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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 February 2016 (18.02.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/024261 A1

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

(21) International Application Number:
PCT/IL2015/000037

(22) International Filing Date:
10 August 2015 (10.08.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/037,466 14 August 2014 (14.08.2014) US

(71) Applicant: **OPISOFTCARE LTD.** [IL/IL]; 34 Habarzel St., 6971050 Tel-Aviv (IL).

(72) Inventors: **ALTER, Alon**; 6 Sphnoza St., 4358808 Raanana (IL). **TOZHOVEZ, Oksana**; 5/6 Bilu St., 4946207 Petah-Tiqwa (IL).

(74) Agent: **ROSENTHAL, Ytzhak**; 21 Hapninim St., 6019000 Neve Efrayim (IL).

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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*

(54) **Title**: METHOD AND SYSTEM FOR SEARCHING PHRASE CONCEPTS IN DOCUMENTS

(57) **Abstract**: A system and method for fast concept search in multiple documents where the concept is expressed by plurality of words, all of which have to be in the same sentence and within specified range. The system automatically finds equivalent expressions of the same concept, and returns as search results all documents in which the concept is contained.

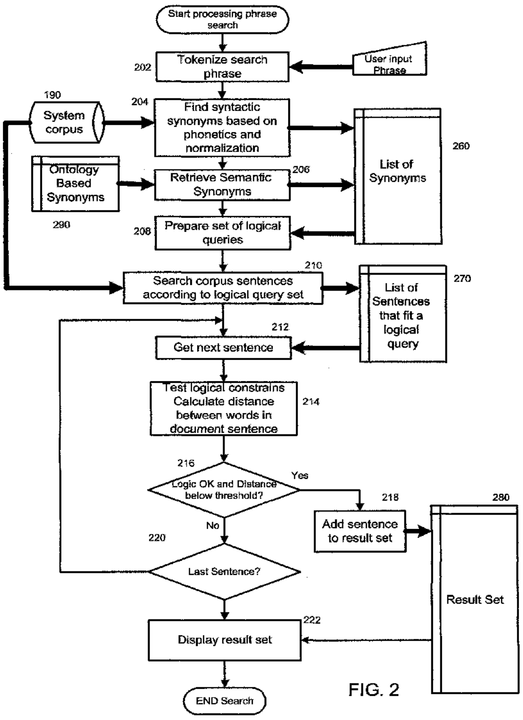
Start processing phrase search
202
Tokenize search phrase
User input Phrase
190
System corpus
204
Find syntactic synonyms based on phonetics and normalization
260
List of Synonyms
Ontology Based Synonyms
290
206
Retrieve Semantic Synonyms
208
Prepare set of logical queries
210
Search corpus sentences according to logical query set
270
List of Sentences that fit a logical query
212
Get next sentence
Test logical constrains Calculate distance between words in document sentence
214
216
Logic OK and Distance below threshold?
Yes
No
218
Add sentence to result set
280
Result Set
220
Last Sentence?
222
Display result set
END Search

FIG. 2

METHOD AND SYSTEM FOR SEARCHING PHRASE CONCEPTS IN DOCUMENTS

FIELD OF THE INVENTION

The invention generally relates to the field of information retrieval and more particularly to retrieving answers to the concepts expressed in the search queries

BACKGROUND OF THE INVENTION

In recent years there has been a massive movement towards computerizing medical data for various health service e organizations. However, making doctors write down their examination documents and their diagnostics using specific codes and sentences to write down the prognosis of each patient, will inevitably lower their productivity. Thus, most modern systems designed for computerizing medical data today go the path of natural language processing (NLP), allowing the doctors to write down their prognosis the way they are used to, and using computer analysis to extract vital information such as information about a patient, about illnesses, treatments etc. through the use of natural language processing (NLP).

Naturally, this process presents many problems. One of them is the need to analyze and normalize sentences – for example "there is no sign of a hernia"; This prognosis can be written in many forms in natural language – for example "hernia has been ruled out", or "no apparent sign of a hernia" and so on. These variations appear in different documents, and they all express the same concept.

Most algorithms, such as the ones described in the public Stanford NLP pages and in many patents, refer to web searches. In these cases users fail to choose effective query terms. Often documents that satisfy user's information need may use different words than the query terms. We are interested in professional information retrieval system aimed to be used by professional community, such as health data retrieval system. In this case the query is expressed with the exact terms, but the meaning of the query depends on the whole phrase. In many cases the query defines allowed distances between words, but they do not require that that words in the phrase are in the same sentence. Thus wrong results can be retrieved.

SUMMARY OF THE INVENTION

The disclosed invention assumes that an meaningful information that is searched by a user is expressed in a sentence, thus when a set of keywords are searched for, they are all expected to be in the same sentence. Usually, search engines define maximum distance between the words in the query regardless of the sentence limits. Hence, in the first phase of the processing, each new document that is added to the corpus is analyzed and broken into sentences so that for every word information as to its position in the document and to the sentence in which it appears is kept. In addition to the indexing information, normalized version and phonetic representation of the word are saved.

From the phrase query entered by the user, many search phrases are derived. These search phrases are generated by finding dictionary synonyms to all query words, and retrieving semantic synonyms from an ontology. Phonetic representation is prepared for each word in the derived search phrases. From this data a comprehensive set of logical queries is prepared.

It is an object of the disclosed invention to retrieve maximal set of relevant documents that relates to the query phrase

It is another objective of the disclosed invention is to enable a professional user, who is not familiar with complex query structure, to retrieve information he is interested in.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a top level flow chart of the search preparation process.

FIG. 2 presents a flow chart of the processing of the query.

DETAILED DESCRIPTION

The invention will be described more fully hereinafter, with reference to the accompanying drawings, in which a preferred embodiment of the invention is shown. The embodiment refers to a corpus containing medical documents.

The invention may, however, be embodied in many different forms and should not be construed as limited to the embodiment set forth herein; rather this embodiment is provided so that the disclosure will be thorough and complete, and will fully convey the scope of the invention to those skilled in the art.

Before describing the processing that each document goes through, it is important to explain the corpus of the system. The corpus of the system is a database that stores information on each document ever entered the system, documents that constitute the search domain. Among the information on each word the system corpus keeps a list of all words and their locations within the sentences as well as the sentence number within the document where that word is located, referred to as the search indexes. It also contains a phonetic representation for each word as well as statistical information on the word.

The top level flow chart of the preparation process is shown in Fig. 1. Each new medical document, which is part of the search domain, goes through the preparation process. The system reads a new document in step 100. The document is split into sentences – step 102. The sentences are temporarily saved in a List of Sentences 170. Each sentence is processed by steps 104 to 122.

A sentence is retrieved from the List of Sentences 170 and an index to the sentence is added – step 104. The retrieved sentence is tokenized in step 106 and a temporary

List of Words and symbols in the sentence 180 is prepared. Each word in the sentence is processed in steps 108 to 120 as described hereafter.

A new word is retrieved – step 108 from the List of words in a sentence 180. In step 110 the system corpus 190 is searched to find out if the word is already known. If the word is new, as checked in step 112, than the new word is processed in step 114, where it is normalized and goes through phonetic conversion and is added to the corpus – 190 and the processing proceeds with step 116. If the retrieved word is not new, processing continues in step 116. In step 116 the statistics related to the word is updated, and the search index of the word in the sentence is updated in step 118.

If the retrieved word was not the last word in the sentence, as tested in step 120, then the processing returns to step 108, where processing of a new word begins. If the retrieved word was the last word in the sentence, then step 122 is executed. If the retrieved sentence was not the last one, then the processing of new sentence is executed, starting with step 104. Otherwise the processing of the new document terminates.

Fig. 2 describes the processing of the phrase query. For the purpose of explanation we assume that there are 3 documents in the corpus that contain the following sentences respectively, "there is no sign of Carcinoma", "Carzinoma has been ruled out", and "no apparent sign of cancer". These three sentences clearly express the same idea. The user wants to find out the cases where cancer was suspected but was not found. The professional user enters the query "no carcinoma". These words all have to be in the same sentence, but they do not have to be consecutive. The expression "ruled out" is synonym for "no" and it may appear after the subject "carcinoma" in the sentence and it gives the sentence the same meaning. Skin cancer, carcinoma, SCC are all semantic synonyms, and carcinoma is frequently misspelled as carzinoma, carsinome etc. The process as described hereafter can find all wording combinations that have the same meaning.

The incoming search query is tokenized in step 202. For each word in the query, syntactic synonyms based on phonetic similarity and normalization are generated in step 204 and are temporarily saved in a List of Synonyms 260. The synonyms are looked for in the corpus 190. Referring to the above give example, in this step the words carcinoma, carzinoma, are found because they are similar from phonetic point of view. This similarity is determined by the distance between these words measured by Jaro-Winkler algorithm.

Semantic synonyms for each word in the query are derived in step 206 from an ontology 290, and are added to the List of Synonyms 260. Again, referring to the above given example, in this step the words cancer, SCC are semantic synonyms for carcinoma, and the words ruled-out, without, not and negative are semantic synonyms for "no".

Using the stored list of synonyms 260, in step 208 a set of logical queries is prepared. The query set is comprised of all combinations of search phrases that express the same concept of the query. A search query within the set can include, in addition to the words, also logical constrains such as distance between the words in a sentence, or define that a specific word has to precede another one etc. For example, the query can include multiple phrases with logical operators that determine the relationship between them, e.g. hypertension OR [edema extremities]. Note that every query in the set includes the constraint that the words have to be in the same sentence. In step 210, the set of queries are applied to the documents in the system corpus 190, and a list of all sentences that contain the required words is prepared and these sentences are temporarily saved in a list 270.

A candidate search result sentence saved in the list 270 is popped from the list 270 in step 212. The logical constraints and the distance between words are evaluated in step 216. The maximum distance is checked against predefined threshold. If the logical constraints are met and the distance between the words in the sentence is below the query defined threshold, then the sentence with its relevant data, such as its document number is added – step 218 to the result set 280. When all searched sentences in the list 270 have been processed, the test in step 214 indicates that there is no new sentence, and the search results are displayed to the user – step 222.

what has been described above is just one embodiment of the disclosed innovation. It is of course, not possible to describe every conceivable combination of components and/or methodologies, but one of ordinary skill in the art may recognize that many further combinations and permutations are possible. Accordingly, the innovation is intended to embrace all such alterations, modifications, and variations that fall within the spirit and scope of the appended claims.

Furthermore, to the extent that the term "includes" is used in either the detailed description or the claims, such term is intended to be inclusive in a manner similar to the term "comprising" as "comprising" is interpreted when employed as a transitional word in a claim.

CLAIMS

What is claimed is:

1. A method for performing search to retrieve phrase concepts from documents stored in a corpus, the method is comprised of the following steps:
    a. Splitting all documents in the search domain into sentences; splitting each sentence to its words; keeping for each word its phonetic representation and its indexes;
    b. receiving a query regarding the search subject from the user;
    c. finding syntactic and semantic synonyms to all words of the query;
    d. preparing set of logical queries for all synonym combinations;
    e. retrieving all sentences that respond to at least one query;
    f. calculating a score for each retrieved sentence, and
    g. displaying documents that contain sentences having a score higher than a pre-defined threshold.

2. The method according to claim 1 wherein semantic synonyms are derived from an ontology database;

3. The method according to claim 1 wherein the syntactic synonyms are derived from the words stored in the corpus, by finding similar phonetic representation between a word in the query and a word in the corpus and measuring the distance between these two words.

4. The method according to claim 3 wherein Jaro-Winkler algorithm is used to compute the distance between words having similar phonetic representation.

5. The method according to claim 1 wherein the user can update the ontology.

6. The method according to claim 3 wherein the Jaro-Winkler algorithm is adapted to the Hebrew Language.

7. The method according to claim 1 wherein sentence splitting is based on syntactic analysis and noun-phrase analysis.

8. A system comprising one or more computers configured to perform operations for retrieving findings from documents stored in a corpus, operations comprising:
    a. Splitting all documents in the search domain into sentences; splitting each sentence to its words; keeping for each word its phonetic representation and its indexes;
    b. receiving a query regarding the search subject from the user;
    c. finding syntactic and semantic synonyms to all words of the query;
    d. preparing set of logical queries for all synonym combinations;
    e. retrieving all sentences that respond to at least one query;
    f. calculating a score for each retrieved sentence, and
    g. displaying documents that contain sentences having a score higher than a pre-defined threshold.

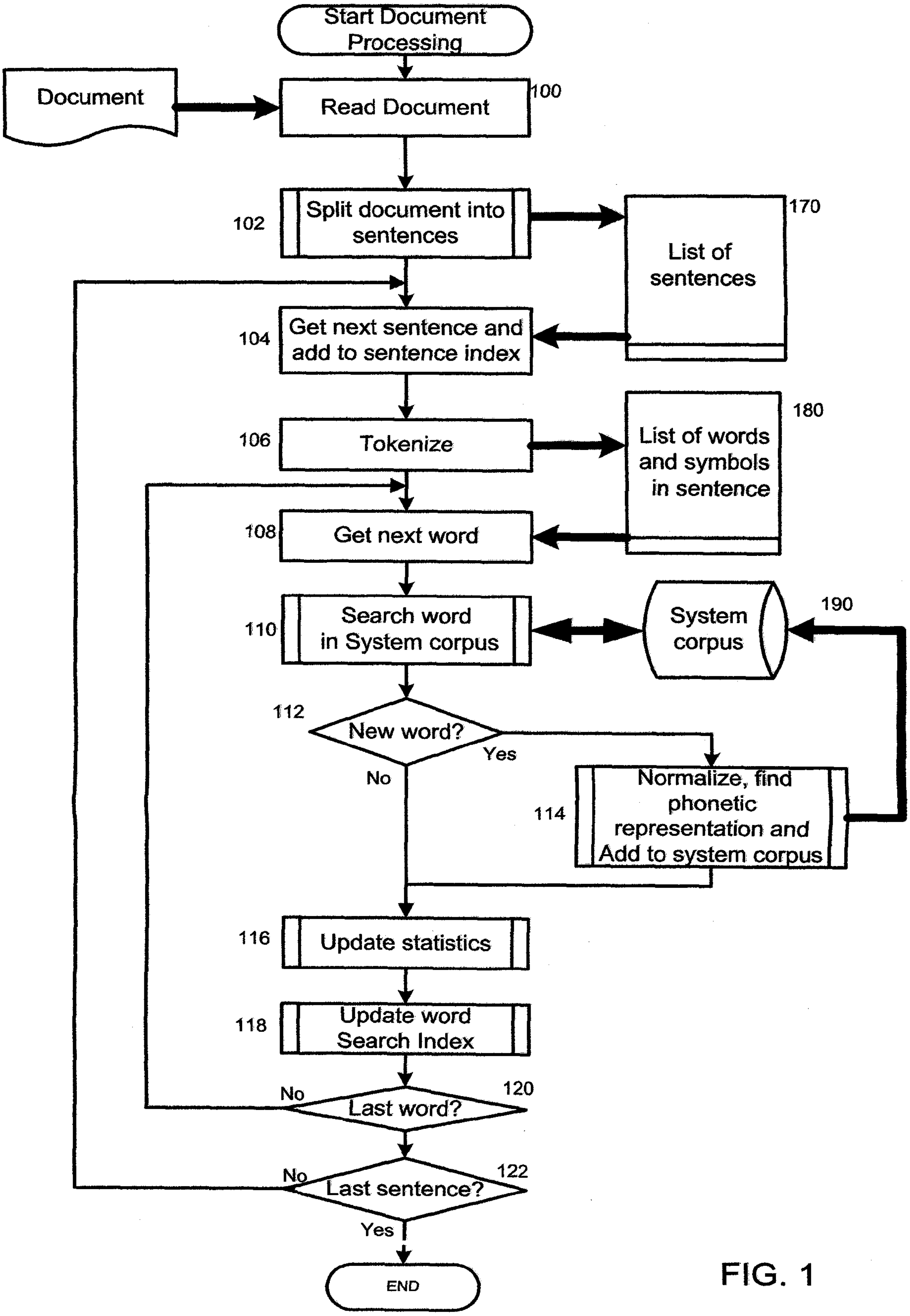
Start Document Processing
Document
Read Document
100
102
Split document into sentences
170
List of sentences
104
Get next sentence and add to sentence index
106
Tokenize
180
List of words and symbols in sentence
108
Get next word
110
Search word in System corpus
System corpus
190
112
New word?
Yes
No
114
Normalize, find phonetic representation and Add to system corpus
116
Update statistics
118
Update word Search Index
No
Last word?
120
No
Last sentence?
122
Yes
END

FIG. 1

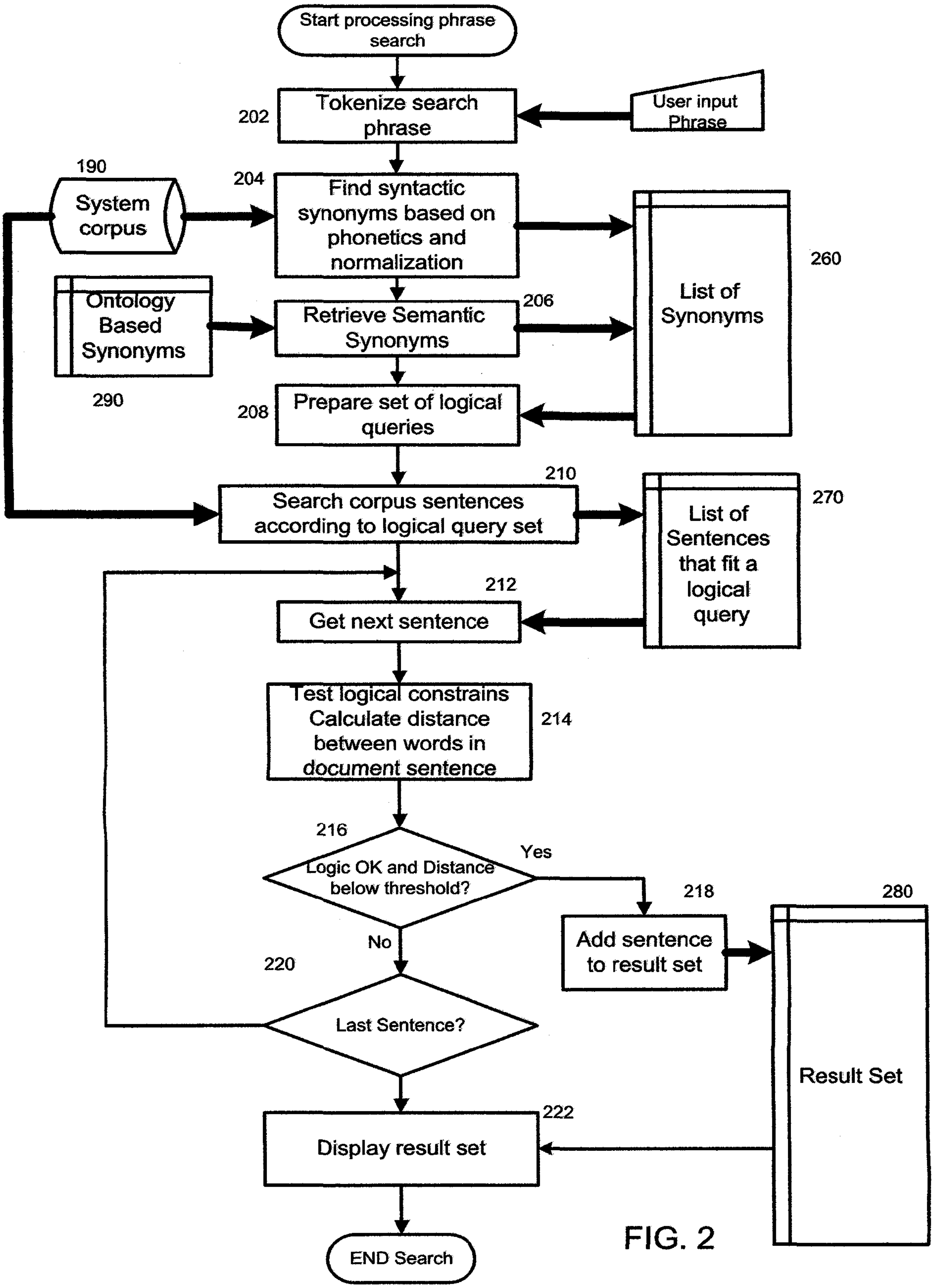
Start processing phrase search
202
Tokenize search phrase
User input Phrase
190
System corpus
204
Find syntactic synonyms based on phonetics and normalization
260
List of Synonyms
Ontology Based Synonyms
206
Retrieve Semantic Synonyms
290
208
Prepare set of logical queries
210
Search corpus sentences according to logical query set
270
List of Sentences that fit a logical query
212
Get next sentence
Test logical constrains Calculate distance between words in document sentence
214
216
Logic OK and Distance below threshold?
Yes
218
280
Add sentence to result set
No
220
Last Sentence?
Result Set
222
Display result set
END Search

FIG. 2

INTERNATIONAL SEARCH REPORT

International application No.

PCT/IL2015/000037

A. CLASSIFICATION OF SUBJECT MATTER

IPC (2015.01) G06F 17/30

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC (2015.01) G06F 17/30

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

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

Databases consulted: Google Scholar, PatBase

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 20090204605 A1 13 Aug 2009 (2009/08/13) par. 0014, 0015, 0016	1-8
Y	US 20090076839 A1 19 Mar 2009 (2009/03/19) par. 0019, 0021, 0024	1-8
A	US 7636714 B1 22 Dec 2009 (2009/12/22) abstract, the whole doc.	1-8
A	US 20060122997 A1 08 Jun 2006 (2006/06/08) abstract, the whole doc.	1-8
A	US 20060206476 A1 14 Sep 2006 (2006/09/14) abstract, the whole doc.	1-8

□ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search 25 Nov 2015	Date of mailing of the international search report 01 Dec 2015
Name and mailing address of the ISA: Israel Patent Office Technology Park, Bldg.5, Malcha, Jerusalem, 9695101, Israel Facsimile No. 972-2-5651616	Authorized officer RUSS Eran Telephone No. 972-2-5651701

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/IL2015/000037

Patent document cited search report	Publication date	Patent family member(s)	Publication Date
US 20090204605 A1	13 Aug 2009	US 2009204605 A1	13 Aug 2009
		US 8392436 B2	05 Mar 2013
US 20090076839 A1	19 Mar 2009	US 2009076839 A1	19 Mar 2009
US 7636714 B1	22 Dec 2009	US 7636714 B1	22 Dec 2009
US 20060122997 A1	08 Jun 2006	US 2006122997 A1	08 Jun 2006
		CN 1783089 A	07 Jun 2006
		TW I336850 B	01 Feb 2011
US 20060206476 A1	14 Sep 2006	US 2006206476 A1	14 Sep 2006
		US 7574436 B2	11 Aug 2009
		CN 101137985 A	05 Mar 2008
		CN 101882149 A	10 Nov 2010
		EP 1856641 A1	21 Nov 2007
		JP 2008533596 A	21 Aug 2008
		JP 5114380 B2	09 Jan 2013
		KR 20070110868 A	20 Nov 2007
		KR 101157349 B1	03 Jul 2012
		KR 20120065423 A	20 Jun 2012
		WO 2006099331 A1	21 Sep 2006