for the noun, verb and adjective forms of the same root. For example "amuse," "amusement," and "amusing" may all be assigned the same semcode associated with the shared concept. This should be distinguished from the example of the word "sound" above where the noun for noise, the verb for measuring, and the adjective for good are not related concepts, and therefore are assigned different semcodes.

[0136] Weighting factors may also be assigned to parts of speech (POS) to reflect their informational value for use in calculating document concept profiles and document summarizations. For example, nouns may convey more information in a sentence than do the verbs and adjectives in the sentence, and verbs may carry more information than do adjectives. A non-limiting example of POS weights that may be assigned to terms is shown in the following table.

**Table 6**

| POS | Weight |
|---|---|
| Noun | 1.0 |
| Verb | 0.8 |
| Adjective | 0.5 |

[0137] Like other weighting factors and parameters used in the present invention, the POS weights may be adjusted to tailor the analysis to different types of texts or genre. Such adjustments may be made, for example, parametrically, by operator input, according to the type of document being analyzed, or by means of a learning system.

[0138] Within the present invention it is anticipated that the POS weights may be adjusted to match the genre of the documents being analyzed or the genre of documents in which a comparison/search is to be conducted. For example, a number of POS weight tables may be stored, such as for technical, news, general, sports, literature, etc., that can be selected for use in the analysis or comparison/search. The POS weight table may be adjusted to reflect weights for proper nouns (e.g., names). The appropriate table may be selected by the operator by making choices in a software window (e.g., by selecting a radio button to indicate the type of document to be analyzed, compared or searched for). Appropriate POS weights for various genres may be obtained by statistically analyzing various texts, hand calculating standard texts and then adjusting weights until desired results are obtained, or training the system on standard texts. These, and any other mechanism for adjusting or selecting POS weights are contemplated by the present invention.

[0139] **Polysemy Count Weights**

[0140] Many terms have multiple different meanings. Terms with more than one part of speech are called syntactically ambiguous and terms with more than one meaning (semcode) within a part of speech are called polysemous. The present invention makes use of the observation that terms with few meanings generally carry more information in a sentence or paragraph than terms with many meanings. An example of a term with many meanings is "sound" and an example of a term with only one meaning is "tuberculosis." The presence of "tuberculosis" in a sentence generally communicates that the sentence unambiguously concerns or is related to this disease. In contrast, a sentence containing "sound" may concern or relate to any of 34 different concepts based upon the various parts of speech, meanings and associated concepts in the example taxonomy. Thus, other terms must be considered in order to determine whether "sound" in a sentence concerns noise, a body of water, determining depth, a state or quality, or any of the other concepts encompassed by "sound." In other words, the concept conveyed by "sound" in a sentence has many possible syntactic and semantic interpretations which must be disambiguated based upon other terms in the sentence and/or other contextual information. Consequently, the "polysemy count" of "sound" is high and its inherent information content is low.

[0141] The "polysemy count weight" (wPCT) is a weighting factor used in the present invention for assigning a relative information value based on the number of different meanings that each term has. The polysemy count and the polysemy count weighting factor are in inverse relation, with low polysemy counts translating to high polysemy count weights. As the preceding paragraph illustrates, terms with one or just a few different meanings have a low polysemy count and are therefore assigned a high polysemy count weight indicative of high inherent information value, since such terms communicate just one or few meanings and are hence relatively unambiguous. In contrast, terms with many different meanings have a high polysemy count and are therefore assigned a low polysemy count weight, indicative of low inherent information value and high degree of ambiguity, since such terms by themselves communicate all their meanings and additional context must be considered in order to determine the concepts communicated in the sentence. Thus, the word "tuberculosis," with only one meaning, is assigned the highest polysemy count weight, and "sound" with many meanings is assigned a relatively low polysemy count weight.

[0142] The polysemy count weighting factor may be assigned using a look up table or factors directly associated with terms populating a taxonomy. An example of a polysemy weighting factor table is provided the table below in which a step function of weights are assigned based upon the number of meanings of a term.

**Table 7**

| PolysemyCount Range Attribute | Value |
|---|---|
| 1 to 3 | 1.0 |
| 4 to 7 | 0.8 |
| 8 to 12 | 0.6 |
| 13 to 18 | 0.4 |
| 19 and above | 0.2 |

**[0143] Semantico-Syntactic Codes, Classes, and Weights**

[0144] Developers of machine translation technologies recognized that the meaning of a term in a given sentence depends upon its semantic associations within that sentence as well as its syntactic (grammatical) function. Natural language exhibits immense complexity due in part to this interplay of semantics and syntax in sentences. While humans easily master such complexities, their presence poses formidable barriers to computer analysis of natural language. To cope with such complexity, the present invention employs a system of semantico-syntactic codes assigned to terms. These semantico-syntactic codes capture the syntactic implications of a term's semantics, and thus help in reducing syntactic ambiguity during analysis of a text. For example, consider the two phrases "gasoline pump" and "pump gasoline." It is clear to a human that the term "pump" has entirely different syntactic function in each phrase. To enable a computer to decide that "pump" is a verb, not a noun, in "pump gasoline", the invention will make use of the "mass noun" semantic-syntactic code assigned to "gasoline." One of the properties of a "mass noun" is that the noun can be the object of a verb despite its singular form. This is not true of "count" nouns, e.g., in the phrase "pump meter,' the term "pump" would not be allowed as a verb where the "count" noun "meter" is singular.

[0145] An arrangement of the terms of a language in an hierarchical, semantico-syntactic ordering may be regarded as an ontology or taxonomy. An example of such an ontology is the Semantico-Syntactic Abstract Language developed by the inventor at Logos Corporation for the Logos Machine Translation System, and which has been further adapted for the present invention. The Semantico-Syntactic Abstract Language is described in "The Logos Model: An Historical Perspective," Machine Translation 18:1-

72, 2003, by Bernard Scott, which is hereby incorporated by reference in its entirety. The Semantico-Syntactic Abstract Language (SAL) reduces all parts of speech to roughly 1000 elements, referred to as "SAL codes." This SAL taxonomy focuses on how semantics and syntax intersect (i.e., on the syntactic implications of various semantic properties, as was seen in the example, above).

[0146] An additional use of the SAL codes in the present invention concerns assignment of a relative information value to terms of a document. For this purpose, the SAL codes have been modified to form a Semantico-Syntactic Class (SSC), each member of which makes a different weight contribution to the calculation of a term's information value. For example, consider the phrase "case of measles." Intuitively, one recognizes that in this phrase, "measles" carries inherently higher information value than does the word "case." The different Semantico-Syntactic Classes and their associated weights assigned to "case" and "measles" in the present invention allows the system to compute lower information values for "case" and higher values for "measles." In this example, "case" is an 'Aspective' type noun in the SAL ontology, which translates to Class E in the SSC scheme of the present embodiment, with the lowest possible weight. "Measles" in the SAL ontology is a "Condition" type noun which translates to Class A in the SSC scheme, with the highest possible weight. Thus, the SSC weight contributes significantly to determining relative information values of terms in a document. Other examples of "Aspective" nouns that translate into very low SSC weights are "piece" as in "piece of cake," and "row" as in "row of blocks." In all these examples, the words "case", "piece" and "row" all convey less information in the phrase than the second noun "measles", "cake" and "blocks". Thus, "Aspective" nouns are a class that are assigned a lower Semantico-Syntactic class weighting factor.

[0147] The Semantic-Syntactic Class (SSC) weight is also useful in balancing word-count weights for common words that tend to occur frequently. For example, in a document containing dialog, like a novel or short story, the verb "said" will appear a large number of times. Thus, on word count alone, "said" may be given inappropriately high information value. However, "said" is a Semantico-Syntactic Class E word and thus will have very low information value as compared to other words in the statement that was uttered. Thus, when the Semantico-Syntactic Class weighting factors are applied, the indicated information value of "said" will be reduced compared to these other words, despite its frequency.

[0148] As shown in Fig. 8, the SSC, such as 812B, for each term is stored in semantic database 102. An example of weights for the various SSC classes is 1806, shown in Fig. 18. Like other weighting factors and parameters used in the present invention, the SSC weights may be adjusted to tailor the analysis to different types of texts or genre. Such adjustments may be made, for example, parametrically, by operator input, according to the type of document being analyzed, or by means of a learning system.

[0149] **Word and Concept Frequency Statistics and Weights.**

[0150] The frequency at which terms (words/phrases) and concepts (semcodes) appear both in general usage and in a document being analyzed convey useful data regarding the information value of the term or concept. The present invention takes advantage of such frequency statistics in computing the information value of terms and concepts. Frequency statistics are considered under three aspects: (1) frequency of terms and concepts (semcodes) in general usage; (2) frequency of terms and concepts (semcodes) in a given document; (3) the statistical relationship between concepts considered in (1) and (2). In the following discussion, "terms" refer exclusively to open-class words and phrases, such as nouns, verbs, and adjectives; "concepts" refer to their respective semcodes.

[0151] Concerning frequency in general usage, the present invention assumes that there is an inverse relationship between frequency of a term or concept in general usage and its inherent information value. For example, the adjectives "hot" and "fast" are common, frequently occurring terms, and thus provide information of less value compared to the more infrequent "molten" and "supersonic." Thus, the relative frequency of a term or concept in general usage can be used in assigning a weighting factor in developing the information value of the associated semcodes.

[0152] Concerning frequency within a given document, the present invention assumes there is a direct relationship between the frequency of a term in a particular document and the information value of term to that document For example, a document including the word "tuberculosis" a dozen times is more likely to be more concerned with that disease than is a document that mentions the word only once. Accordingly, computation of the informational value of "tuberculosis" to the document must reflect such frequency statistics. In sum, the frequency of a term or semcode in a document conveys information regarding the relevance of that term/semcode to the content of the document. Such semcode frequency information may be in the form of a count, a fraction (e.g., number of times appearing / total number of terms in the document), a percentage, a logarithm of the count or fraction, or other statistical measure of frequency.

**[0153]** A third statistical frequency weighting factor to be used in computing the information value of a term or concept (semcode) may be obtained by comparing the frequency of the term or concept in a document with the frequency of the term or concept in general use in the language or genre. For example, in a document where the word "condition" appears frequently and the word "tuberculosis" appears far less frequently, the conclusion could be falsely drawn that 'condition' was the informationally more valuable term of the two. This incorrect conclusion can be avoided by referring to real-world usage statistics concerning the relative frequency of these two words. Such statistics would indicate that the word 'tuberculosis' occurs far more infrequently than does "condition," allowing the conclusion that, even if the word "tuberculosis" appears only once or twice in a document, its appearance holds more information value than the word "condition." Thus, weights based on relative real-world usage statistics may help offset the misleading informational value of more frequently occurring terms.

**[0154]** Comparative frequency measures require knowledge of the average usage frequency of terms in the language or particular genre or domain of the language. Frequency values for terms can be obtained form from public sources, or calculated from a large body of documents or corpus. For example, one corpus commonly used for measuring the average frequency of terms is the Brown Corpus, prepared by Brown University, which includes a million words. Further information regarding the Brown Corpus is available in the Brown Corpus Manual available at:

<http://helmer.aksis.uib.no/icame/brown/bcm.html>. Another corpus suitable for calculating average frequencies is the British National Corpus (BNC) which contains 100 million words and is available at http://www.natcorp.ox.ac.uk/.

[0155] This term-frequency weighting factor (term statistics weight or wTSTAT) may be assigned to a term or calculated based upon the frequency statistics by any number of methods known in the art, including, for example, a step function based upon a number range or fraction above or below the average, a fraction of the average (e.g., frequency in document/frequency in corpus), a logarithm of a fraction (e.g., log(freq. in document/frequency in corpus)), a statistical analysis, or other suitable formula. For example, wTSTAT may be derived from the classic statistical calculation:

$$\text{wTSTAT} = \text{tf} * \text{idf}, \qquad \text{Eq. 1}$$

where tf is the frequency of the term in a given document; and

idf is the inverse frequency of the term in general use.

[0156] Comparable resources for global statistics on semcode frequency, as opposed to statistics on term frequency, are not generally available, and may be developed inductively as a by-product of the present invention and its ability to resolve semantically ambiguous terms to a specific meaning (semcode).

[0157] Instead of general-usage frequency measures, term/semcode frequency statistics may be calculated for a particular set of documents, such as all documents within a particular library or database to be indexed and/or searched. Such library-specific frequency statistics may provide more reliable indicators of the importance of particular terms/semcodes in documents within the library or database

[0158] **Concept Co-Occurrence Statistics**

[0159] An additional use of general language statistics in the present invention concerns semcode co-occurrence statistics. Semcode co-occurrence statistics here refer to the relative frequency with which each of the concepts (semcodes) in the entire taxonomy co-occurs with each of the other concepts (semcodes) in the taxonomy. Statistical data on semcode co-occurrence in general, or in specific genre or domains, may typically be maintained in a two dimensional matrix optimized as a sparse matrix. Such a semcode co-occurrence matrix may comprise semcodes at any taxonomic level, or, to keep matrix size within more manageable bounds, at taxonomic levels 2 or 3 (L2 or L3). For example, a co-occurrence matrix maintained at L3 may be a convenient size, such as approximately 1000 by 1000.

[0160] Statistical data on the co-occurrence of semcodes may be used by the present invention to help resolve part-of-speech and term-sense ambiguities. For example, assume that a given term in a document has two unresolved meanings (semcodes), one of which the invention must now attempt to select as appropriate to the context. One of the ways this is done in the present invention is to consider each of the two unresolved semcodes in relationship to each of the semcodes of all other terms in the document, and then to consult the general co-occurrence matrix to determine which of these many semcode combinations (co-occurrences) is statistically more probable. The semcode combination (semcode co-occurrence) in the document found to have greater frequency in the general semcode co-occurrence matrix will be given additional weight when calculating its information value, which value in turn provides the principal basis for selecting among the competing semcodes. Thus, general co-occurrence statistics may influence one of the weighting factors used in resolution of semantic ambiguity.

**[0161]** Statistics on the co-occurrence of semcodes are not generally available, but may be compiled and maintained inductively through the on-going processing of large volumes of text. To collect these co-occurrence statistics, the analysis of corpora is at the level of term meaning, where such statistics are not generally available and no reliable automated method has been established for this purpose.

**[0162]** Term and semcode frequency analysis is also useful when assessing the relevance of documents to a search query or comparing the relatedness of two documents. In such comparisons, the co-occurrence of terms or semcodes provides a first indication of relevance or relatedness based on the assumption that documents using the same words concern the same or similar concepts. A second indication of relevance or relatedness is the relative information value of shared concepts. As stated, one possible weighting factor in calculating information value is local and global frequency statistics.

**[0163]** Frequency analysis may be applied to semcodes above level 1 (L1) to yield a frequency of concepts statistic. To calculate such statistics, semcodes for the terms in the corpus must be determined at the level (e.g., L2 or L3). In essence, this weighting factor reflects the working assumption that concepts appearing more frequently in a document than in common usage are likely to be of more importance regarding the subject matter of the document than other concepts that occur at average or less than average frequency. Using such semcode frequency statistics may resolve the small-number (sparseness) problem associated with rarely used synonyms of common concepts. Semcode-frequency weighting factors may be calculated in a manner similar to term frequency as discussed above.

[0164] As a further alternative, various combinations of term frequency and L1, L2 or L3 semcode frequency weighting factors may be applied. Applying weighting factors for both term frequency and semcode (i.e., concept) frequency at one or more of these levels would permit the resulting information value to be useful for disambiguating the meanings (semcodes) of terms, for profiling a document, for comparing the relatedness of two documents for ranking purposes, and for document summarization. In some genres, such as technical, scientific and medical publications, frequency statistics may be most useful for identifying important concepts, particularly when a word reflects a term of art or a formal descriptor of a particular concept.

[0165] **Grammatical Function Weights**

[0166] Not all terms of a text have equal information value to the semantic import of that text. One feature that further serves to differentiate the information value of a term in a text is the term's grammatical function. Grammatical function (GF) is the syntactic role that a particular term plays within a sentence. For example, a term may be (or be part of) the object, subject, predicate or complement of a clause, and that clause may be a main clause, dependent clause or relative clause, or be a modifier to the subject, object predicate or complement in those clauses. Further, a weighting factor reflecting the grammatical function of a word is useful in establishing term information value for purposes of profiling, similarity-ranking and summarization of texts.

[0167] A GF tag may be assigned to a term to identify the grammatical function played by a term in a particular sentence. Then, a GF weight value may be assigned to terms to reflect the grammatical information content of the term in the sentence based upon its function. During the processing of a document according to an embodiment of the present invention, parser 104 will add a tag to each term to indicate its grammatical function in the parsed sentence (step 202 of Fig 2).

[0168] After parsing, a GF weight may be assigned to the term, such as by looking the GF tag up in a data table and assigning the corresponding weighting factor. For example, GF weights may be assigned to the grammatical functions parametrically, by learning, by operator input, or according to genre. The relative information content for various grammatical functions may vary based upon the genre (e.g., scientific, technical, literature, news, sports, etc.) of a document, so different GF weights may be assigned based upon the type of document in order to refine the analysis of the document. As another example, the GF weight value may be assigned by a parser at the time the sentence is analyzed and the grammatical function is determined.

[0169] It is important to note that while the present invention has been described in the context of a fully functioning data processing system, those of ordinary skill in the art will appreciate that the processes of the present invention are capable of being distributed in the form of a computer readable medium of instructions and a variety of forms and that the present invention applies equally regardless of the particular type of signal bearing media actually used to carry out the distribution. Examples of computer readable media include recordable-type media such as floppy disc, a hard disk drive, RAM, and CD-ROM's, as well as transmission-type media, such as digital and analog communications links.

**[0170]** Although specific embodiments of the present invention have been described, it will be understood by those of skill in the art that there are other embodiments that are equivalent to the described embodiments. Accordingly, it is to be understood that the invention is not to be limited by the specific illustrated embodiments, but only by the scope of the appended claims.

# CLAIMS

What is claimed is:

1. A method of semantic profiling of documents comprising:

receiving a document to be profiled, the document comprising a plurality of terms;

for each of at least a portion of the plurality of terms in the document:

determining a part of speech and a grammatical function of the term,

obtaining senses of the term,

selecting a sense as a most likely meaning of the term, and

calculating an information value of the term; and

generating a semantic profile of the document comprising at least some of the calculated information values.

2. The method of claim 1, wherein a sense is selected as a most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document.

3. The method of claim 2, wherein a sense is selected as a most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document by:

determining at what level of a semantic taxonomy hierarchy the senses of each term are related to the senses of the other terms; and assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to senses of the other terms.

4. The method of claim 3, wherein the lower the level at which each sense of the term is related to senses of the other terms, the higher the weighting assigned to the relation.

5. The method of claim 3, wherein the determination of taxonomic relationships of the senses of the term with senses of other terms in the document is done using a moving window based on the term.

6. The method of claim 3, wherein a sense is selected as a most likely meaning of the term based on statistics of co-occurrence of senses at a selected taxonomic level.

7. The method of claim 6, wherein a sense is selected as a most likely meaning of the term based on a most probable meaning of each sense based on statistics of occurrences of meaning.

8. The method of claim 7, wherein a sense is selected as a most likely meaning of the term by:

selecting as the most likely meaning of the term either:

the sense determined to be the most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document; or

the sense determined to be the most likely meaning of the term based on statistics of co-occurrence of senses at a selected taxonomic level; or

the sense determined to be the most likely meaning of the term based on the most probable meaning of each sense based on statistics of occurrences of meaning.

9. The method of claim 1, wherein the information value is calculated by:

obtaining information relating to the selected sense of the term;

weighting the information relating to the selected sense of the term; and

calculating the information value based on the weighted information relating to the selected sense of the term.

10. The method of claim 9, wherein the information relating to the selected sense of the term comprises semantic and syntactic information relating to the term and to the selected sense of the term.

11. The method of claim 9, wherein the information relating to the selected sense of the term comprises at least one of:

information identifying a part of speech of the term;

information identifying a grammatical function of the term;

information identifying a semantico-syntactic class of the term; and

information identifying a polysemy count of the term.

12. The method of claim 11, wherein the information relating to the selected sense of the term further comprises:

a level of a semantic taxonomy hierarchy at which the senses of each term are related to selected senses of other terms in the document.

13. The method of claim 12, wherein the information relating to the selected sense of the term is weighted by:

applying a weighting to the at least one of information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, and information identifying a polysemy count of the term; and

assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to a sense of at least one other term.

14. The method of claim 9, wherein the information relating to the selected sense of the term comprises:

information identifying a part of speech of the term;

information identifying a grammatical function of the term;

information identifying a semantico-syntactic class of the term;

information identifying a polysemy count of the term; and

a level of a semantic taxonomy hierarchy the senses of each term are related.

15. The method of claim 14, wherein the information relating to the selected sense of the term is weighted by:

applying a weighting to the information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, and information identifying a polysemy count of the term; and

assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to a sense of at least one other term.

16. The method of claim 15, wherein the information value is calculated according to:

$$PICW = wPOS\frac{(wPCT + wSSC)}{2}\sum_{i=N}^{N} wGF\,,$$

$$ICW = PICW \times TW\text{, and}$$

$$InfoVal = \frac{\sum_{i=N}^{N} ICW_i}{\sum ICW}\,;$$

wherein wPOS is the weighted information identifying a part of speech of the term, wGF is the weighted information identifying a grammatical function of the term, wSSC is the weighted information identifying a semantico-syntactic class of the term, wPCT is the weighted information identifying a polysemy count of the term, and TW is the taxonomic weight.

17. A method for performing document-based searching comprising:

receiving a search document, the search document comprising a plurality of terms;

for each of at least a portion of the plurality of terms in the search document:

determining a part of speech and a grammatical function of the term,

obtaining senses of the term,

selecting a sense as a most likely meaning of the term, and

calculating an information value of the term; and

generating a semantic profile of the search document comprising at least some of the calculated information values; and

accessing a database comprising a plurality of semantic profiles of documents to retrieve documents having semantic profiles that are similar to the semantic profile of the search documents, each semantic profile in the database comprising a plurality of information values of terms included in the document.

18. The method of claim 17, wherein a sense is selected as a most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document.

19. The method of claim 18, wherein a sense is selected as a most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document by:

determining at what level of a semantic taxonomy hierarchy the senses of each term are related to the senses of the other terms; and

assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to senses of the other terms.

20. The method of claim 19, wherein the lower the level at which each sense of the term is related to senses of the other terms, the higher the weighting assigned to the relation.

21. The method of claim 19, wherein the determination of taxonomic relationships of the senses of the term with senses of other terms in the document is done using a moving window based on the term.

22. The method of claim 19, wherein a sense is selected as a most likely meaning of the term based on statistics of co-occurrence of senses at a selected taxonomic level.

23. The method of claim 22, wherein a sense is selected as a most likely meaning of the term based on a most probable meaning of each sense based on statistics of occurrences of meaning.

24. The method of claim 23, wherein a sense is selected as a most likely meaning of the term by:

selecting as the most likely meaning of the term either:

the sense determined to be the most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document; or

the sense determined to be the most likely meaning of the term based on statistics of co-occurrence of senses at a selected taxonomic level; or

the sense determined to be the most likely meaning of the term based on the most probable meaning of each sense based on statistics of occurrences of meaning.

25. The method of claim 17, wherein the information value is calculated by:

obtaining information relating to the selected sense of the term;

weighting the information relating to the selected sense of the term; and

calculating the information value based on the weighted information relating to the selected sense of the term.

26. The method of claim 25, wherein the information relating to the selected sense of the term comprises semantic and syntactic information relating to the term and to the selected sense of the term.

27. The method of claim 25, wherein the information relating to the selected sense of the term comprises at least one of:

information identifying a part of speech of the term;

information identifying a grammatical function of the term;

information identifying a semantico-syntactic class of the term; and

information identifying a polysemy count of the term.

28. The method of claim 27, wherein the information relating to the selected sense of the term further comprises:

a level of a semantic taxonomy hierarchy at which the senses of each term are related to selected senses of other terms in the document.

29. The method of claim 28, wherein the information relating to the selected sense of the term is weighted by:

applying a weighting to the at least one of information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, and information identifying a polysemy count of the term; and

assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to a sense of at least one other term.

30. The method of claim 25, wherein the information relating to the selected sense of the term comprises:

information identifying a part of speech of the term;

information identifying a grammatical function of the term;

information identifying a semantico-syntactic class of the term;

information identifying a polysemy count of the term; and

a level of a semantic taxonomy hierarchy the senses of each term are related.

31. The method of claim 30, wherein the information relating to the selected sense of the term is weighted by:

applying a weighting to the information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, and information identifying a polysemy count of the term; and

assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to a sense of at least one other term.

32. The method of claim 31, wherein the information value is calculated according to:

$$PICW = wPOS\frac{(wPCT + wSSC)}{2}\sum_{i=N}^{N} wGF,$$

$ICW = PICW \times TW$, and

)1

$$InfoVal = \frac{\sum_{i=N}^{N} ICW_i}{\sum ICW};$$

wherein wPOS is the weighted information identifying a part of speech of the term, wGF is the weighted information identifying a grammatical function of the term, wSSC is the weighted information identifying a semantico-syntactic class of the term, wPCT is the weighted information identifying a polysemy count of the term, and TW is the taxonomic weight.

33. The method of claim 17, further comprising:

ranking the retrieved documents based on similarity to the search document.

34. The method of claim 33, wherein the retrieved documents are ranked by:

comparing semantic profiles of the retrieved documents to the semantic profile of the search document.

35. The method of claim 17, further comprising:

generating a summary of each of at least some of the retrieved documents.

36. The method of claim 35, wherein the summary of each of at least some of the retrieved documents is generated by:

calculating an information value for each sentence in the document;

deleting sentences having an information value below a threshold value;

deleting non-main clauses having an information value below a threshold value;

normalizing to declarative form each remaining sentence and clause;

deleting noun phrases and verb phrases having an information value below a threshold value from each remaining sentence and clause;

selecting at least a portion of the remaining sentences and clauses as kernel phrases; and

replacing terms present in the kernel phrases with terms relating to similar concepts selected from a taxonomic hierarchy to form the summary of the document.

37. The method of claim 35, wherein terms present in the kernel phrases are replaced with terms relating to similar concepts selected from a taxonomic hierarchy by:

identifying a subject, verb, and object terms of each kernel phrase;

determining intersections of the subject, verb, and object terms at a level of the taxonomic hierarchy;

obtaining concept labels of the level of the taxonomic hierarchy at which the intersections; and

combining the obtained concept labels to form sentences that summarize the kernel phrases.

38. A method of summarizing a textual document comprising:

calculating an information value for each sentence of the document;

deleting from consideration for a summary sentences having an information value below a first threshold value to form retained sentences;

deleting from the retained sentences non-main clauses having information values below a second threshold value to form retained clauses;

normalizing the retained clauses to declarative form;

deleting modifiers having information values below a third threshold value from the normalized retained clauses to from kernel phrases;

selecting at least a portion of the kernel phrases; and

replacing at least portions of the kernel phrases with terms relating to similar concepts selected from a taxonomic hierarchy.

39. The method of claim 38, wherein the information value for a sentence is calculated by:

calculating an information value for at least some terms in the sentence;

summing the calculated information values for the terms in the sentence; and

dividing the sum by a number of the terms for which the information values were summed.

40. The method of claim 38, wherein the information value for a term is calculated by:

determining a part of speech and a grammatical function of the term,

obtaining senses of the term,

selecting a sense as a most likely meaning of the term, and

calculating an information value of the term.

41. The method of claim 40, wherein a sense is selected as a most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document.

42. The method of claim 41, wherein a sense is selected as a most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document by:

determining at what level of a semantic taxonomy hierarchy the senses of each term are related to the senses of the other terms; and

assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to senses of the other terms.

43. The method of claim 42, wherein the lower the level at which each sense of the term is related to senses of the other terms, the higher the weighting assigned to the relation.

44. The method of claim 42, wherein the determination of taxonomic relationships of the senses of the term with senses of other terms in the document is done using a moving window based on the term.

45. The method of claim 42, wherein a sense is selected as a most likely meaning of the term based on statistics of co-occurrence of senses at a selected taxonomic level.

46. The method of claim 45, wherein a sense is selected as a most likely meaning of the term based on a most probable meaning of each sense based on statistics of occurrences of meaning.

47. The method of claim 46, wherein a sense is selected as a most likely meaning of the term by:

selecting as the most likely meaning of the term either:

the sense determined to be the most likely meaning of the term based on taxonomic relationships of the senses of the term with senses of other terms in the document; or

the sense determined to be the most likely meaning of the term based on statistics of co-occurrence of senses at a selected taxonomic level; or

the sense determined to be the most likely meaning of the term based on the most probable meaning of each sense based on statistics of occurrences of meaning.

48. The method of claim 40, wherein the information value is calculated by:

obtaining information relating to the selected sense of the term;

weighting the information relating to the selected sense of the term; and

calculating the information value based on the weighted information relating to the selected sense of the term.

49. The method of claim 48, wherein the information relating to the selected sense of the term comprises semantic and syntactic information relating to the term and to the selected sense of the term.

50. The method of claim 48, wherein the information relating to the selected sense of the term comprises at least one of:

information identifying a part of speech of the term;

information identifying a grammatical function of the term;

information identifying a semantico-syntactic class of the term; and

information identifying a polysemy count of the term.

51. The method of claim 50, wherein the information relating to the selected sense of the term further comprises:

a level of a semantic taxonomy hierarchy at which the senses of each term are related to selected senses of other terms in the document.

52. The method of claim 51, wherein the information relating to the selected sense of the term is weighted by:

applying a weighting to the at least one of information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, and information identifying a polysemy count of the term; and assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to a sense of at least one other term.

53. The method of claim 48, wherein the information relating to the selected sense of the term comprises:

information identifying a part of speech of the term;

information identifying a grammatical function of the term;

information identifying a semantico-syntactic class of the term;

information identifying a polysemy count of the term; and

a level of a semantic taxonomy hierarchy the senses of each term are related.

54. The method of claim 53, wherein the information relating to the selected sense of the term is weighted by:

applying a weighting to the information identifying a part of speech of the term, information identifying a grammatical function of the term, information identifying a semantico-syntactic class of the term, and information identifying a polysemy count of the term; and

assigning a taxonomic weight to relationships among the senses based on the level of the semantic taxonomy at which each sense of a term is related to a sense of at least one other term.

55. The method of claim 54, wherein the information value is calculated according to:

$$PICW = wPOS\frac{(wPCT + wSSC)}{2}\sum_{i=N}^{N} wGF,$$

$ICW = PICW \times TW$, and

$$InfoVal = \frac{\sum_{i=N}^{N} ICW_i}{\sum ICW};$$

wherein wPOS is the weighted information identifying a part of speech of the term, wGF is the weighted information identifying a grammatical function of the term, wSSC is the weighted information identifying a semantico-syntactic class of the term, wPCT is the weighted information identifying a polysemy count of the term, and TW is the taxonomic weight.

56. The method of claim 38, wherein the information value for a clause is calculated by:

calculating an information value for at least some terms in the clause;

summing the calculated information values for the terms in the clause; and

dividing the sum by a number of the terms for which the information values were summed.

57. The method of claim 38, wherein the information value for a modifier is calculated by:

calculating an information value for at least some terms in the modifier;

summing the calculated information values for the terms in the modifier; and

dividing the sum by a number of the terms for which the information values were summed.

58. The method of claim 38, wherein the kernel phrases are selected by:

selecting as the selected kernel phrases that portion of the kernel phrases having higher information values than kernel phrases that are not selected.

59. The method of claim 58, wherein the portion of kernel phrases that are selected is a fraction or percentage of the kernel phrases.

60. The method of claim 58, wherein the portion of kernel phrases that are selected is determined so as to form a summary of a selected size.

61. The method of claim 38, wherein the kernel phrases are replaced with terms relating to similar concepts selected from a taxonomic hierarchy by:

identifying subject, verb, and object terms for each kernel phrase;

determining intersections of the identified subject, verb, and object terms at a level of the taxonomic hierarchy;

combining concept labels of the level of the taxonomic hierarchy at which the identified subject, verb, and object terms intersect to form sentences that summarize each kernel phrase; and replacing the kernel phrases with the summary sentences.

Fig. 1

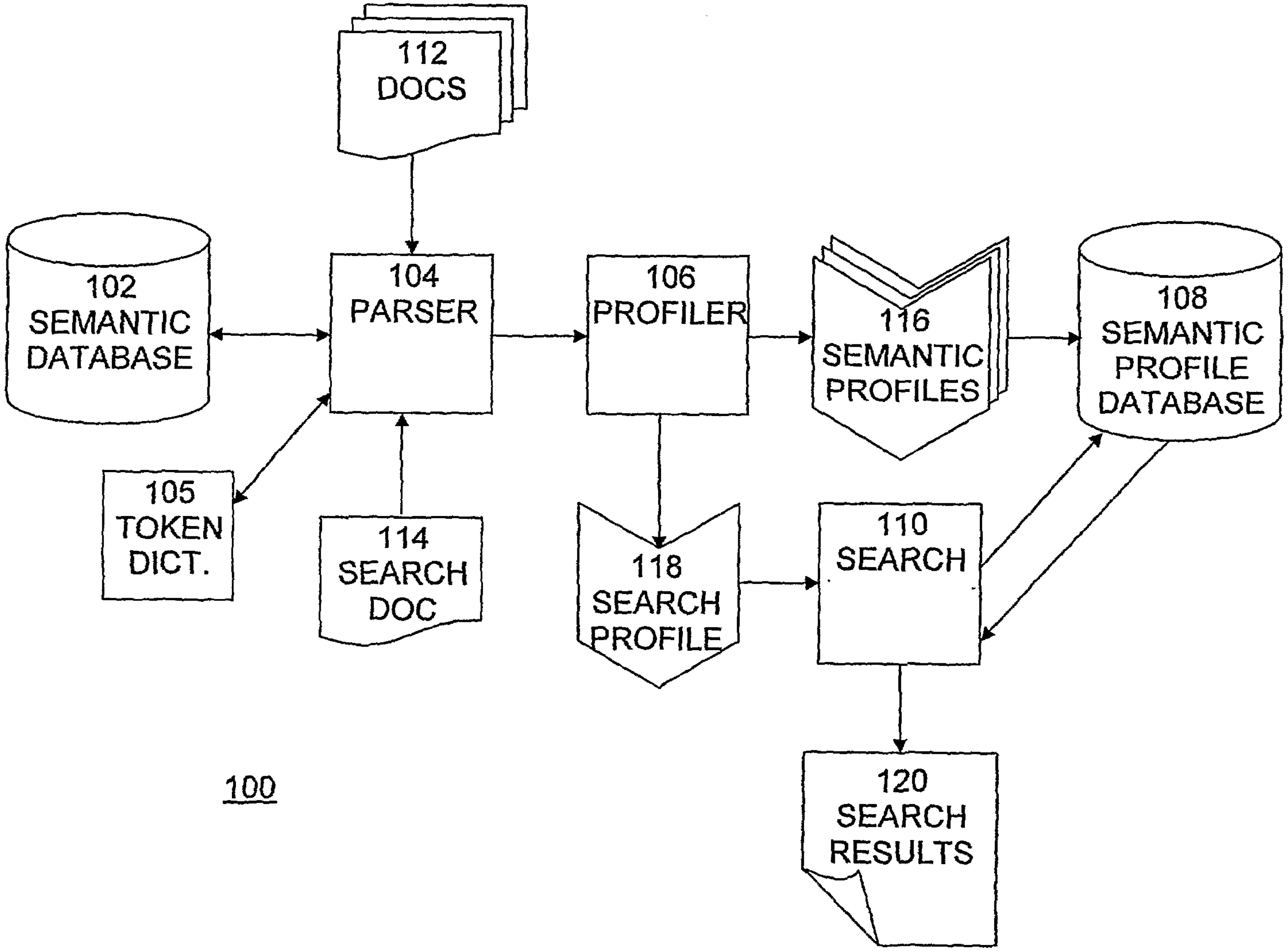

112
DOCS
102
SEMANTIC
DATABASE
104
PARSER
106
PROFILER
116
SEMANTIC
PROFILES
108
SEMANTIC
PROFILE
DATABASE
105
TOKEN
DICT.
114
SEARCH
DOC
118
SEARCH
PROFILE
110
SEARCH
120
SEARCH
RESULTS
100

Fig. 2

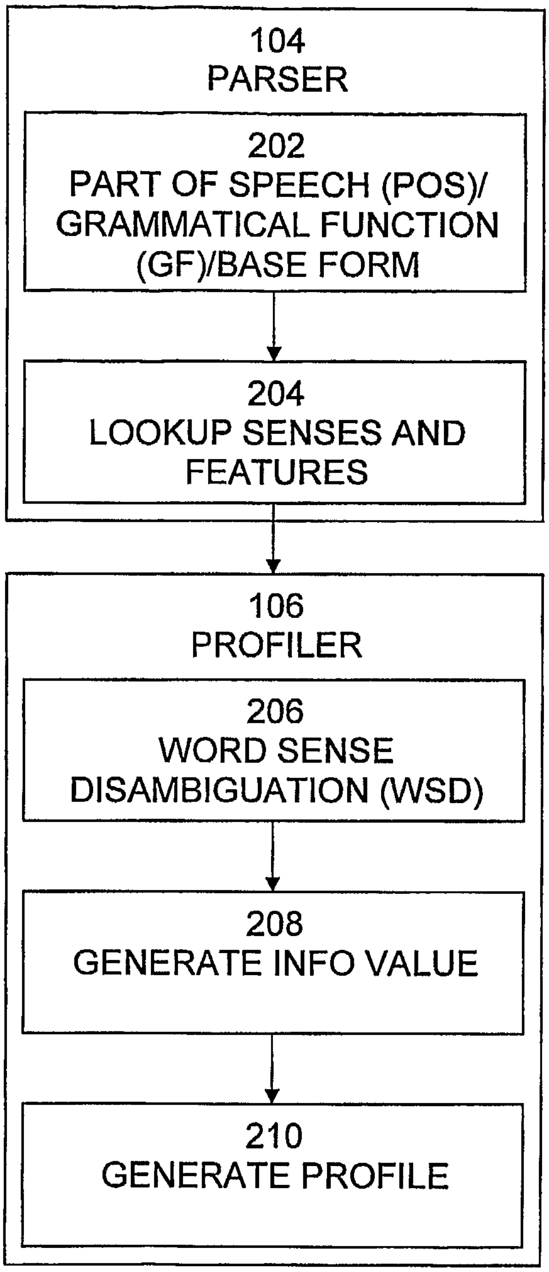

104
PARSER
202
PART OF SPEECH (POS)/
GRAMMATICAL FUNCTION
(GF)/BASE FORM
204
LOOKUP SENSES AND
FEATURES
106
PROFILER
206
WORD SENSE
DISAMBIGUATION (WSD)
208
GENERATE INFO VALUE
210
GENERATE PROFILE

Fig. 3

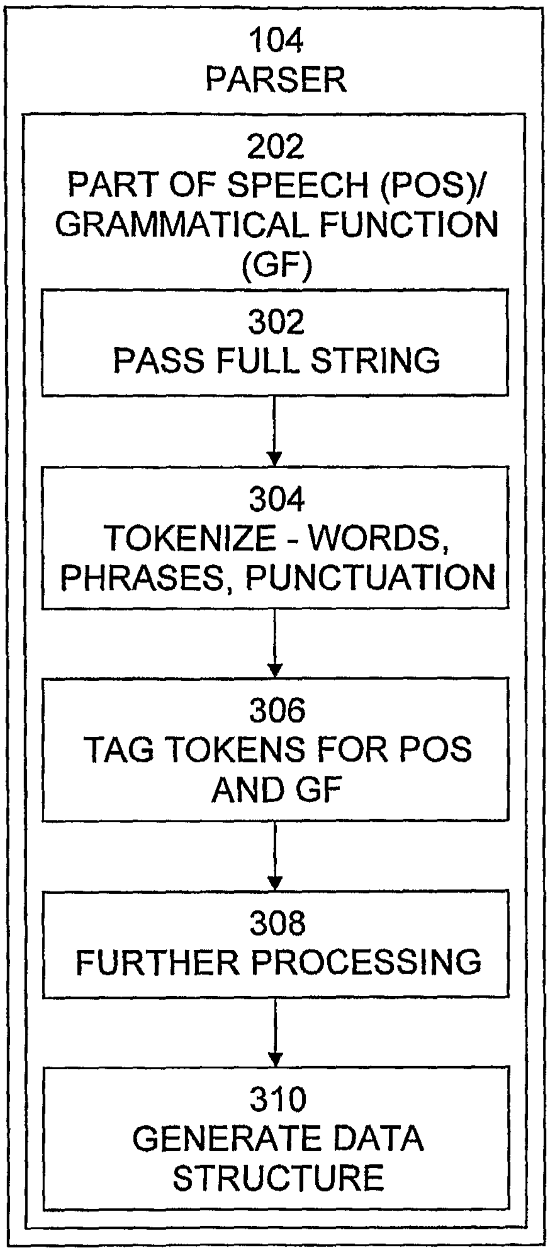

104
PARSER
202
PART OF SPEECH (POS)/
GRAMMATICAL FUNCTION
(GF)
302
PASS FULL STRING
304
TOKENIZE - WORDS,
PHRASES, PUNCTUATION
306
TAG TOKENS FOR POS
AND GF
308
FURTHER PROCESSING
310
GENERATE DATA
STRUCTURE

Fig. 4

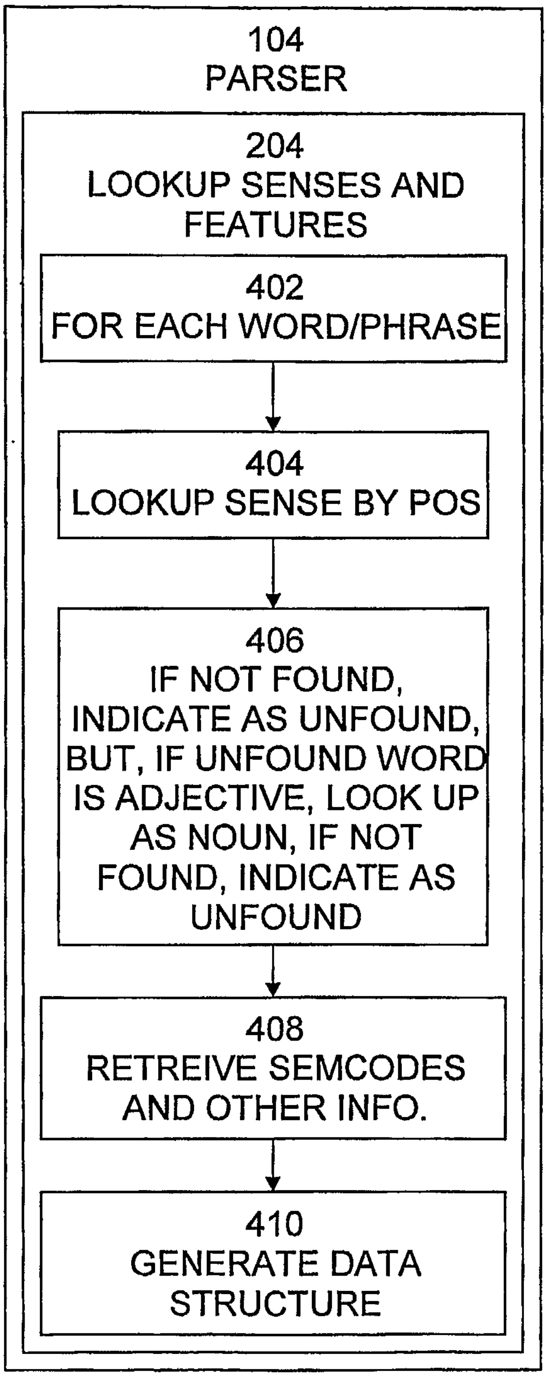

104
PARSER
204
LOOKUP SENSES AND FEATURES
402
FOR EACH WORD/PHRASE
404
LOOKUP SENSE BY POS
406
IF NOT FOUND, INDICATE AS UNFOUND, BUT, IF UNFOUND WORD IS ADJECTIVE, LOOK UP AS NOUN, IF NOT FOUND, INDICATE AS UNFOUND
408
RETREIVE SEMCODES AND OTHER INFO.
410
GENERATE DATA STRUCTURE

Fig. 5

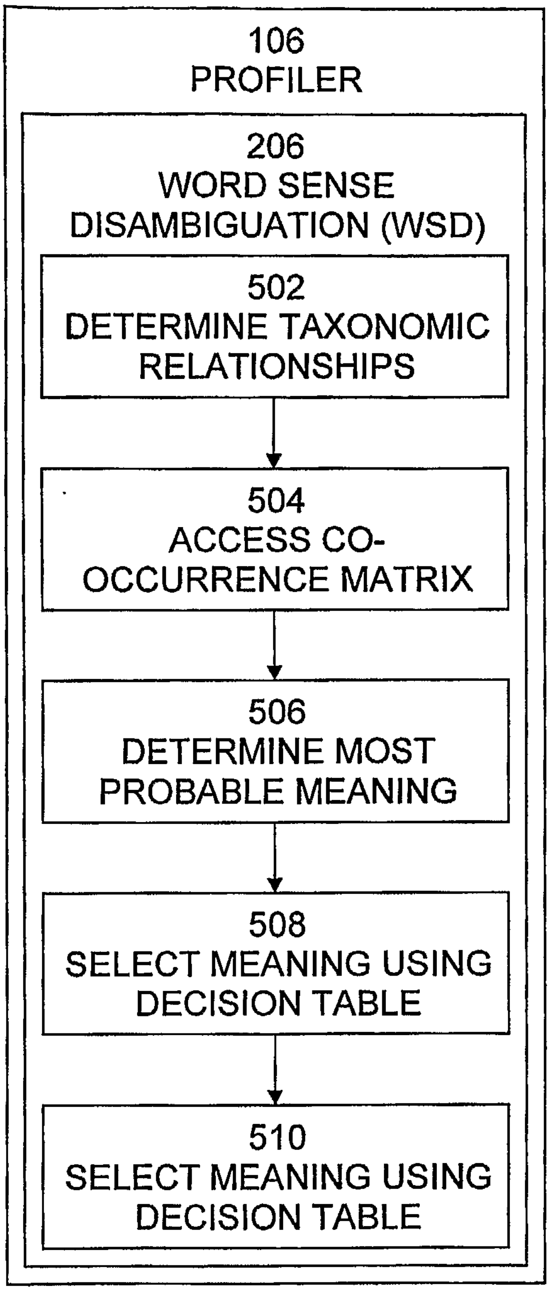

106
PROFILER
206
WORD SENSE DISAMBIGUATION (WSD)
502
DETERMINE TAXONOMIC RELATIONSHIPS
504
ACCESS CO-OCCURRENCE MATRIX
506
DETERMINE MOST PROBABLE MEANING
508
SELECT MEANING USING DECISION TABLE
510
SELECT MEANING USING DECISION TABLE

Fig. 6

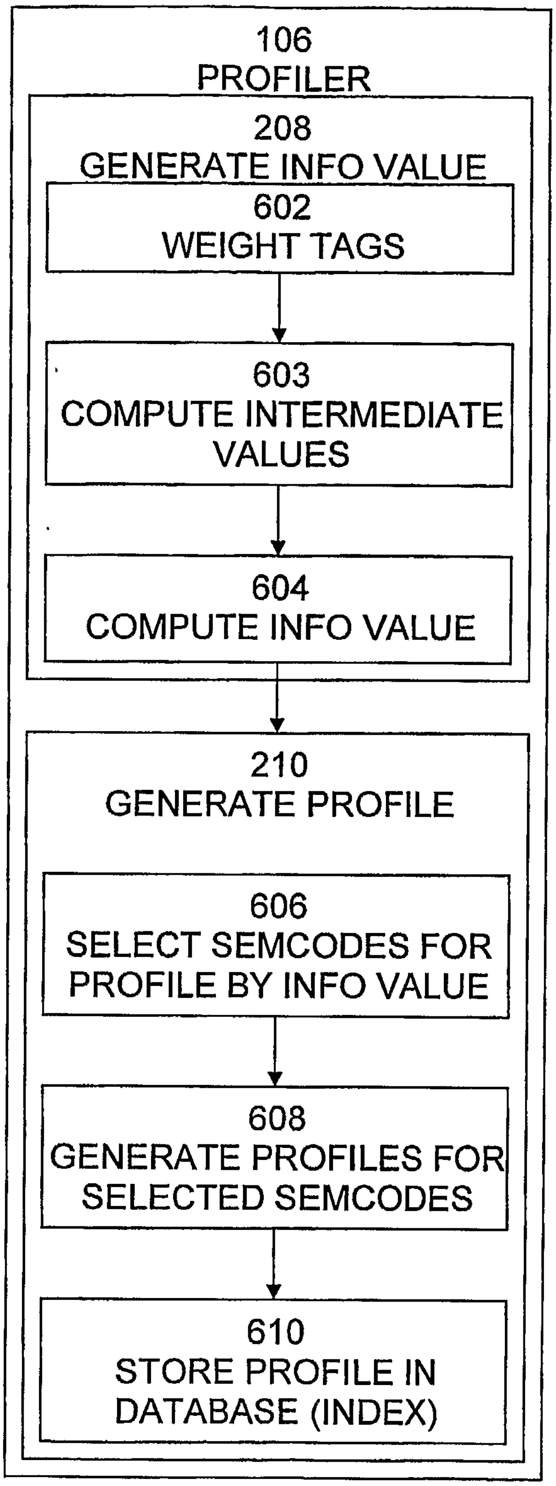
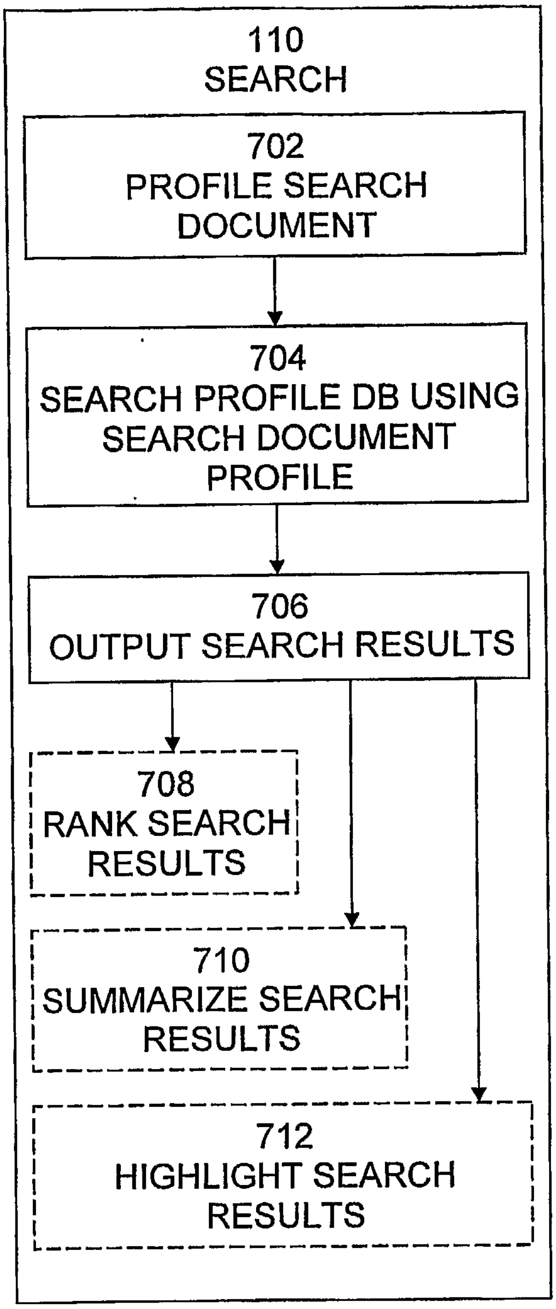
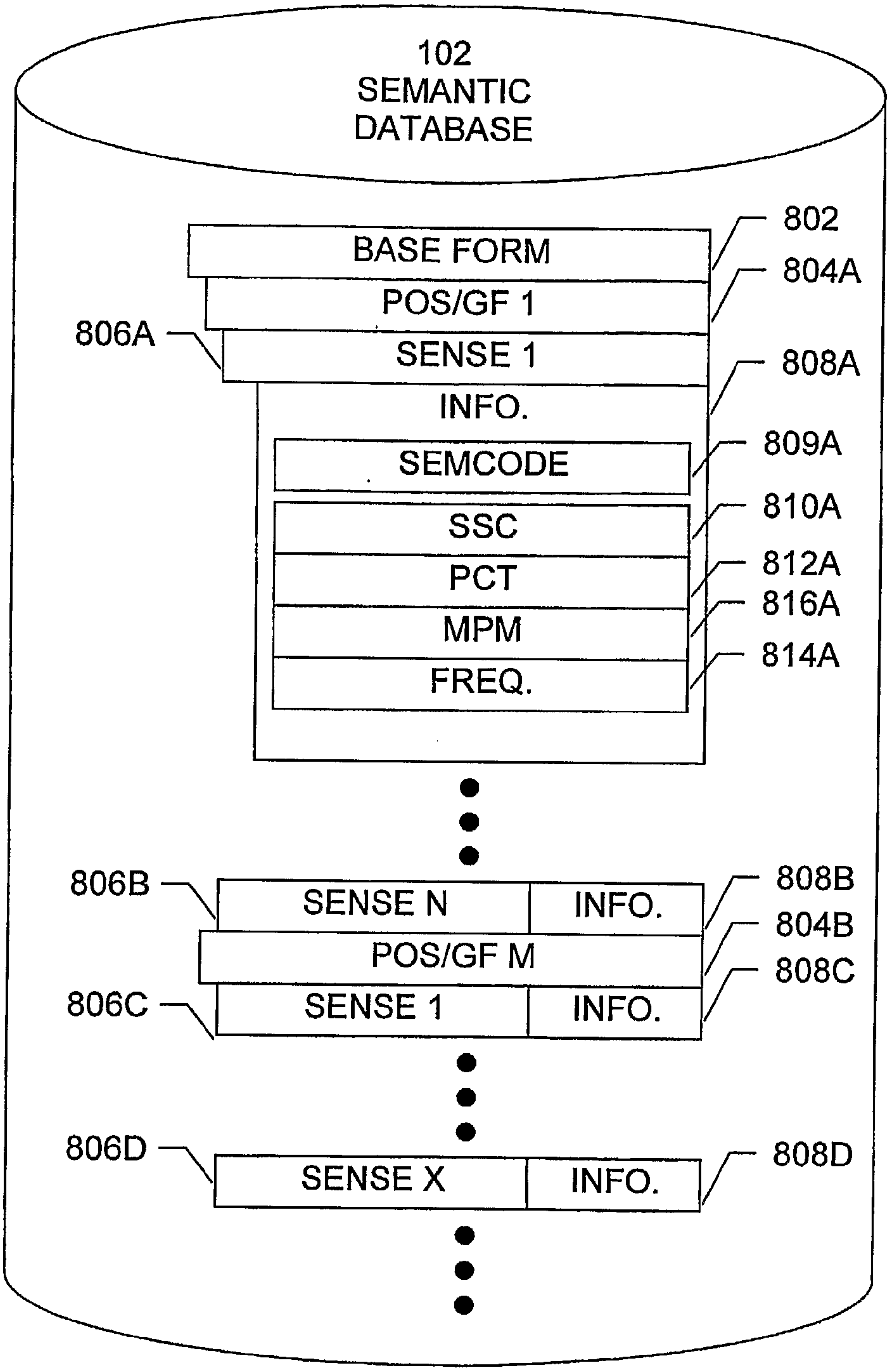

106
PROFILER
208
GENERATE INFO VALUE
602
WEIGHT TAGS
603
COMPUTE INTERMEDIATE VALUES
604
COMPUTE INFO VALUE
210
GENERATE PROFILE
606
SELECT SEMCODES FOR PROFILE BY INFO VALUE
608
GENERATE PROFILES FOR SELECTED SEMCODES
610
STORE PROFILE IN DATABASE (INDEX)

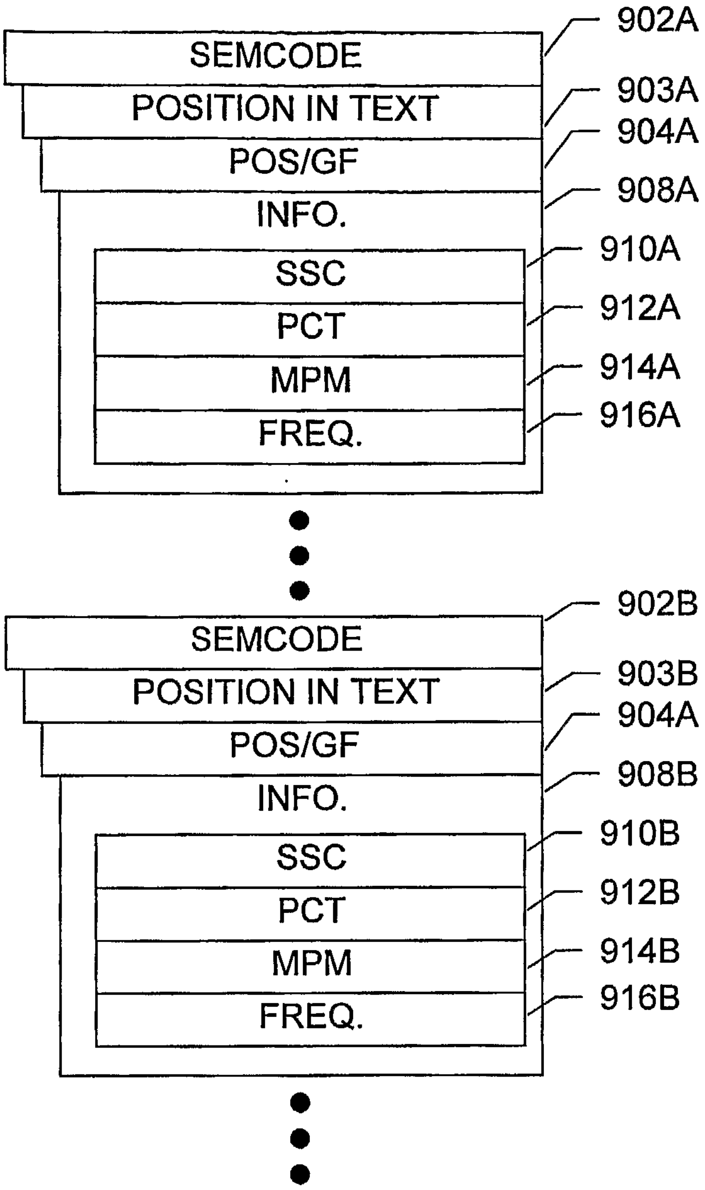
Fig. 7


110
SEARCH
702
PROFILE SEARCH DOCUMENT
704
SEARCH PROFILE DB USING SEARCH DOCUMENT PROFILE
706
OUTPUT SEARCH RESULTS
708
RANK SEARCH RESULTS
710
SUMMARIZE SEARCH RESULTS
712
HIGHLIGHT SEARCH RESULTS

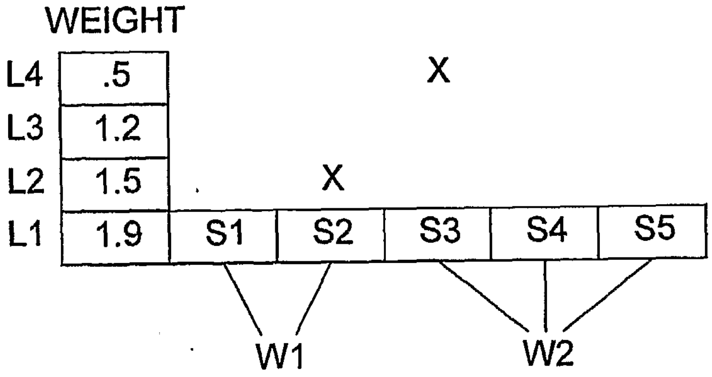
# Fig. 8


102
SEMANTIC
DATABASE
BASE FORM
802
POS/GF 1
804A
806A
SENSE 1
808A
INFO.
SEMCODE
809A
SSC
810A
PCT
812A
MPM
816A
FREQ.
814A
806B
SENSE N
INFO.
808B
POS/GF M
804B
806C
SENSE 1
INFO.
808C
806D
SENSE X
INFO.
808D

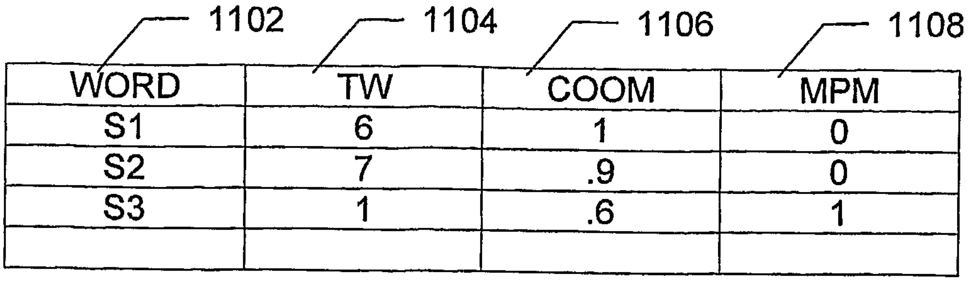
# Fig. 9


SEMCODE
902A
POSITION IN TEXT
903A
POS/GF
904A
INFO.
908A
SSC
910A
PCT
912A
MPM
914A
FREQ.
916A
SEMCODE
902B
POSITION IN TEXT
903B
POS/GF
904A
INFO.
908B
SSC
910B
PCT
912B
MPM
914B
FREQ.
916B

# Fig. 10



# Fig. 11

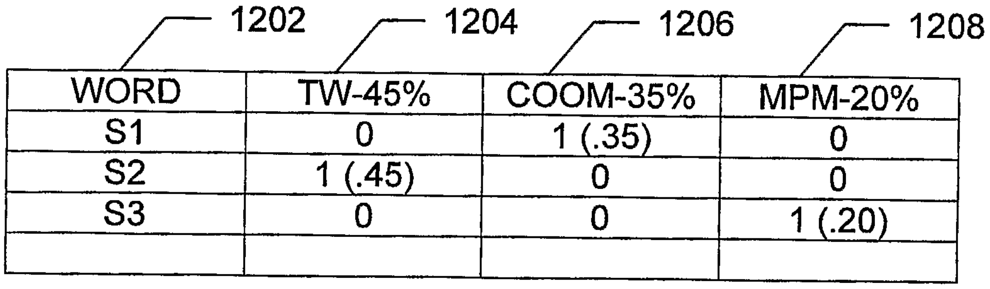


| WORD | TW | COOM | MPM |
|---|---|---|---|
| S1 | 6 | 1 | 0 |
| S2 | 7 | .9 | 0 |
| S3 | 1 | .6 | 1 |
| | | | |

1102
1104
1106
1108

1100

# Fig. 12

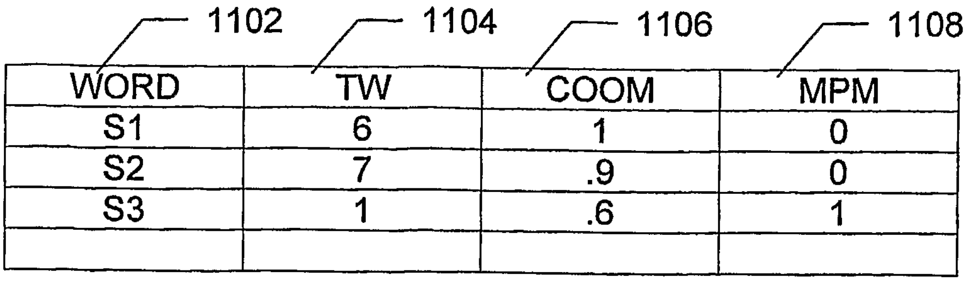


| WORD | TW-45% | COOM-35% | MPM-20% |
|---|---|---|---|
| S1 | 0 | 1 (.35) | 0 |
| S2 | 1 (.45) | 0 | 0 |
| S3 | 0 | 0 | 1 (.20) |
| | | | |

1202
1204
1206
1208

1200

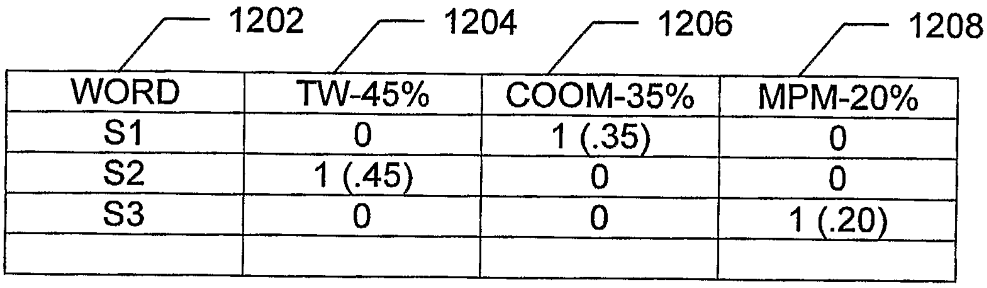
# Fig. 11


1102
1104
1106
1108

| WORD | TW | COOM | MPM |
|---|---|---|---|
| S1 | 6 | 1 | 0 |
| S2 | 7 | .9 | 0 |
| S3 | 1 | .6 | 1 |
| | | | |

1100

# Fig. 12


1202
1204
1206
1208

| WORD | TW-45% | COOM-35% | MPM-20% |
|---|---|---|---|
| S1 | 0 | 1 (.35) | 0 |
| S2 | 1 (.45) | 0 | 0 |
| S3 | 0 | 0 | 1 (.20) |
| | | | |

1200

Fig. 13

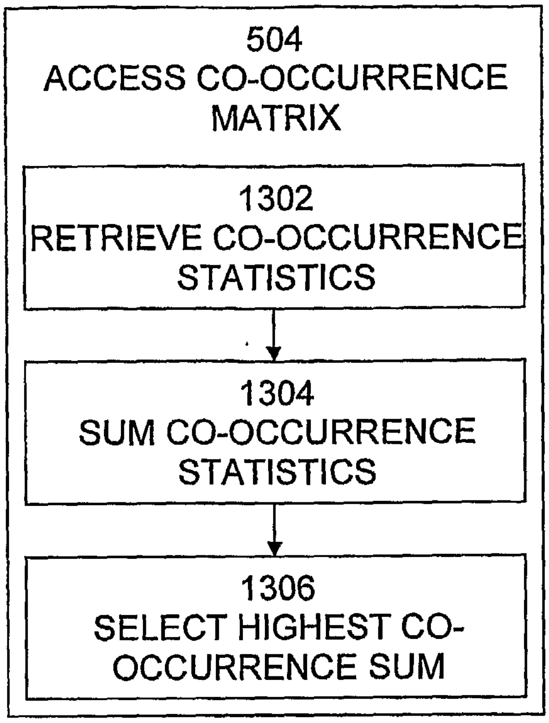
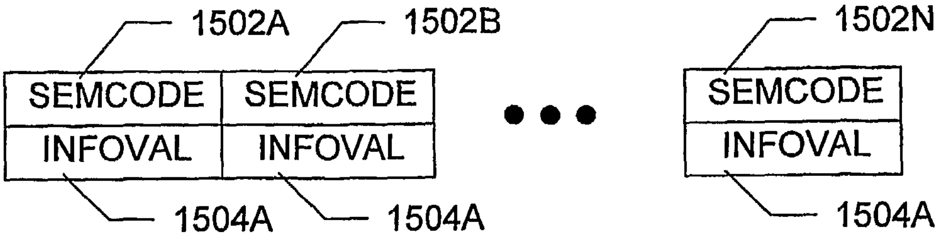

504
ACCESS CO-OCCURRENCE MATRIX
1302
RETRIEVE CO-OCCURRENCE STATISTICS
1304
SUM CO-OCCURRENCE STATISTICS
1306
SELECT HIGHEST CO-OCCURRENCE SUM

Fig. 14

| | $L3_1$ | $L3_2$ | $L3_3$ | $L3_4$ | $L3_5$ |
|---|---|---|---|---|---|
| $L3_1$ | 10 | 3 | 0 | 6 | 2 |
| $L3_2$ | -- | 9 | 4 | 1 | 2 |
| $L3_3$ | -- | -- | 6 | 8 | 0 |
| $L3_4$ | -- | -- | -- | 10 | 2 |
| $L3_5$ | -- | -- | -- | -- | 7 |

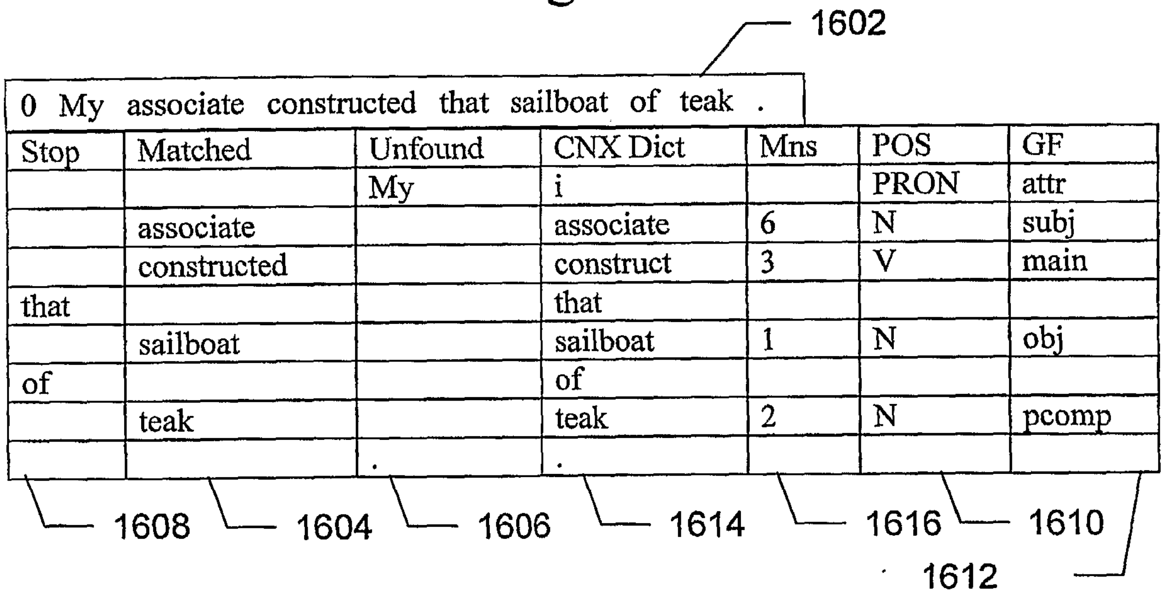
1400

# Fig. 15



1500

# Fig. 16


1602
1608
1604
1606
1614
1616
1610
1612

0 My associate constructed that sailboat of teak .

| Stop | Matched | Unfound | CNX Dict | Mns | POS | GF |
|---|---|---|---|---|---|---|
| | | My | i | | PRON | attr |
| | associate | | associate | 6 | N | subj |
| | constructed | | construct | 3 | V | main |
| that | | | that | | | |
| | sailboat | | sailboat | 1 | N | obj |
| of | | | of | | | |
| | teak | | teak | 2 | N | pcomp |
| | | . | . | | | |

1600

Fig. 17

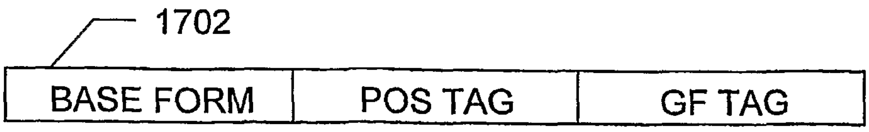


Fig. 18

**Parameters** 1800

POS weights: 1802
Noun: **1.0**
Verb: **0.8**
Adj: **0.5**

Polysemy weights: 1804
from 1 to 3: **1.0**
from 4 to 7: **0.8**
from 8 to 12: **0.6**
from 13 to 18: **0.4**
from 19 and up: **0.2**

SSC weights: 1806
A: **1.1**
B: **1.0**
C: **0.9**
D: **0.65**
E: **0.4**

GF weights: 1808
subj: **1.0**
obj: **1.0**
v-ch: **0.2**
main: **1.0**
subj comp: **1.0**
obj comp: **1.0**
ind obj: **1.0**
prep obj: **0.7**
pcomp-nom: **0.7**
pcomp-verb: **0.9**
pcomp-adverb: **0.5**

Winner weights: 1810
Taxonomy: **1.0**
COOM: **0.0**
MPM: **0.0**
Window size: **100** words
Relevance weight bonus: **15.0%**
Relevance weight for L1 = **1.9**, L2 = **1.5**, L3 = **1.2**, L4 = **1.075**, L5 = **1.02**
Profile levels: **L3 L4**
StrongMeaning threshold: **90.0%**

Fig. 19

**Statistics**

Number of sentences = **1**
Number of words/phrases = **4**
Number of unique words/phrases = **4**

Fig. 20

**Input** *Total: 13*

2002 2004 2006 2008 2010 2012 2014

| N | Word | Type | L6 | L5 | L4 | L3 | L2 | L1 | GF | SSC | Word # | Sent.# |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | associate | Noun | 1 | 2 | 7 | 20 | 3 | 4 | 1 | A | 2 | 0 |
| 2 | associate | Noun | 6 | 3 | 5 | 562 | 25 | 1 | 1 | A | 2 | 0 |
| 3 | associate | Noun | 7 | 5 | 2 | 787 | 1 | 1 | 1 | A | 2 | 0 |
| 4 | associate | Noun | 7 | 5 | 2 | 788 | 11 | 3 | 1 | A | 2 | 0 |
| 5 | associate | Noun | 7 | 5 | 2 | 788 | 12 | 1 | 1 | A | 2 | 0 |
| 6 | associate | Noun | 8 | 2 | 2 | 927 | 23 | 1 | 1 | A | 2 | 0 |
| 7 | construct | Verb | 1 | 3 | 10 | 58 | 1 | 1 | 10 | A | 3 | 0 |
| 8 | construct | Verb | 1 | 10 | 6 | 167 | 3 | 2 | 10 | A | 3 | 0 |
| 9 | construct | Verb | 2 | 3 | 1 | 245 | 1 | 1 | 10 | A | 3 | 0 |
| 10 | sailboat | Noun | 2 | 4 | 6 | 277 | 3 | 1 | 2 | A | 5 | 0 |
| 11 | teak | Noun | 4 | 1 | 3 | 378 | 9 | 1076 | 3 | A | 7 | 0 |
| 12 | teak | Noun | 4 | 3 | 3 | 411 | 50 | 1189 | 3 | A | 7 | 0 |
| 13 | @p | Noun | 0 | 0 | 0 | 0 | 0 | 0 | 0 | C | 9 | 1 |

# Fig. 21

Relevance Weight *Total: 12*

| N | L6 | L5 | L4 | L3 | L2 | L1 | TW | PICW | Word | Type | Word # | KIA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 7 | 20 | 3 | 4 | 1.0 | 0.95 | associate | Noun | 2 | |
| 2 | 1 | 3 | 10 | 58 | 1 | 1 | 1.0 | 0.84 | construct | Verb | 3 | |
| 3 | 1 | 10 | 6 | 167 | 3 | 2 | 1.0 | 0.84 | construct | Verb | 3 | L6+ |
| 4 | 2 | 3 | 1 | 245 | 1 | 1 | 1.0 | 0.84 | construct | Verb | 3 | |
| 5 | 2 | 4 | 6 | 277 | 3 | 1 | 1.0 | 1.05 | sailboat | Noun | 5 | |
| 6 | 4 | 1 | 3 | 378 | 9 | 1076 | 1.0 | 0.735 | teak | Noun | 7 | |
| 7 | 4 | 3 | 3 | 411 | 50 | 1189 | 1.0 | 0.735 | teak | Noun | 7 | L6+ |
| 8 | 6 | 3 | 5 | 562 | 25 | 1 | 1.0 | 0.95 | associate | Noun | 2 | |
| 9 | 7 | 5 | 2 | 787 | 1 | 1 | 1.0 | 0.95 | associate | Noun | 2 | |
| 10 | 7 | 5 | 2 | 788 | 11 | 3 | 1.0 | 0.95 | associate | Noun | 2 | L4+ |
| 11 | 7 | 5 | 2 | 788 | 12 | 1 | 1.0 | 0.95 | associate | Noun | 2 | L3+ |
| 12 | 8 | 2 | 2 | 927 | 23 | 1 | 1.0 | 0.95 | associate | Noun | 2 | |
| | | | | | | Total | | | | | | 0 |

# Fig. 22

2200

Semantic Profile *Total: 21*

2202 2204

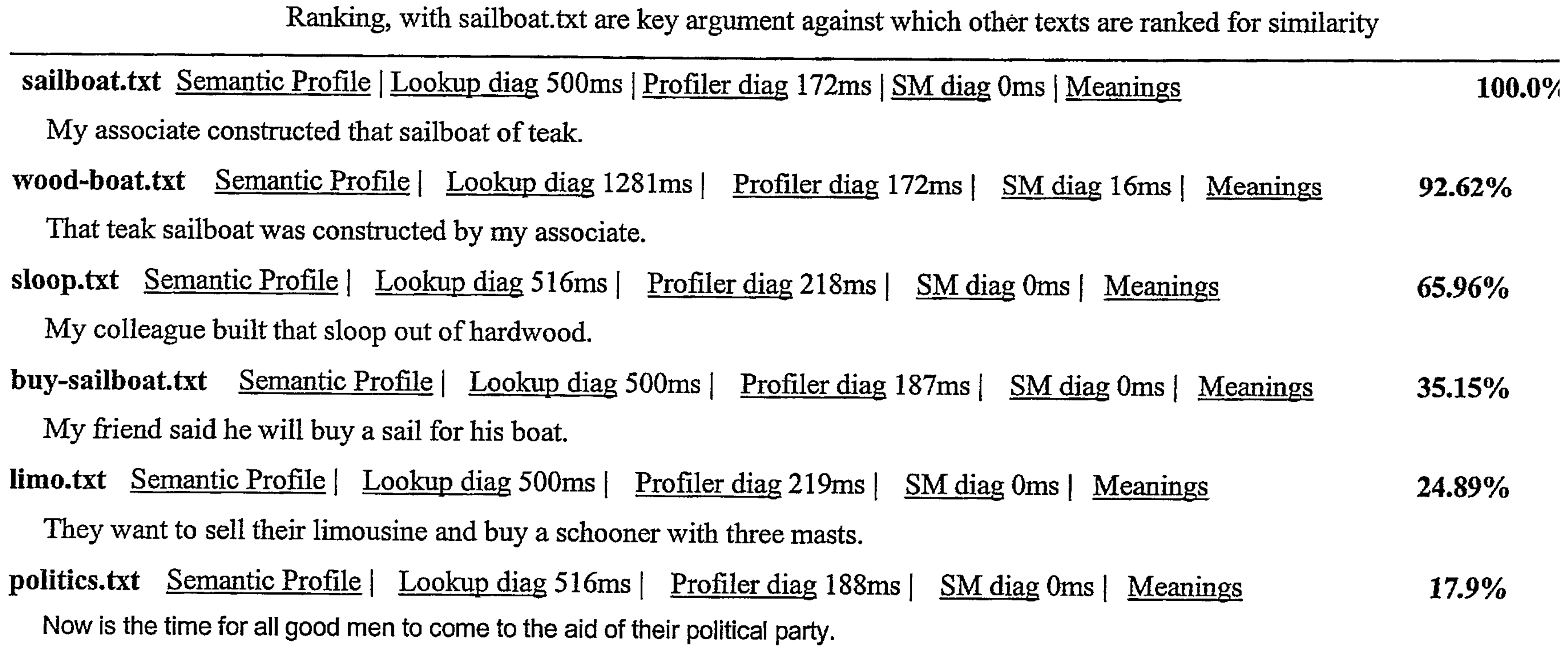
| N | SemCode | InfoValue | Imp. % | Words |
|---|---|---|---|---|
| 1 | 2.4.6 | 0.1188 | 100.0 | sailboat |
| 2 | 1.2.7 | 0.1075 | 90.48 | associate |
| 3 | 6.3.5 | 0.1075 | 90.48 | associate |
| 4 | 7.5.2 | 0.1075 | 90.48 | associate |
| 5 | 8.2.2 | 0.1075 | 90.48 | associate |
| 6 | 2.4.6.277 | 0.1073 | 90.3 | sailboat |
| 7 | 1.2.7.20 | 0.097 | 81.7 | associate |
| 8 | 6.3.5.562 | 0.097 | 81.7 | associate |
| 9 | 7.5.2.787 | 0.097 | 81.7 | associate |
| 10 | 7.5.2.788 | 0.097 | 81.7 | associate |

Fig. 23

**Target Profile** *Total: 18*

| N | SemCode | InfoValue | Imp. % | Words | Match on | Words | InfoValue | Imp. % | Damp | Matched Value |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.4.6.277 | 0.1393 | 100.0 | sloop | # 6 | sailboat | 0.1073 | 90.3 | | 0.1393 |
| 2 | 2.4.6 | 0.1393 | 100.0 | sloop | # 1 | sailboat | 0.1188 | 100.0 | | 0.1393 |
| 3 | 7.5.2.787 | 0.1393 | 100.0 | colleague | # 9 | associate | 0.097 | 81.7 | | 0.1393 |
| 4 | 7.5.2 | 0.1393 | 100.0 | colleague | # 4 | associate | 0.1075 | 90.48 | | 0.1393 |
| 5 | 8.2.2.927 | 0.1393 | 100.0 | colleague | # 11 | associate | 0.097 | 81.7 | | 0.1393 |
| 6 | 8.2.2 | 0.1393 | 100.0 | colleague | # 5 | associate | 0.1075 | 90.48 | | 0.1393 |
| 7 | 1.3.10.58 | 0.1009 | 72.38 | build | # 15 | construct | 0.0858 | 72.24 | | 0.1009 |
| 8 | 1.3.10 | 0.1009 | 72.38 | build | # 12 | construct | 0.095 | 80.0 | | 0.1009 |
| 9 | 1.10.6.167 | 0.1009 | 72.38 | build | # 16 | construct | 0.0858 | 72.24 | | 0.1009 |
| 10 | 1.10.6 | 0.1009 | 72.38 | build | # 13 | construct | 0.095 | 80.0 | | 0.1009 |
| 11 | 2.2.2.197 | 0.1009 | 72.38 | build | | | | | | |
| 12 | 2.2.2 | 0.1009 | 72.38 | build | | | | | | |
| 13 | 1.3.5.38 | 0.0956 | 68.57 | build | | | | | | |
| 14 | 1.3.5 | 0.0956 | 68.57 | build | | | | | | |
| 15 | 2.1.3.184 | 0.0956 | 68.57 | build | | | | | | |
| 16 | 2.1.3 | 0.0956 | 68.57 | build | | | | | | |
| 17 | 4.1.3.378 | 0.0883 | 63.33 | hardwood | # 20 | teak | 0.0751 | 63.21 | | 0.0883 |
| 18 | 4.1.3 | 0.0883 | 63.33 | hardwood | # 18 | teak | 0.0831 | 70.0 | | 0.0883 |
| | **Total** | **2.0** | | | **12** | | | | | **1.4161** |

Fig. 24



Ranking, with sailboat.txt are key argument against which other texts are ranked for similarity

**sailboat.txt** Semantic Profile | Lookup diag 500ms | Profiler diag 172ms | SM diag 0ms | Meanings **100.0%**

My associate constructed that sailboat of teak.

**wood-boat.txt** Semantic Profile | Lookup diag 1281ms | Profiler diag 172ms | SM diag 16ms | Meanings **92.62%**

That teak sailboat was constructed by my associate.

**sloop.txt** Semantic Profile | Lookup diag 516ms | Profiler diag 218ms | SM diag 0ms | Meanings **65.96%**

My colleague built that sloop out of hardwood.

**buy-sailboat.txt** Semantic Profile | Lookup diag 500ms | Profiler diag 187ms | SM diag 0ms | Meanings **35.15%**

My friend said he will buy a sail for his boat.

**limo.txt** Semantic Profile | Lookup diag 500ms | Profiler diag 219ms | SM diag 0ms | Meanings **24.89%**

They want to sell their limousine and buy a schooner with three masts.

**politics.txt** Semantic Profile | Lookup diag 516ms | Profiler diag 188ms | SM diag 0ms | Meanings **17.9%**

Now is the time for all good men to come to the aid of their political party.

Fig. 25

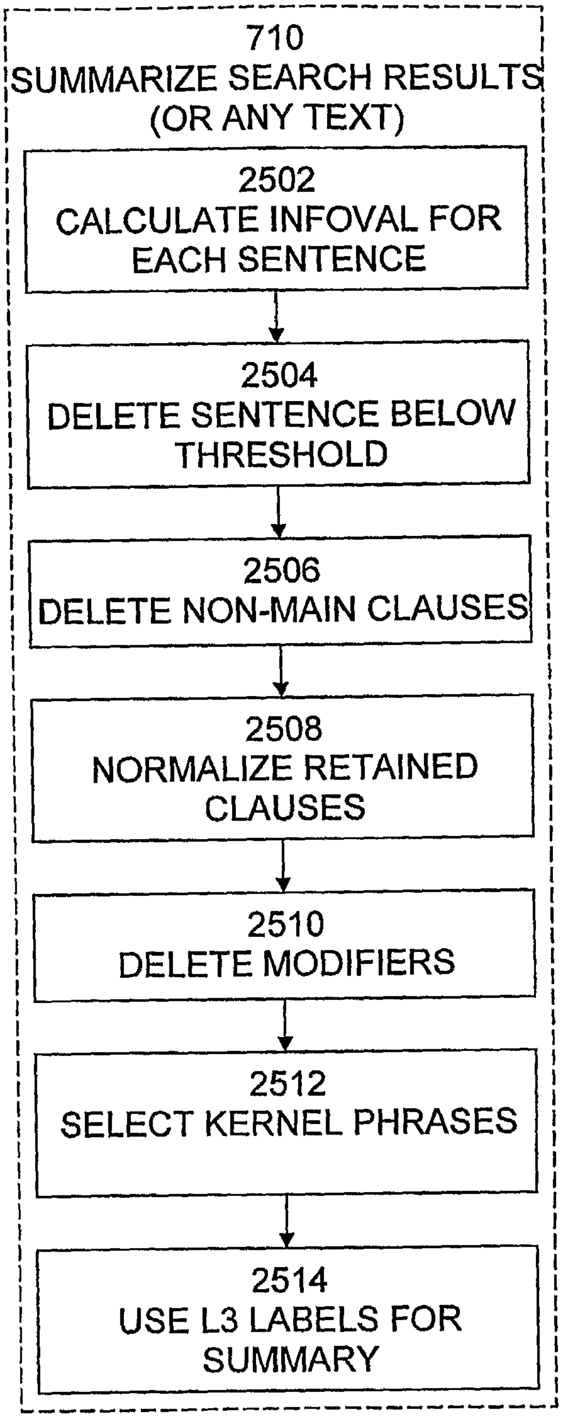
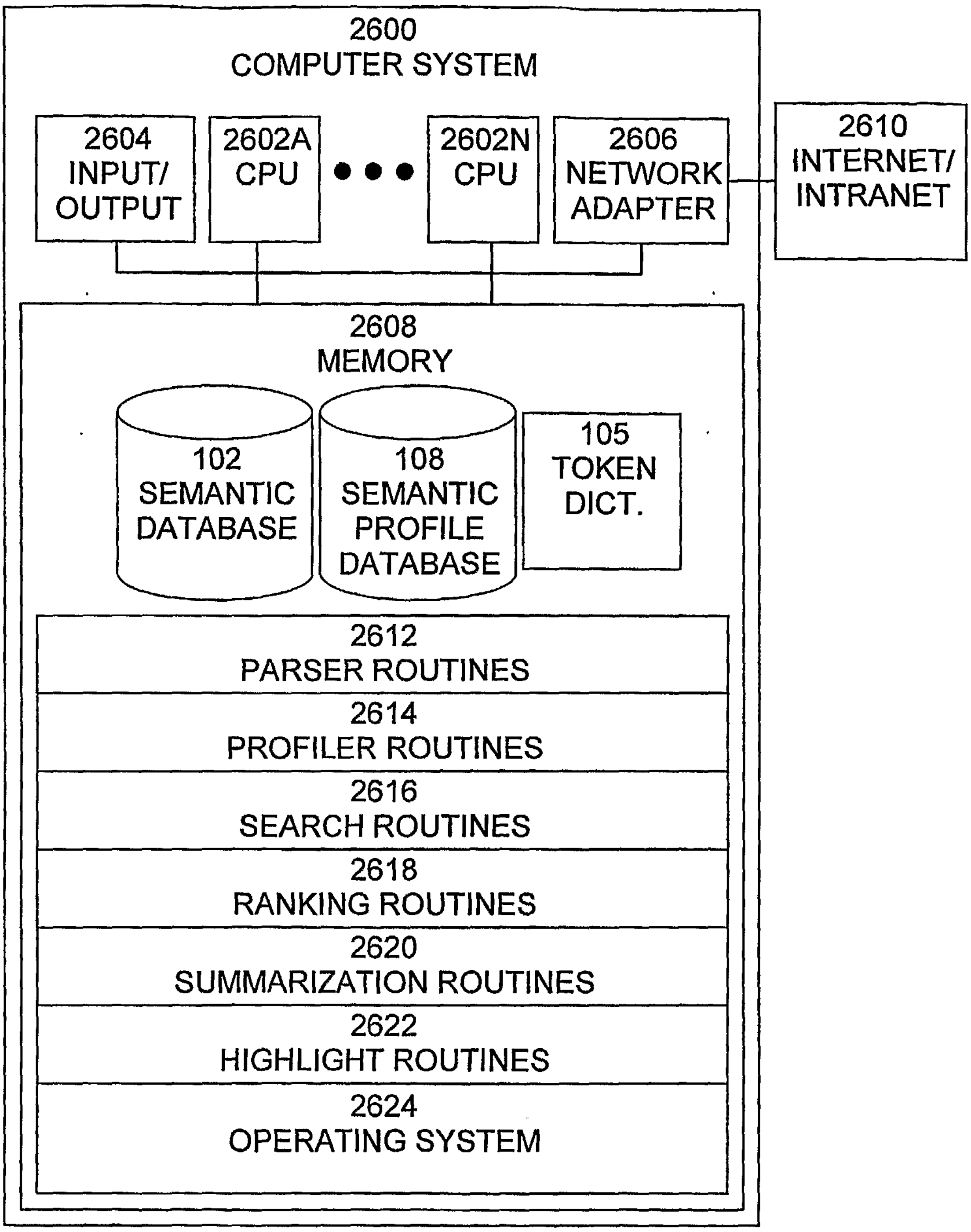

710
SUMMARIZE SEARCH RESULTS
(OR ANY TEXT)
2502
CALCULATE INFOVAL FOR EACH SENTENCE
2504
DELETE SENTENCE BELOW THRESHOLD
2506
DELETE NON-MAIN CLAUSES
2508
NORMALIZE RETAINED CLAUSES
2510
DELETE MODIFIERS
2512
SELECT KERNEL PHRASES
2514
USE L3 LABELS FOR SUMMARY

Fig. 26


2600
COMPUTER SYSTEM
2604
INPUT/
OUTPUT
2602A
CPU
2602N
CPU
2606
NETWORK
ADAPTER
2610
INTERNET/
INTRANET
2608
MEMORY
102
SEMANTIC
DATABASE
108
SEMANTIC
PROFILE
DATABASE
105
TOKEN
DICT.
2612
PARSER ROUTINES
2614
PROFILER ROUTINES
2616
SEARCH ROUTINES
2618
RANKING ROUTINES
2620
SUMMARIZATION ROUTINES
2622
HIGHLIGHT ROUTINES
2624
OPERATING SYSTEM