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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
3 January 2014 (03.01.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/003543 A1

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

(21) International Application Number:
PCT/NL2012/050463

(22) International Filing Date:
29 June 2012 (29.06.2012)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **SOPHEON N.V.** [NL/NL]; Kantoorgebouw OFFICIA 1, De Boelelaan 7, NL-1083 HJ Amsterdam (NL).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **RUTTEN, Hubert Joseph Marie** [NL/NL]; 4, Spauwenstraat, NL-6245 BW Eijsden (NL). **KLEYNENBERG, Steven Ernst** [NL/NL]; 1, Waterloop, NL-DELFT 2614 Xc (NL). **TEN PAS, Elsemiek** [NL/NL]; 242, Rijndijk, NL-2331 AM Leiden (NL).

(74) Agent: **Mertens, H.V.**; P.O. Box 3241, NL-2280 GE Rijswijk (NL).

(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, 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, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, 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, 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, ML, MR, NE, SN, TD, TG).

Published:

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

WO 2014/003543 A1

(54) **Title**: METHOD, SYSTEM AND COMPUTER PROGRAM FOR GENERATING A QUERY REPRESENTATION OF A DOCUMENT, AND QUERYING A DOCUMENT RETRIEVAL SYSTEM USING SAID QUERY REPRESENTATION

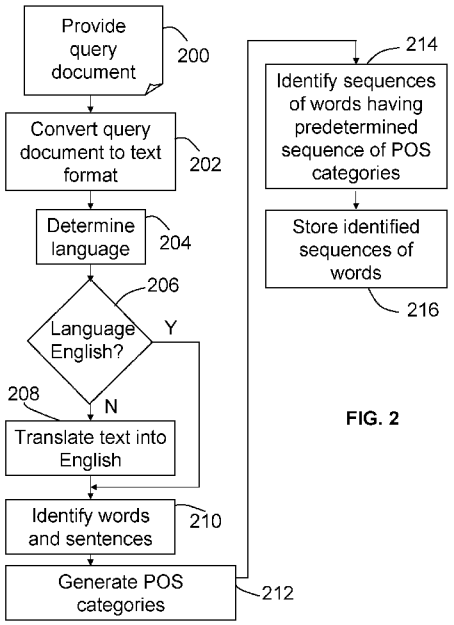


FIG. 2

(57) **Abstract**: In a method and system of generating a query representation of an electronic query document, the query document is processed by a computer processor. The computer processor is configured to identify words and sentences in the query document, generate for each word a corresponding part-of-speech, POS, category of the word, identify each sequence of words having a predetermined sequence of POS categories, and store the identified sequences of words as the query representation of the query document. In a method and system for querying a document retrieval system, the document retrieval system is queried with a plurality of the stored identified sequences of words; and target documents are retrieved from the document retrieval system. The target documents have at least one sequence of words in common with the query document. In a method and system for clustering similar documents in a set of electronic documents, one document of the set of documents is designated as a query document. The query document is processed to store identified sequences of words as a query representation of the query document. Each remaining one of the set of documents is queried with a plurality of the stored identified sequences of words. A similarity value for each query of a remaining one of the set of documents is determined, and documents in the set of documents are clustered based on the similarity values.

Method, system and computer program for generating a query representation of a document, and querying a document retrieval system using said query representation

FIELD OF THE INVENTION

The invention relates to the field of document processing, and more specifically to document processing allowing to query a document retrieval system. More specifically, the invention relates a computer-implemented method, system and computer program for generating a query representation of an electronic query document. The invention also relates to a computer-implemented method, system and computer program for querying a document retrieval system, using the query representation of the electronic query document to retrieve relevant documents or information related to such documents. The invention further relates to a computer-implemented method, system and computer program for clustering similar documents in a set of electronic documents.

Herein, a document retrieval system is any system allowing to query an index of documents with one or more word sequences (including exact phrases) and returning results ranked by relevance. The word sequences may be linked by Boolean operators, and results may be provided with a relevance score. Examples of documents retrieval systems are systems based on Lucene, Oracle Database or SQL server.

BACKGROUND OF THE INVENTION

Ever since people and organizations have had to deal with massive quantities of electronic documents, there are attempts to use smart, cognitive methods to identify and select documents of relevance out of a large number, e.g. millions, of candidate documents. In such methods for finding relevant documents, a specific document, referred to as a reference document or a query document, may be used as a basis for setting up a query for similar, comparable documents. The query document is processed to generate some form of representation of its content, sometimes called a profile or fingerprint.

Several approaches exist to generate a query from a query document. As documents in general mostly contain text, and as texts are made of words, all approaches include some form of text analysis to identify meaningful units of text deemed relevant for building the query.

Some approaches are limited to identifying key phrases (a phrase being a plurality of words), whereas others mention phrases as just one of the types of units to consider, next to single words and other entities.

For phrase identification for the purpose of information retrieval, one or more of the following techniques may be used. Firstly, an 'n-gram technique' may be applied, which does not use any linguistic features, but splits a text in sequences of two or more words (also referred to as character strings) which are evaluated using frequency measures or rules for a beginning and end of a phrase. Secondly, a 'lexicon-based method' may be applied, using a predetermined list of phrases to select matching character strings, or a stopword list to filter out non-relevant words. Thirdly, a 'linguistic approach' may be applied, using linguistic features and patterns to recognize phrases. Whichever of these techniques is used, it is always combined with, or followed by statistical measures.

In discussions on phrase queries, it is often pointed our that there is a risk that phrases are too precise to yield a good recall. Hence, various approaches provide a means to abstract away from the exact wording in a text. Accordingly, fourthly, a 'conceptual approach' may be applied, wherein words and phrases are mapped onto concepts (such as a concept hierarchy) such that similar documents can be found even if a different wording is used therein for the same concepts. An approach measuring conceptual similarity is the Latent Semantic Indexing, LSI, paradigm. LSI uses a mathematical analysis to identify relations below a surface level of language.

US Patent No. 6,026,388 discloses techniques for generating sophisticated representations of the contents of both queries and documents in a retrieval system by using natural language processing (NLP) techniques to represent, index, and retrieve texts at the multiple levels at which humans construe meaning in writing (e.g., the morphological, lexical, syntactic, semantic, discourse, and pragmatic levels). The user enters a query and the system processes the query to generate an alternative representation, which includes conceptual-level abstraction and representations based on complex nominals (CNs), proper nouns (PNs), single terms, text structure, and logical make-up of the query, including mandatory terms. After processing the query, the system displays query information to the user, indicating the system's interpretation and representation of the content of the query. The user is then given an opportunity to provide input, in response to which the system modifies the alternative representation of the query. Once the user has provided desired input, the possibly modified representation of the query is matched to the relevant document retrieval system, and measures of relevance generated for the documents. A set of documents is presented to the user, who is given an opportunity to select some or all of the documents, typically on the basis of such documents being of particular relevance. The user then initiates

the generation of a query representation based on the alternative representations of the selected document(s).

The techniques presented in reference US Patent No. 6,026,388 provide a sophisticated, and thereby complex body of processing of queries and documents for query purposes. For example, the system according to the reference includes a Subject Field Coder which tags content-bearing words in a document text with a disambiguated subject code using an online lexical resource of words, as a step to prepare a query to be performed on one or more databases containing documents which have been pre-processed to provide a representation thereof to be used in the query.

In general, the prior art provides integrated systems to, on the one hand, build a query and, on the other hand, provide query results whereby the query document and the retrieved documents must be subject to the same indexing process. Such systems do not allow to use whichever document as query document, independent from the retrieval system.

SUMMARY OF THE INVENTION

It would be desirable to provide a simple, powerful, improved method, system and computer program to characterize a query document. It would also be desirable to provide a simple, powerful, improved method, system and computer program to query any existing document retrieval system, to retrieve one or more target documents or other information, such as bibliographic information, therefrom, without any specific previous processing of the (documents in the) document retrieval system to optimally respond to the query. It would further be desirable to provide a simple, powerful, improved method, system and computer program to cluster (i.e. to group together) documents retrieved from any existing document retrieval system, in parallel or sequentially, based on their degree of similarity.

To better address one or more of these concerns, in a first aspect of the invention a computer-implemented method of generating a query representation of an electronic query document is provided. The method comprises: processing the query document by a computer processor configured to: identify words and sentences in the query document; generate for each word a corresponding part-of-speech, POS, category of the word; identify each sequence of words having a predetermined sequence of POS categories; and store the identified sequences of words as the query representation of the query document.

The method serves to build a query from a document, here referred to as a query document, and will produce a set of word sequences. This set of word sequences may serve to understand the document, and further may advantageously be used as a list of search terms to query a document retrieval system. The query document is separated from any target document of the document retrieval system, both conceptually and physically.

In a second aspect of the invention, a computer-implemented method for querying a document retrieval system is provided. The method comprises: providing an electronic query document; processing the query document by a computer processor configured to: identify words and sentences in the query document; generate for each word a corresponding part-of-speech, POS, category of the word; identify each sequence of words having a predetermined sequence of POS categories; and store the identified sequences of words as a query representation of the query document, querying the document retrieval system with a plurality of the stored identified sequences of words; and retrieving from the document retrieval system target documents having at least one sequence of words in common with the query document.

The method of the invention yields exceptionally good query results by selecting the predetermined sequences of POS categories, in other words: by selecting the predetermined sequences of POS categories, to select word sequences from a query document. A high quality of search is obtained by the cumulation of sequences of POS categories which have been found to be highly representative of the content of the documents searched for. Contrary to the common notion that both the query documents and the target documents need to be indexed using the same sophisticated and complex techniques, the invention focuses on the user (searcher) side to build, in fact generate a query from a document which may be almost universally applied to existing, industry standard document retrieval systems without the need for being part of the index of the document retrieval system concerned. The method of the invention is scalable in the sense that an unlimited number of document retrieval systems may be addressed, even in parallel and synchronously. The method of the invention is independent from proprietary, linguistic or semantic indexes and databases. The target documents contained in the document retrieval systems may be indexed in a standard way for the document retrieval system concerned, without a need for linguistic or semantic manipulations on the indexes.

In a third aspect of the invention, a computer-implemented method for clustering similar documents in a set of electronic documents is provided. The method comprises: (a) designating one document of the set of documents as a query document; (b) processing the query document by a computer processor configured to: identify words and sentences in the query document; generate for each word a corresponding part-of-speech, POS, category of the word; identify each sequence of words having a predetermined sequence of POS categories; and store the identified sequences of words as a query representation of the query document; (c) querying each remaining one of the set of documents with a plurality of the stored identified sequences of words; (d) providing a similarity value for each query of a remaining one of the set of documents; (e) repeating steps (a) – (d) for each remaining one of the set of documents by designating each remaining one of the set of documents as a query

document; and (f) clustering documents in the set of documents having a similarity value exceeding a threshold.

In a fourth aspect of the invention, a system for generating a query representation of an electronic query document is provided. The system comprises: a processor for processing the query document to: identify words and sentences in the query document; generate for each word a corresponding part-of-speech, POS, category of the word; and identify each sequence of words having a predetermined sequence of POS categories. The system further comprises: a memory configured to store the identified sequences of words as the query representation of the query document.

In a fifth aspect of the invention, a system for querying a document retrieval system is provided. The system comprises: a terminal configured to provide an electronic query document; a processor for processing the query document to: identify words and sentences in the query document; generate for each word a corresponding part-of-speech, POS, category of the word; and identify each sequence of words having a predetermined sequence of POS categories, a memory configured to store the identified sequences of words as a query representation of the query document. The processor is further configured to: query the document retrieval system with a plurality of the identified sequences of words; and retrieve target documents having at least one sequence of words in common with the query document from the document retrieval system.

In a sixth aspect of the invention, a system for clustering similar documents in a set of electronic documents is provided. The system comprises: a processor configured to: (a) designate one document of the set of documents as a query document; (b) process the query document to: identify words and sentences in the query document; generate for each word a corresponding part-of-speech, POS, category of the word; identify each sequence of words having a predetermined sequence of POS categories; and store the identified sequences of words as a query representation of the query document; (c) query each remaining one of the set of documents with a plurality of the stored identified sequences of words; (d) provide a similarity value for each query of a remaining one of the set of documents; (e) repeat steps (a) – (d) for each remaining one of the set of documents by designating each remaining one of the set of documents as a query document; and (f) cluster documents in the set of documents based on the similarity values.

In a seventh aspect of the invention, a computer program is provided. The computer program comprises computer instructions enabling a processor executing the computer instructions to carry out any of the methods of the invention. The computer software may be executed locally, in a user terminal or client terminal, or remotely, in a server.

These and other aspects of the invention will be more readily appreciated as the same becomes better understood by reference to the following detailed description and considered

in connection with the accompanying drawings in which like reference symbols designate like parts.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 depicts a diagram of an embodiment of a computer-implemented information system to illustrate possible implementations of the present invention.

Figure 2 depicts a flow diagram illustrating an embodiment of a method of the present invention to provide a query representation of a query document.

Figure 3 depicts a flow diagram illustrating an embodiment of a method of the present invention to retrieve a target document from a document retrieval system.

Figure 4 depicts a flow diagram illustrating an embodiment of a method of the present invention to enable clustering of similar documents.

Figure 5 depicts a matrix or table illustrating a clustering of documents based on evaluation of similarity values.

DETAILED DESCRIPTION OF EMBODIMENTS

Unless otherwise stated, herein the term “word” is to be taken to include single words, as delimited in a text by spaces.

Further, herein a part-of-speech, POS, tagger is to be taken as any commercially available or privately developed software, to identify text elements and provide tags, also referred to as labels or indicators, to such text elements in electronic documents. A label is a data element representing information. In particular, each word may be given a label. A label may indicate the word to be a noun, an adjective, an article, a preposition, etcetera.

Figure 1 depicts a diagram of interconnected or coupled data processing entities in an embodiment of the invention. Lines between entities indicate a wired or wireless connection or coupling between such entities. Each connection or coupling is configured to carry data traffic for communication between entities to exchange information, and to control processing at the entities.

A client terminal 100 comprises a data processor 101, a memory 102, at least one input device 103 to input data and control commands at the client terminal 100, at least one output device 104 to output data at the client terminal 100, and a communication unit 105 configured to provide communication with other computer devices. The client terminal 100 may comprise a desktop computer, a laptop computer, a tablet computer, a personal digital assistant, PDA, a smartphone, and/or any other device comprising a data processor 101, a memory 102, input device 103, output device 104 and communication unit 105. The data

processor 101 may comprise one or more data processing units. The memory 102 may comprise permanent memory, such as read only memory, ROM, and non-permanent memory, such as random access memory, RAM. The input device 103 may comprise a keyboard, a mouse device, a touch detection device such as a touch screen, a microphone to control speech recognition software, a camera, a scanner, and/or any other device to provide input to the client terminal 100. The output device 104 may comprise a display such as a screen display or a touch screen, a speaker, and/or any other device to provide output from the client terminal. The communication unit 105 is configured to transmit and receive data from other devices, possibly in various data formats and according to various standards or protocols.

The client terminal 100 may include or be connected or coupled to a database 110 storing a plurality of documents. Alternatively, the client terminal 100 may store a plurality of documents in the memory 102 thereof.

The client terminal 100 may be connected or coupled to a data communication network 120, such as the Internet, a local area network, LAN, a telecommunication network, or any other network configured to provide data communication between electronic devices.

Through the data communication network 120, the client terminal 100 may access information retrieval systems, such as server device 130, which may include or be connected or coupled to a database 140 storing a plurality of indexed documents, and server device 150, which may include or be connected or coupled to a database 160 storing a plurality of indexed documents.

Figure 2 shows a flow diagram of steps executed by computer software comprising computer instructions configured for, when run in a processor of a computer system, performing these steps. The steps are to extract a plurality of word sequences from the query document, where the plurality of word sequences forms a query representation of the query document. The query representation may form the basis of a query to be performed in one or more document retrieval systems.

In a first step 200, a given document, called a query document, is provided. The query document may be any electronic document in any electronic format from which a text may be extracted. The query document may have any length, and may be written in any language. The query document may contain images or other graphical features, formatting tags or other tags, or any other non-text data. The query document may be provided as a pre-existing document file, may be provided as a temporary file representing a document file or a cut part thereof, may be provided by inputting a text in any way, for example by typing or otherwise forming text characters, or by speech recognition, by optical character recognition, OCR, or may be provided in any other way suitable for processing by a computer system. Herein, any

such way of providing text will be related to as providing a query document as indicated in step 200.

At step 202 following step 200, the query document is converted into a text format. In this step, images and other graphical features, formatting or other unuseful tags are ignored and/or removed.

At step 204 following step 202, the language of the query document is determined from the text thereof. Computer tools for recognition of the language of a text are known and available to the person of ordinary skill in the art.

At step 206 following step 204, a decision is made depending on the recognized language. If the recognized language of the query document is not the English language, a step 208 is performed. If the recognized language of the query document is the English language, then the step 208 may be by-passed to continue with step 210.

At step 204, multiple languages may be determined to be present in the query document. In such a case, steps 204 – 208 may be performed for each text in a different language.

At step 208, a machine translation of the query document, or a part thereof, being in a non-English language, is performed to translate the query document, or the part thereof, into the English language. Computer tools for translation of a text are known and available to the person of ordinary skill in the art.

At step 210, words and sentences are identified in the query document. First, the query document is split into tokens, such as numbers, punctuation, symbols, and words. To perform this step, a tokenizer as known and available to the person of ordinary skill in the art is selected and used. Then, the query document is split into sentences. To perform this step, a sentence splitter as known and available to the person of ordinary skill in the art is selected and used.

At step 212, for each word a corresponding part-of-speech, POS, category of the word is generated. To perform this step, a POS tagger as known and available to the person of ordinary skill in the art is selected and used.

A POS tagger performs a morphological and syntactical analysis of sentences to determine the POS category for each word in a sentence and to label the words accordingly. More in detail, the POS tagger analyzes the sentences morphologically and syntactically, and provides an output comprising a sequence of tokens (words) from each sentence, and a corresponding sequence of the POS categories of the tokens (words). Examples of POS categories are noun, adjective, adverb, preposition, determiner, etcetera. Table I below lists different sequences of POS categories to be used in generating a query representation of a query document, and illustrates an example of a sequence of words having such sequence of

POS categories. The set of POS sequences listed in Table I is primarily directed to the English language. For other languages, a different set of POS sequences may be used.

Table I – Predetermined sequences of POS categories – English language

POS SEQUENCE	EXAMPLE WORD SEQUENCE
adjective noun	final version
adjective noun noun	final document version
adjective noun noun noun	electronic document management system
adjective adjective noun	final electronic version
adjective adjective noun noun	final electronic document version
noun noun	document version
noun noun noun	document management system
noun noun noun noun	web content management system
noun adjective noun	document-centric system
noun preposition noun	publication of documents
noun past participle noun	browser-based system
present participle noun	corresponding version
present participle noun noun	corresponding document version
past participle noun	published document
past participle noun noun	published document version

At step 214, predetermined sequences of POS categories are compared with the sequences of POS categories identified in the query document in the previous step 212, based on the POS category. Herein, the term “sequence” is to be taken as comprising at least two elements, where the order of the elements is also taken into account.

At step 216, when a sequence of POS categories identified in the query document matches a predetermined sequence of POS categories, the sequence of words in the query document corresponding to this matched sequence of POS categories in the query document are stored in a memory. When at different instances in the query document a sequence of POS categories identified in the query document matches a predetermined sequence of POS categories, each sequence of words in the query document corresponding to the matched sequence of POS categories is stored in the memory. If a particular sequence of words, identified at a second instance in the query document, is already stored in the memory for reason that the same sequence of words was identified earlier at a first instance in the query document, it may be decided to discard the sequence of words identified at the second

instance without storing the sequence, since this would result in storing the same sequence of words twice.

Step 216 may be repeated until, for each predetermined sequence of POS categories, all corresponding sequences of POS categories identifiable in the query document have been identified and the corresponding, mutually different, sequences of words in the query document have been stored. This generally means that a same word may appear in more than one sequence of words.

When a large, or a maximum number of word sequences are extracted from a query document, based on a predetermined set of predetermined sequences of POS categories, this may lead to a more than 100% representation of the query document, in the sense that the number of extracted word sequences exceeds the number of word sequences intended by the author of the query document. For example, 'high blood pressure' is one concept from the author, but with a sequence of POS categories "noun noun" and "adjective noun", it also leads to the extraction of 'blood pressure' and 'high blood'. However, even if some word sequences might not be relevant for the meaning of the text of the query document, they do not disturb the quality of the query document because of the linguistic phenomenon of redundancy: the value of the composition of the whole is not disturbed by the presence of non-relevant or less relevant word sequences. Optionally, a user may remove word sequences, in particular non-relevant or less relevant word sequences, from the stored set of word sequences.

It is noted that in the method illustrated by reference to Figure 2, it has been assumed that a query document, or a part thereof, being in a non-English language, is to be translated into the English language before taking further steps. In other embodiments, such a conditional translation step may be omitted, and the original language of the query document, or a part thereof, may be retained in further steps. In particular, a POS tagger for the identified language may be used to generate, for each word identified in the query document, a corresponding POS category of the word.

When the processing of the query document has resulted in a representation of the query document in an extracted and stored set of word sequences as described above, a query can be built using the extracted word sequences, to retrieve target documents from a document retrieval system, such as a server device 130, 150 coupled to a database 140, 160, respectively. This document retrieval is illustrated by reference to the flow diagram of Figure 3.

At step 300, a query document is provided, i.e. defined, indicated, prepared or selected. For example, a query document may be an existing document, a part thereof, a newly created document, a pointer to a document, etc.

At step 302, a query representation of this query document is made in accordance with the method as explained above by reference to Figure 2.

At step 304, all of the word sequences of the query representation of the query document are submitted to the document retrieval system. Alternatively, a limited number of word sequences are selected from among all of the word sequences extracted from the query document, to be submitted to the document retrieval system. The document retrieval system will respond by providing information about target documents matching the query document.

At step 306, provided that there is sufficient similarity between the query document and at least one target document, the at least one target document is retrieved.

In this query, each of the word sequences extracted from the query document are submitted to a document retrieval system. The document retrieval system should be full-text searchable and index-based, should allow word sequence searches, and should return results ranked by relevance. The extracted word sequences may be linked by a Boolean 'or' statement to form a search expression. Since the vast majority of document retrieval systems has such properties, the query based on the representation of the query document obtained as described above is widely useable, and yields fast results. For the query method of the present invention, a document retrieval system does not require any special preprocessing or indexing based on semantic or linguistic techniques.

The word sequences extracted from the query document, may be submitted, in series (sequentially) or in parallel, possibly 'or'-ed, to (an index of) a document retrieval system, resulting in relevance scores for target documents retrieved. For a given query document, the number of queries may equal the number of extracted word sequences, without a limit. However, for performance reasons it may be decided to limit the number of word sequences that are submitted to the document retrieval system, because the redundancy of language usage in documents allows for substantial minorization of the identified word sequences, while still having relevant retrieval results.

In some embodiments, the system may enable users to expand one or more word sequence queries using a stemming or morphological algorithm to retrieve morphological variants of one or more word sequence constituents. In some embodiments, a lexical resource like a dictionary or thesaurus may be used to enrich a query by including synonyms, translations or related concepts.

The hit scores for target documents for all queried word sequences may be collected and calculated in a statistical overview which, by means of accumulation, allows to rank the target documents by relevance.

The present invention further allows for determining the relative degree of similarity of a query document and a target document, on the basis of the relative overlap of word sequences between the query document and the target document, for a plurality of target

documents. If 100% of the word sequences extracted from the query document are matched in a target document with high relevance scores for each word sequence, then a target document has been identified that is completely identical to the query document. Accordingly, the present invention uses the extracted word sequences in their combination and accumulation as a representation of the query document, and thereby guarantees to identify the most similar target document with high linguistic preciseness and completeness.

It is noted that for the person of ordinary skill in the art, various tools and methods are known and available that can be used and configured to perform a relevance score calculation.

It is further noted that bibliographic information associated with each of the target documents may be retrieved. As an example, an expert in a specific field may be found by providing a query document containing a description of subject matter, where the document retrieval system to be queried comprises correspondence, publications, curricula vitae and/or other documents of potential experts. Thus, the present invention may be used to find one or more experts.

The present invention provides not only a method and system for querying a document retrieval system by comparing one query document with a plurality of target documents, and calculating a degree of similarity between the query document and (each of) the target documents. As illustrated in the flow diagram of Figure 4, the present invention can also be used in calculating a degree of similarity between documents in a set of documents.

At step 400, a set of documents is provided.

At step 402, one document of the set of documents is selected to be a query document.

At step 404, a query representation of this query document is made in accordance with the method as explained above by reference to Figure 2.

At step 406, if it is assumed that the set of documents contains N documents, the query defined by the query document is performed on the remaining N-1 target documents of the set of documents.

At step 408, for each of the N-1 target documents, a similarity value, relating each of the N-1 target documents to the query document, may be calculated, resulting in a series of N-1 similarity values.

As indicated by a dashed arrow in Figure 4, the steps 402, 404, 406 and 408 may be repeatedly performed for each of the N documents. At each repetition, a new one of the documents in the set of documents is selected as a query document in step 402, and a new series of similarity values may be calculated in step 408 for each one of the remaining N-1 target documents.

As an example, Figure 5 depicts a matrix containing similarity values (relevance scores) for different documents Doc1, Doc2, Doc3, Doc4, Doc5, Doc6 and Doc7. Such a matrix delivers the raw data for clustering of similar documents in the set of documents.

Returning to Figure 4, at step 410, one or more clusters of documents are determined based on the series of similarity values generated in the repetition of the steps 402, 404, 406 and 408.

As an example, the similarity values may be compared with a threshold. If the similarity value is below the threshold, it is ignored. If the similarity value is above the threshold, this implicates a relevant degree of similarity between the query document concerned and the target document concerned.

As explained above, in a method and system of generating a query representation of an electronic query document, the query document is processed by a computer processor. The computer processor is configured to identify words and sentences in the query document, generate for each word a corresponding part-of-speech, POS, category of the word, identify each sequence of words having a predetermined sequence of POS categories, and store the identified sequences of words as the query representation of the query document. In a method and system for querying a document retrieval system, the document retrieval system is queried with a plurality of the stored identified sequences of words; and target documents are retrieved from the document retrieval system. The target documents have at least one sequence of words in common with the query document. In a method and system for clustering similar documents in a set of electronic documents, one document of the set of documents is designated as a query document. The query document is processed to store identified sequences of words as a query representation of the query document. Each remaining one of the set of documents is queried with a plurality of the stored identified sequences of words. A similarity value for each query of a remaining one of the set of documents is determined, and documents in the set of documents are clustered based on the similarity values.

As required, detailed embodiments of the present invention are disclosed herein; however, it is to be understood that the disclosed embodiments are merely exemplary of the invention, which can be embodied in various forms. Therefore, specific structural and functional details disclosed herein are not to be interpreted as limiting, but merely as a basis for the claims and as a representative basis for teaching one skilled in the art to variously employ the present invention in virtually any appropriately detailed structure. Further, the terms and phrases used herein are not intended to be limiting, but rather, to provide an understandable description of the invention.

The terms "a" or "an", as used herein, are defined as one or more than one. The term plurality, as used herein, is defined as two or more than two. The term another, as used

herein, is defined as at least a second or more. The terms including and/or having, as used herein, are defined as comprising (i.e., open language, not excluding other elements or steps). Any reference signs in the claims should not be construed as limiting the scope of the claims or the invention.

The mere fact that certain measures are recited in mutually different dependent claims does not indicate that a combination of these measures cannot be used to advantage.

The term coupled, as used herein, is defined as indirectly connected, although not necessarily mechanically.

A single processor or other unit may fulfil the functions of several items recited in the claims.

The terms software, software program, software application, and the like as used herein, are defined as a sequence of instructions designed for execution on a computer system by a processor of the computer system. A program, computer program, or software application may include a subroutine, a function, a procedure, an object method, an object implementation, an executable application, an applet, a servlet, a source code, an object code, a shared library/dynamic load library and/or other sequence of instructions designed for execution on a computer system.

A computer program may be stored and/or distributed on a suitable medium, such as an optical storage medium or a solid-state medium supplied together with or as part of other hardware, but also be distributed in other forms, such as via the Internet or other wired or wireless telecommunication systems.

CLAIMS

1. A computer-implemented method of generating a query representation of an electronic query document, the method comprising processing the query document by a computer processor configured to:

identify words and sentences in the query document;

generate for each word a corresponding part-of-speech, POS, category of the word;

identify each sequence of words having a predetermined sequence of POS categories; and

store the identified sequences of words as the query representation of the query document.

2. The method of claim 1, wherein the predetermined sequence of POS categories indicates a sequence of at least two words having different POS categories.

3. The method of claim 1, wherein the predetermined sequence of POS categories indicates a sequence of POS categories comprising at least one noun.

4. The method of claim 3, wherein the predetermined sequence of POS categories indicates a sequence of POS categories comprising N nouns, wherein N is integer and $1 \leq N \leq 4$.

5. The method of claim 3, wherein a predetermined sequence of POS categories indicates a sequence of POS categories further comprising at least one of:

- an adjective;
- a preposition;
- a present participle;
- a past participle.

6. The method of claim 3, wherein the predetermined sequence of POS categories indicates a sequence of POS categories selected from a group of sequences of POS categories comprising the sequences of POS categories:

- adjective noun;
- adjective noun noun;
- adjective noun noun noun;
- adjective adjective noun;
- adjective adjective noun noun

- noun noun;
- noun noun noun;
- noun noun noun noun;
- noun adjective noun;
- noun preposition noun;
- noun past participle noun;
- present participle noun;
- present participle noun noun;
- past participle noun;
- past participle noun noun.

7. The method of claim 1, further comprising, after identifying a sequence of words having a predetermined sequence of POS categories:

comparing the identified sequence of words with a stored sequence of words having the same predetermined sequence of POS categories and, if the identified sequence of words is equal to the stored sequence of words, discarding the identified sequence of words without storing it.

8. The method of claim 1, further comprising:

prior to identifying words in the query document, if the language of the query document or a part thereof differs from English, machine-translating the query document or part thereof into the English language.

9. A computer-implemented method for querying a document retrieval system, the method comprising:

providing an electronic query document;

processing the query document by a computer processor configured to:

identify words and sentences in the query document;

generate for each word a corresponding part-of-speech, POS, category of the word;

identify each sequence of words having a predetermined sequence of POS categories; and

store the identified sequences of words as a query representation of the query document,

querying the document retrieval system with a plurality of the stored identified sequences of words; and

retrieving from the document retrieval system target documents having at least one sequence of words in common with the query document.

10. The method of claim 9, further comprising:

retrieving bibliographic information associated with each of the target documents.

11. A computer-implemented method for clustering similar documents in a set of electronic documents, the method comprising:

(a) designating one document of the set of documents as a query document;

(b) processing the query document by a computer processor configured to:

identify words and sentences in the query document;

generate for each word a corresponding part-of-speech, POS, category of the word;

identify each sequence of words having a predetermined sequence of POS categories; and

store the identified sequences of words as a query representation of the query document;

(c) querying each remaining one of the set of documents with a plurality of the stored identified sequences of words;

(d) providing a similarity value for each query of a remaining one of the set of documents;

(e) repeating steps (a) – (d) for each remaining one of the set of documents by designating each remaining one of the set of documents as a query document; and

(f) clustering documents in the set of documents based on the similarity values.

12. A system for generating a query representation of an electronic query document, the system comprising:

a processor for processing the query document to:

identify words and sentences in the query document;

generate for each word a corresponding part-of-speech, POS, category of the word; and

identify each sequence of words having a predetermined sequence of POS categories,

the system further comprising:

a memory configured to store the identified sequences of words as the query representation of the query document.

13. A system for querying a document retrieval system, the system comprising:

a terminal configured to provide an electronic query document;

a processor for processing the query document to:

identify words and sentences in the query document;

generate for each word a corresponding part-of-speech, POS, category of the word; and

identify each sequence of words having a predetermined sequence of POS categories,

a memory configured to store the identified sequences of words as a query representation of the query document,

wherein the processor is further configured to:

query the document retrieval system with a plurality of the identified sequences of words; and

retrieve target documents having at least one sequence of words in common with the query document from the document retrieval system.

14. A system for clustering similar documents in a set of electronic documents, the system comprising:

a processor configured to:

(a) designate one document of the set of documents as a query document;

(b) process the query document to:

identify words and sentences in the query document;

generate for each word a corresponding part-of-speech, POS, category of the word;

identify each sequence of words having a predetermined sequence of POS categories; and

store the identified sequences of words as a query representation of the query document;

(c) query each remaining one of the set of documents with a plurality of the stored identified sequences of words;

(d) provide a similarity value for each query of a remaining one of the set of documents;

(e) repeat steps (a) – (d) for each remaining one of the set of documents by designating each remaining one of the set of documents as a query document; and

(f) cluster documents in the set of documents based on the similarity values.

15. A computer program comprising computer instructions enabling a processor executing the computer instructions to carry out the method of claim 1, 9 or 11.

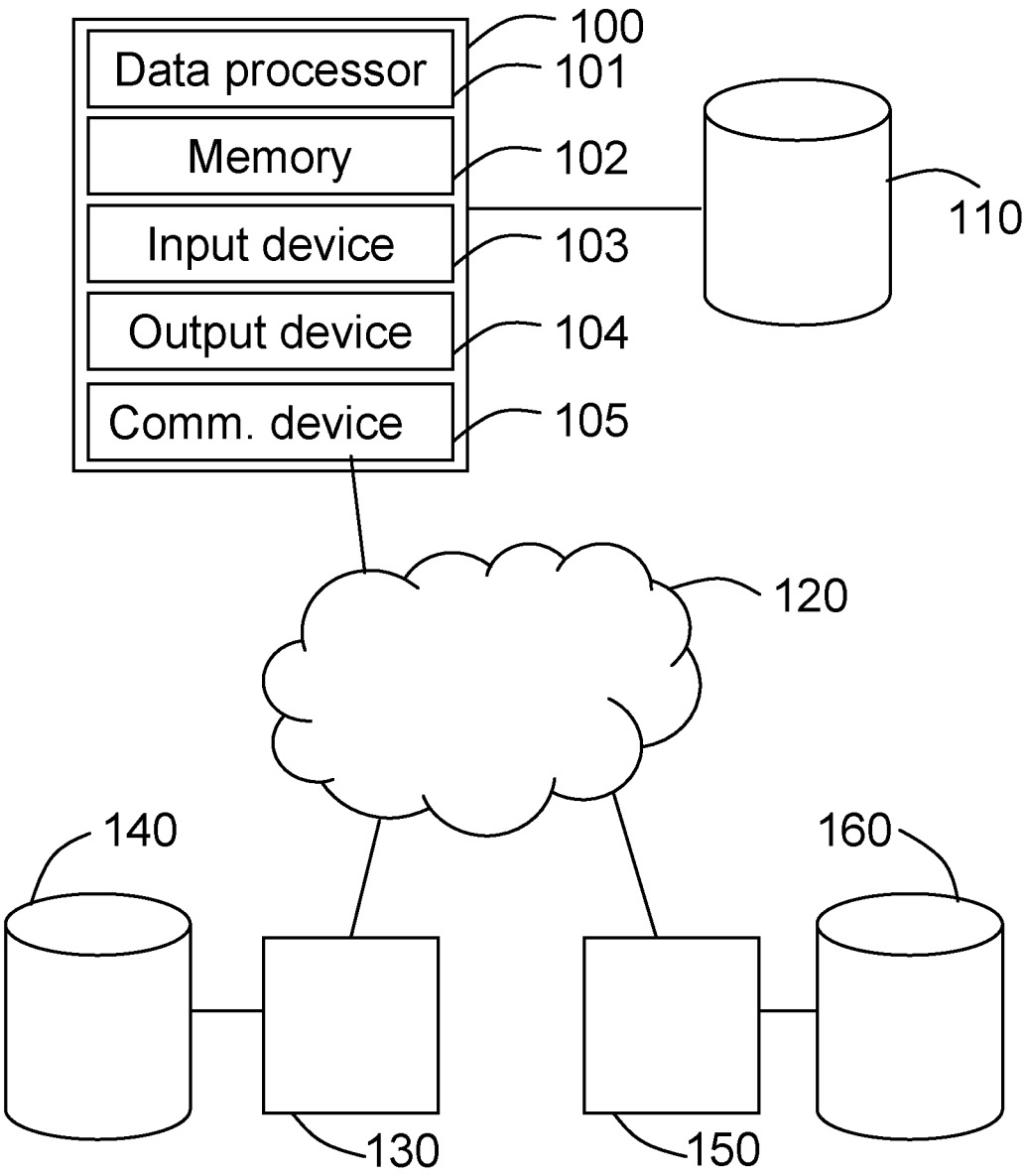

Data processor
Memory
Input device
Output device
Comm. device
100
101
102
103
104
105
110
120
130
140
150
160

FIG. 1

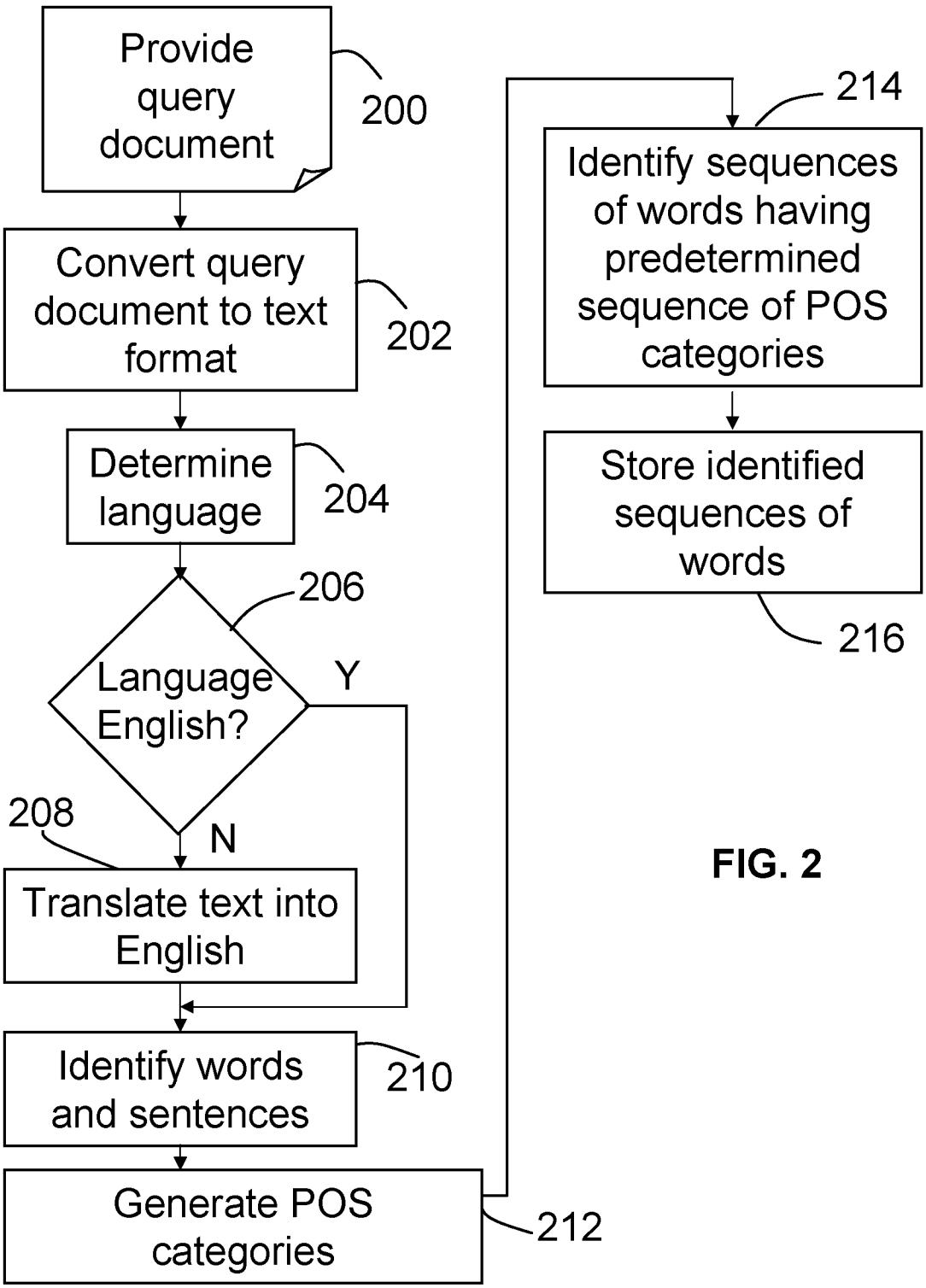

Provide query document
200
Convert query document to text format
202
Determine language
204
206
Language English?
Y
N
208
Translate text into English
Identify words and sentences
210
Generate POS categories
212
214
Identify sequences of words having predetermined sequence of POS categories
Store identified sequences of words
216

FIG. 2

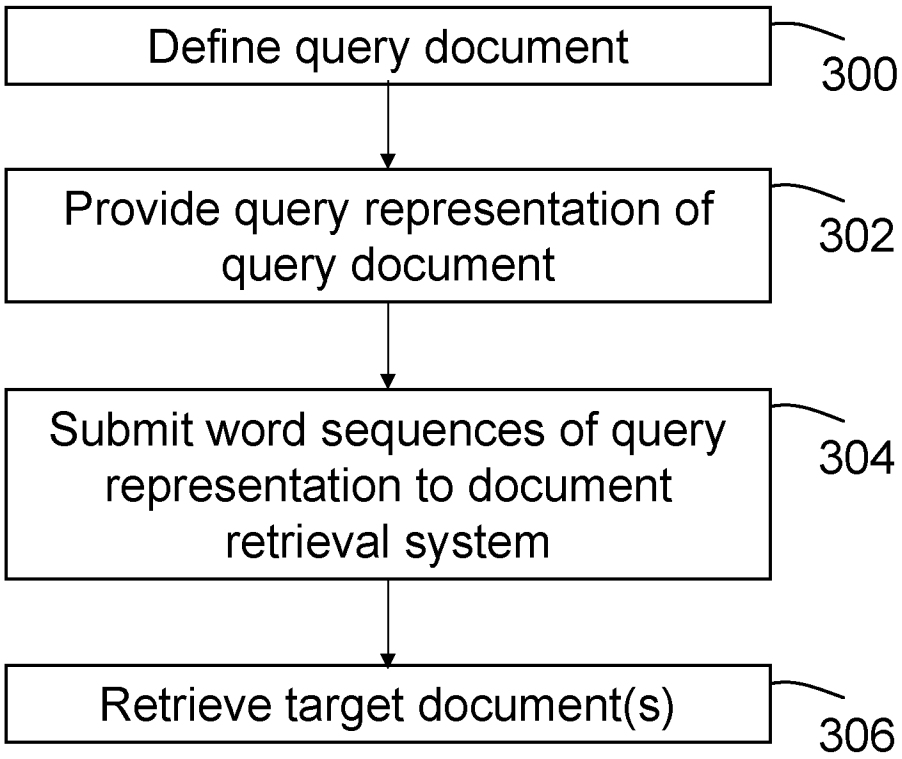

Define query document
300
Provide query representation of query document
302
Submit word sequences of query representation to document retrieval system
304
Retrieve target document(s)
306

FIG. 3

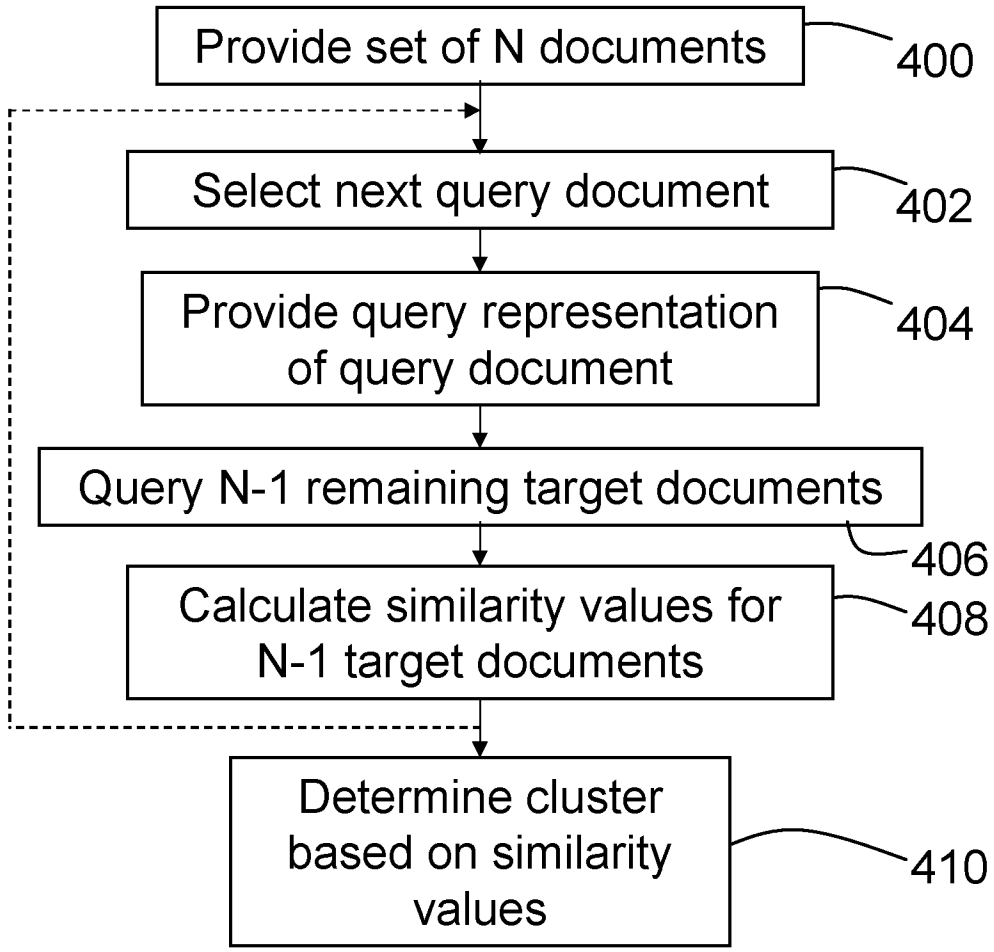

Provide set of N documents
400
Select next query document
402
Provide query representation of query document
404
Query N-1 remaining target documents
406
Calculate similarity values for N-1 target documents
408
Determine cluster based on similarity values
410

FIG. 4

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7
Doc1	1.00	0.23	0.26	0.77	0.41	0.05	0.12
Doc2	0.23	1.00	0.17	0.02	0.62	0.38	0.34
Doc3	0.26	0.17	1.00	0.31	0.11	0.17	0.44
Doc4	0.77	0.02	0.31	1.00	0.23	0.84	0.91
Doc5	0.41	0.62	0.11	0.23	1.00	0.56	0.39
Doc6	0.05	0.38	0.17	0.84	0.56	1.00	0.81
Doc7	0.12	0.34	0.44	0.91	0.39	0.81	1.00

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No
PCT/NL2012/050463

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F17/30
ADD.

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)
G06F

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)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 696 962 A (KUPIEC JULIAN M [US]) 9 December 1997 (1997-12-09) column 1, line 65 - column 2, line 13 column 2, line 65 - column 3, line 9 column 5, line 7 - line 8 column 5, line 42 - line 46 column 7, line 63 - column 8, line 3 column 8, line 18 - column 9, lines 7, 29-41, 48-52 -----	1-15
A	EP 0 597 630 A1 (CONQUEST SOFTWARE INC [US]) 18 May 1994 (1994-05-18) page 5, line 35 - line 58 page 7, line 46 - page 8, line 50 page 12, line 45 - page 15, line 4 ----- -/--	1-15

[X] Further documents are listed in the continuation of Box C.

[X] 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	Date of mailing of the international search report
1 November 2012	09/11/2012

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Haffner, Ronald

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/NL2012/050463

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2002/103799 A1 (BRADFORD ROGER BURROWES [US] ET AL) 1 August 2002 (2002-08-01) paragraph [0045] paragraph [0064] - paragraph [0066] -----	1-15
A	US 7 672 831 B2 (TODHUNTER JAMES [US] ET AL) 2 March 2010 (2010-03-02) column 4, line 30 - line 62 column 6, line 9 - line 14 column 7, line 54 - line 57 column 8, line 64 - column 9, line 2 -----	1-15

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/NL2012/050463

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5696962 A	09-12-1997	JP 7056933 A	03-03-1995
		US 5696962 A	09-12-1997
EP 0597630 A1	18-05-1994	AT 221683 T	15-08-2002
		AU 1998497 A	03-07-1997
		AU 5047893 A	19-05-1994
		CA 2102325 A1	05-05-1994
		DE 69332166 D1	05-09-2002
		EP 0597630 A1	18-05-1994
		IL 107482 A	30-10-1998
US 2002103799 A1	01-08-2002	NONE	
US 7672831 B2	02-03-2010	EP 1941405 A2	09-07-2008
		KR 20080084803 A	19-09-2008
		US 2007094006 A1	26-04-2007
		WO 2007051109 A2	03-05-2007