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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 October 2012 (26.10.2012)

WIPO | PCT

(10) International Publication Number
WO 2012/143839 A1

(51) International Patent Classification:
*G06F 17/27* (2006.01)

(21) International Application Number:
PCT/IB2012/051870

(22) International Filing Date:
16 April 2012 (16.04.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/476,917 19 April 2011 (19.04.2011) US

(72) Inventor; and
(71) Applicant : **GREYLING, Abraham, Carel** [ZA/ZA]; 48 Beverly Hills Crescent, Centurion Gold Estate, Tshwane, 0157 Centurion (ZA).

(74) Agent: **VON SEIDELS INTELLECTUAL PROPERTY ATTORNEYS**; P O Box 440, Century City, 7446 Cape Town (ZA).

(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, MD, 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).

**Declarations under Rule 4.17**:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

**Published**:

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

*[Continued on next page]*

(54) **Title**: A COMPUTERIZED SYSTEM AND A METHOD FOR PROCESSING AND BUILDING SEARCH STRINGS

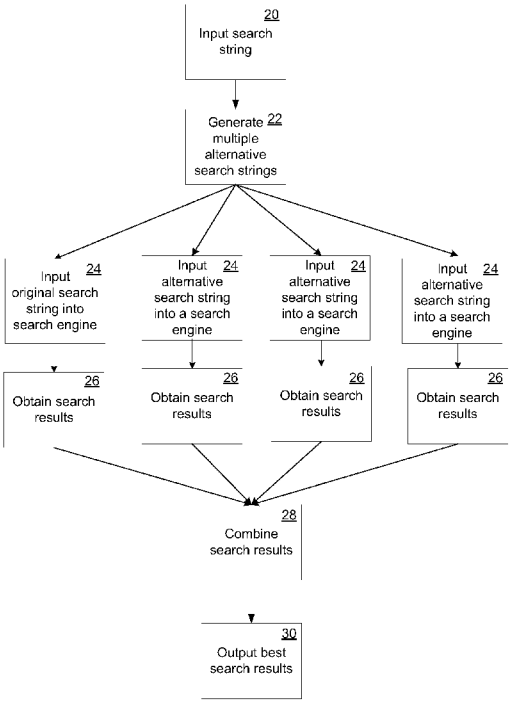


Figure 1

(57) **Abstract**: A method is provided for processing an input search string and building multiple alternative search strings to improve computerized search. The method includes extracting text from web pages, forming a word-relationship database in which every unique word is associated with fields which represent other words found to occur adjacent to that word, processing the word relationship database so as to determine a forward and reverse signature for each word, and combining the forward and reverse signatures to form a signature database. Two-word groups in the input search string are linked and the forward and reverse signatures for each two-word group obtained. These signatures are compared with the signature database to find single words that have signatures that closely match the signature of the two-word group, and those words identified as alternative words that are semantically similar to the two-word group, so as to generate alternate search strings.

WO 2012/143839 A1

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

# A COMPUTERIZED SYSTEM AND A METHOD FOR PROCESSING AND BUILDING SEARCH STRINGS

## FIELD OF THE INVENTION

This invention relates to query processing, and more specifically relates to the semantic analysis of search query strings to generate multiple alternative strings to facilitate improved computerized search.

## BACKGROUND TO THE INVENTION

Enabling computers to understand language remains one of the hardest problems in artificial intelligence. Language is highly contextual. Often the same words have different meanings in different contexts and small differences in sentence structure can lead to totally different meanings. At the same time, a great number of different sentence structures can have the same meaning.

Most internet search engines use text-based input search queries. To return accurate search results, the search engine must be able to apply some form of language interpretation to the search string entered by a user. The search engine must also apply language interpretation when it indexes web pages or other documents, so that the search string can be matched to web pages by a ranking algorithm that only delivers the most relevant results to a user.

One simple form of language interpretation is to analyze each word in the search string and to also search for the synonyms for certain of the words. Thus the word “picture” may have the synonym “photo” so that if a user searches for “picture of Grand Canyon”, the search engine must also return results for web pages in which the words, “photo of Grand Canyon” appear.

However, merely applying synonyms can also lead to wrong results. For example, if a user searches for “history of motion pictures” then the word “pictures” must not be substituted with “photos” because the string “history of motion photos” is meaningless. As another example, if a user were to search for “HP wide screen monitor” and the search engine were also to substitute the synonym “detector” for “monitor”, and “shutter” for “screen”, completely irrelevant results would be delivered. From the above it is clear that a search engine also needs to be able to perform contextual analysis so that it knows, for example, that the string “HP wide screen monitor” has nothing to do with shutters or detectors and that the term “motion photo” is not the same as “motion picture”.

As a further illustration of this problem, even words which are normally interchangeable can lead to totally different meanings when used in different contexts. A search for “arm reduction” probably has to do with cosmetic surgery whereas “arms reduction” relates to reducing stockpiles of weaponry. When longer sentences are involved, the permutations become exponentially more complex.

Even the most advanced search engines available today, such as the Google™ search engine, are generally not able to accurately interpret longer search queries so as to deliver meaningful results. Therefore, a search on Google™ for “Software companies founded before 1990 with a current turnover of more than $100 million” yields a list of largely irrelevant references, even though the search query is perfectly clear to a human and the information is doubtless available on the Internet. Because existing search engines rely primarily on keywords rather than the context of words, most people have learned through experience to write search queries in what has been dubbed “caveman speak”; where, for example, a user wanting to know about popular seafood restaurants in Seattle might search for “seafood Seattle” rather than “Where I can I find a good seafood restaurant in

Seattle?". Existing search engines are not able to properly analyze complex contextual meanings created by combining words into a sentence structure.

The "Semantic Web" refers to a structure for the Internet in which machine-readable data (or meta-data) is available that tells a computer unambiguously what a web page, a document or a topic is about. This meta-data enables computers to understand the meaning of information directly, without the interpretation problems that plague current search engines. Currently, certain defined domains – for example, airline booking systems – operate in this way. Thus the term "JFK" in an airline booking system refers only to John F Kennedy International airport in New York, not to the former US president or other terms that may have these three letters as their acronym.

Some search engines, such as Bing™, identify categories based on the search terms, and a user is able to filter out irrelevant results by only selecting certain categories. Thus a search for "chicken" might identify categories of "animals" and "recipes" and allow the user to filter so as to only search within one of the two categories.

However, the goal of the Internet itself being semantic has not yet been realized, despite ongoing efforts to index and associate concepts on the Internet. The main problem is the enormity of the task involved in performing such identification and association on the open Internet, which requires a huge structured database to be built.

It would be advantageous to have a completely autonomous system that is able to build a contextual language model so that search strings can be interpreted more accurately by search engines, so as to deliver more relevant and targeted search results without the need to categorize or index existing content.

## SUMMARY OF THE INVENTION

In accordance with the invention there is provided a method for processing an input search string and building multiple alternative search strings, the method comprising:

extracting text from a multitude of electronically accessible documents or web pages, the text including words;

forming a word relationship database in which each unique word identified in the text is stored in the database and is associated with a number of fields which each represent other words which were found to occur adjacent to that word in the text, each field also including a frequency sub-field which indicates how frequently that other word was found to occur adjacent to the associated word;

processing the word relationship database so as to determine a forward signature and reverse signature for each word, the forward signature including a ranked list of the words that were found to come after that word in the text, and the reverse signature including a ranked list of the words that were found to come before that word in the text;

combining the forward and reverse signatures of each word to form an ambidextrous signature for each word, and storing the ambidextrous signatures in a signature database;

optionally removing popular words from the input search string, being those words with a total frequency higher than a predetermined threshold;

linking the remaining words of the input search string into two-word groups from left to right with the second word of any preceding two-word group forming the first word of the next two-word group;

for each two-word group, carrying out the following steps:

querying the word relationship database to determine the reverse signature of the first word and the forward signature of the second word;

combining the forward and the reverse signatures obtained in the previous step into an ambidextrous signature representative of that two-word group;

comparing the ambidextrous signature of the previous step with the signature database to find the closest matches;

identifying the words in the signature database with the closest signature matches as alternative words that are semantically similar to the two-word group;

substituting one or more of the two-word groups in the input search string with the identified alternative words;

analyzing whether each substituted word fits grammatically into the input search string by querying the word relationship database to see whether the word preceding and the word following the substituted word in the input search string are words that are associated with the substituted word to a predefined extent; and

if the substituted word or words do fit grammatically, identifying the string with the substituted words as an alternative search string that is both semantically similar to the original search string and grammatically correct.

Further features of the invention provide for the text to be extracted from the web pages or documents by autonomous web crawling programs; for the word relationship database to be continually updated as more and more text is extracted; and for the signature database to be periodically rebuilt.

Still further features of the invention provide for the method to include an additional step of, immediately after extracting text and before forming a word relationship database, parsing the text into sentence portions which start and end with sentence delimiters.

Yet further features of the invention provide for techniques to be employed that keep the growth of the word relationship database in check, for example by gradually reducing the frequency sub-fields so that only those words that are frequently incremented will develop large frequencies, or by periodically discarding words with low frequency sub-fields.

Further features of the invention provide for the ambidextrous signature of the two-word group to be matched with ambidextrous signatures in the signature database by calculating a matching score that looks for matches between the fields of the signatures and applies decreasing weighting factors to fields that are associated with lower frequency sub-fields.

Still further features of the invention provide for inputting the multiple alternate search strings into a search engine simultaneously or in rapid succession and comparing the results of each separate search so as to rank the overall results and present those results which were obtained in the greatest number of separate searches as the most relevant search results.

The language of the text is preferably identified so that separate word relationship databases and signature databases can be built for each separate language.

The invention extends to a system for processing an input search string and building multiple alternative search strings, comprising:

a processor in the form of a server which is able to access a multitude of web pages or other documents through the Internet and extract text, the text including words;

a word relationship database coupled to the processor and having each unique word identified in the text stored thereon, each unique word in the word relationship database being associated with a number of fields which each represent other words which were found to occur adjacent to that word in the text, each field also including a frequency sub-field which indicates how frequently that other word was found to occur adjacent to the associated word;

a signature database coupled to the server, the signature database being formed by the server processing the word relationship database so as to determine a forward signature and reverse signature for each word, the forward signature including a ranked list of the words that were found to

come after that word in the text, and the reverse signature including a ranked list of the words that were found to come before that word in the text, and combining the forward and reverse signatures of each word to form an ambidextrous signature for each word that is stored in the signature database;

and computer software stored on the processor and configured to enable the processor to carry out the following:

optionally removing popular words from the input search string, being those words with a total frequency higher than a predetermined threshold;

linking the remaining words of the input search string into two-word groups from left to right with the second word of any preceding two-word group forming the first word of the next two-word group;

for each two-word group, carrying out the following:

querying the word relationship database to determine the reverse signature of the first word and the forward signature of the second word;

combining the forward and the reverse signatures obtained in the previous step into an ambidextrous signature representative of that two-word group;

comparing the ambidextrous signature of the previous step with the signature database to find the closest matches;

identifying the words in the signature database with the closest signature matches as alternative words that are semantically similar to the two-word group;

substituting one or more of the two-word groups in the input search string with the identified alternative words;

analyzing whether each substituted word fits grammatically into the input search string by querying the word relationship database to see whether the word preceding and the word following the substituted word in the input search string are words that are associated with the substituted word to a predefined extent; and

if the substituted word or words do fit grammatically, identifying the string with the substituted words as an alternative search string that is both semantically similar to the original search string and grammatically correct.

Further features of the invention provide for the server to be configured to input the multiple alternative search strings into a search engine simultaneously or in rapid succession, and to compare the results of each separate search so as to rank the overall results and present those results which are obtained in the greatest number of separate searches as the most relevant search results.

## BRIEF DESCRIPTION OF THE DRAWINGS

The invention will now be described, by way of example only with reference to the accompanying representations in which:

Figure 1 is a flowchart that illustrates the overall steps performed in obtaining improved search results where multiple alternative search strings are generated according to the method of the invention;

Figure 2 is a schematic diagram showing the system for generating multiple alternative search strings according to the invention;

Figure 3 is a flowchart that illustrates the steps performed in creating and updating a word relationship database;

Figure 4 is a flowchart that illustrates the steps performed in creating a signature database based on the word relationship database; and

Figure 5 is a flowchart that illustrates the steps performed in using the signatures in the signature database to create multiple alternative search strings according to the method of the invention.

## DETAILED DESCRIPTION WITH REFERENCE TO THE DRAWINGS

Figure 1 is a flowchart that illustrates the overall steps performed and results obtained by the method and system of the invention. As provided by the invention, at a first stage (20), a search string is input into the system of the invention. At a next stage (22), multiple alternative search strings are generated, with each search string being semantically similar to the original search string and containing correct grammar. The original search string and each alternative search string are then input into a search engine at the next stage (24). The search engine may be any computerized search engine, including a web based internet search engine that facilitates keyword searching. The results of each internet search obtained using the search strings are obtained at the next stage (26). These results are then combined at the next stage (28) so as to identify the most relevant search results and output those results at stage (30).

Figure 2 illustrates a system (100) which enables the multiple alternative search strings to be generated, which was stage (22) in Figure 1. The system includes a processor in the form of a server (102) which is able to access a multitude of web pages (104) or other documents through the Internet (106) by means of web crawling programs (not shown). The server is also coupled to a word relationship database (108) and a signature database (110). The word relationship database is built from content obtained from the Internet by the web crawling programs, and the word relationship database is used to create the signature database as will be explained below. These databases can then be used to obtain multiple alternative search strings for a given

input search string according to the method of the invention which will be explained in detail below.

**A. Creating and Updating a Word Relationship Database**

The first stage in the method of the invention is to create a word relationship database where every word in a particular language is associated with words adjacent to it. A flow chart illustrating the steps to create and update a word relationship database is shown in Figure 3.

At a first stage (40), software programs commonly referred to as "spiders" or "robots", automatically access content on thousands or millions of webpages on the Internet and, from each webpage, extract text by accessing the underlying HTML code. This text need not be in English, although it would be advantageous to identify the language of specific sites or domains so that separate word relationship databases can be built for different languages. For the purposes of this description, it will be assumed that English text is being analyzed and input into the word relationship database. Exactly the same methods can be applied to any other language in which context plays an important part in meaning – i.e. where the meaning conveyed by words depends on the words surrounding them.

An example of a small portion of text extracted by a software spider from a given website might be:

> "The Semantic Web is a technology waiting to be actualized. Application areas are experiencing intensified interest due to the rapid growth in the use of the Web. Information content technologies (such as search engines) are constantly being improved, with the hope of the actualization of powerful search technologies."

At a next stage (42), text is then parsed into sentence portions which start and end with sentence delimiters. The following list of ASCII characters are generally regarded as sentence delimiters:

| ASCII no. | Character |
|---|---|
| 9 | TAB |
| 10 | LF (line forward) |
| 13 | CR (Carriage return) |
| 33 | ! |
| 40 | ( |
| 41 | ) |
| 44 | , |
| 45 | - |
| 46 | . |
| 47 | / |
| 58 | : |

| ASCII no. | Character |
|---|---|
| 59 | ; |
| 60 | < |
| 61 | = |
| 62 | > |
| 63 | ? |
| 91 | [ |
| 92 | \ |
| 93 | ] |
| 123 | { |
| 124 | \| |
| 125 | } |

Table 1: Common Sentence Delimiters

Using the example text above, the delimiters in the text are ".", "(", ")" and ",". The text can therefore be parsed into the following sentence portions:

a) The Semantic Web is a technology waiting to be actualized

b) Application areas are experiencing intensified interest due to the rapid growth in the use of the Web

c) Information content technologies

d) such as search engines

e) are constantly being improved

e) with the hope of the actualization of powerful search technologies

Using the parsed sentence portions, a word relationship database can then be formed which shows how adjacent words are related to each other in the body of text analyzed by the web spiders.

At stage (44), the word relationship database is formed as a two--dimensional matrix in which each row represents a particular word (the "row word"), and has a number of row fields that represent specific words that were found to occur *after* the row word in the body of text that was analyzed. Each row field also includes an indication of the frequency, or number of times, that the word was found to occur after the row word in the body of content. This can schematically be illustrated as follows:

**<RowWord1>:** <WordAfter1>, <Freq1> | <WordAfter2>, <Freq2>…
**<RowWord2>:** <WordAfter1>, <Freq1> | <WordAfter2>, <Freq2>…
**<RowWord3>:** <WordAfter1>, <Freq1> | <WordAfter2>, <Freq2>…
**….**

Each word is assigned a unique reference number, and within each row the row fields are ranked according to frequency. For example, consider a very small portion of the two-dimensional matrix, only the words that follow alphabetically between "actuality" and "actualizes". From the sentence portions (e) above, only the word "of" was found to follow the word "actualization". In the sentence portion (a) no word was found after "actualized". If the only text input into the two-dimensional matrix were the sentence portions (a) – (e) above, the matrix portion might look as follows:

**502("actualization"):** | 3488("of"), 1
**503("actualized"):** (1)

The word "of", which has reference number 3488, was found to come after the word "actualization" (reference number 502) only once. Although the word "actualized" (reference number 503) was found, no words after it were found.

Stages (40), (42) and (44) are then cycled through repeatedly as more and more text is extracted from the Internet, parsed into sentence portions, and added to the word relationship database.

Consider how the word relationship database might look after several thousand or hundreds of thousands of words have been parsed into sentence portions and input into the matrix. The words that follow alphabetically between "actuality" and "actualizes" might then have the following entries:

**501("actuality"):** 6638("in"), 13 | 662("at"), 8 | 465("about"), 6 | 3488("of"), 6, 1392("for"), 5
**502("actualization"):** 2227("towards"), 2 | 3488("of"), 1 | 465("about"), 1
**503("actualized"):** 49731("work"), 7 | 1392("for"), 4
**504("actualizes"):** 663 ("anything"), 4 | 3811 ("something"), 3 (2)

Each row represents a unique word ("a row word") and the rows are alphabetically sorted with incrementing reference numbers, words 501-504 ("actuality" – "actualizes"). Each row word has a number of row fields after it. In the case of word 501 ("actuality"), there is an array of 5 row fields following it, whereas for words 503 and 504 ("actualizes" and "actualizes") there are only 2 row fields following each of these words.

Each row field includes two items of information, the reference number of a word that was found to come after it in the body of searched text, and a frequency number which shows the number of times that referenced word was found to come after the row word in the body of searched text. In this case, the word "in" (6638) was found to come after the row word "actuality" (501) the most (13 times), whereas the word "for" (1392) came after "actuality" (501) only 5 times. The phrase "actuality in" was therefore more popular than the phrase "actuality for" in the body of searched text. In the

case of row word 503 (“actualized”), the only words found to follow it were “work” (49731) and “for” (1392). The row fields are sorted according to the frequency with which the words are found, from the highest frequency on the left to the lowest frequency on the right.

It will be appreciated that the row words and field words have been written between inverted commas to aid understanding. In the actual machine-readable word relationship database, only the reference numbers will be used. The portion of the matrix in (2) above therefore looks as follows:

501: 6638, 13 | 662, 8 | 465, 6 | 3488, 6 | 1392, 5
502: 2227, 2 | 3488, 1 | 465, 1
503: 49731, 7 | 1392, 4 (3)
504: 663, 4 | 3811, 3

This extract from the word relationship database is, of course, greatly simplified for illustrative purposes. As more text is searched and indexed by the web crawling programs, the word relationship database increases in size with the frequency numbers growing rapidly and the number of row fields also growing, although not as quickly. To keep this growth in check, a number of techniques can be employed, such as techniques that gradually reduce the frequency fields so that only those field words that are frequently incremented will develop large frequencies. Algorithms for periodically discarding the row fields that have very low frequencies can also be used so as to keep the number of row fields in check, in addition to algorithms that compress the matrix density (the number of row fields multiplied by their frequencies).

Once populated with content from a large number of web pages and other documents, the word relationship database provides an accurate view of the relationship that each word has to the words that come after it in a particular language (such as English), provided of course that the bulk of the content

accessed by the web spiders is not garbled or meaningless, which it should not be if ordinary content on the Internet is being accessed. One would therefore expect the relationship between certain two-word groups to be strong, such as “founding fathers”, while there would be a very weak (or non-existent) relationship between other, even very similar, two-word groups such as “found fathers”.

**B. Creating a Signature Database based on the Word Relationship Database**

Next, according to the method of the invention, a signature database is created that is based on the word relationship database. Figure 4 illustrates the steps to create a signature database based on the word relationship database.

As explained in the previous section, the words in the row field of the word relationship database only indicate words that come *after* the row word. Each row can therefore be thought of as a signature for the words that follow the row word, where the signature tells you the relationship of the row word to other words following it, ranked according to popularity.

At a first stage (50) in Figure 4, the word relationship database is queried to obtain the “reverse signature” of every word, i.e. an indication of the popularity of words that *precede* the word of interest. This can be done by searching the entire word relationship database for every instance where the word of interest appears in a row field, and identifying the row word associated with that row field as the preceding word.

For example, assume that the word “work” has the following row in the word relationship database:

**54731(“Work”):** | 2284(“on”), 15 | 4433(“hard”), 12 | 5333(“towards”), 5 (4)

To find out what the "reverse signature" of "work" is, the entire word relationship database is searched to find every instance where "work" appears in any one of the row fields. In this example, assume that only the following rows are found in which the word "work" appears in the row fields:

**503("actualized"):** 49731("work"), 7 | 1392("for"), 4
**4332("hard"):** 2290("life"), 12 | 49731("work"), 10 | 902("to") (5)
**27332("start"):** 9221("living"), 14 | 49731 ("work"), 13 | 3323 ("button"), 9

Three rows were found with "work" in the row fields. It will be appreciated that the row words of these three rows ("actualized", "hard" and "start") represent the words that were found to come *before* the word "work", and that the number of times they were found to come before work is given by the frequency field in each row field in which "work" appears.

Therefore, the word "actualized" came before "work" 7 times, "hard" came before work 10 times and "start" came before work 13 times. The reverse signature of "work" can therefore be constructed as:

**54731("Work"):** 27332("start"), 13 | 4332("hard"), 10 | 503("actualized"), 7 (6)

In the same way, the "reverse signature" of each word in the word relationship database can be obtained.

At the next stage (52), the forward and reverse signatures of each word are combined into an "ambidextrous signature". To do this, the information about whether the field word came before or after the word of interest is discarded, and the number of times each field word came before or after the word is also discarded, while nevertheless maintaining a ranking based on the number of times each field word came before or after the word. As an example, the "forward signature" of "work" given at (4) and the "reverse"

signature of “work” given at (6) are combined into the following “ambidextrous signature”:

**54731(“Work”):** 4433(“hard”) | 2284(“on”) | 27332(“start”) | 503(“actualized”) | 5333(“towards”) (7)

It will be noted that the frequency fields no longer appear, but that the row fields are still ranked according to the original frequencies in the forward and reverse signatures. The reason that “hard” appears first is because it appeared in both the forward and reverse signatures, and the frequency of both fields (12 and 10 respectively) were added together, yielding a frequency of 22 which is greater than the next highest frequency of 15 for the word “on”.

The ambidextrous signature (7) is therefore a word relationship signature which shows which words are contextually close to the word “work”, in that those words often appear adjacent to the word “work” (either before or after) in the English language.

By processing the word relationship database, word relationship signatures like the one in (7) are then created for every word in the word relationship database. These word relationship signatures are saved in a second database, called the signature database, at stage (54). The signature database was illustrated at (110) in Figure 2. As the word relationship database is continually updated with more and more content obtained by the web crawling programs, and adapted by the various algorithms to control its growth, the signature database is similarly automatically updated by being periodically rebuilt (stage (56)) and having stages (50) to (54) repeating. Over time, therefore, words which previously may have had no relationship to each other, such as “Lady” and “Gaga”, may become strongly associated with each other if lots of content on web pages starts appearing with those words adjacent to each other. In this way, the word relationship signatures in the

signature database change to reflect language as it is commonly written and used.

It will be appreciated that the word relationship signatures only reflect the relationship of specific words to those words that come immediately before or after them, not to more distant word relationships. Using only a two-word relationship means the word relationship database and signature database are two-dimensional matrices, rather than 3-,4- or higher-order matrices. This simplicity is important because it keeps the size of the word relationship database and signature database manageable and makes it very scalable.

**C. Using the Signatures in the Signature Database to Create Multiple Alternative Search Strings**

Finally, according to the method of the invention, the word relationship signatures in the signature database are used to create multiple alternative search strings that are semantically similar to an input search string and grammatically correct. This is the step that was indicated broadly by stage (22) in Figure 1 and which will now be described in detail. The various stages involved in generating the multiple alternative search strings are illustrated in Figure 5.

At a first stage (60), popular words are removed from the input search string. Popular words are identified as those words with a total frequency in the entire word relationship database that is higher than a predetermined threshold – in other words, those words that appear very commonly in the total body of text accessed by the web crawling programs. As an example, consider the search string “Where can I get cool spring water?”. The words “where”, “can”, “I” and “get” will likely be identified as popular words, with the remaining words “cool spring water” being non-popular words.

Next, at stage (62), the non-popular words are linked in two-word groups from left to right with the last word of any preceding two-word group forming the first word of the next two-word group. In this case, there are two two-word groups, namely “cool spring” and “spring water”.

Next, at stage (64), each two word group is analyzed as follows: the reverse signature of the first word and the forward signature of the second word are obtained. Then, at stage (66), the forward and reverse group signatures are combined into a single ambidextrous “word-group” signature.

For example, if the forward signature of “spring” in the word relationship database is the following:

**42551 (“spring”):** 2211 (“day”), 21 | 53342 (“was”), 15 | 3321 (“morning”), 4 (8)

and the reverse signature of “cool” is the following:

**1221 (“cool”):** 49923 (“very”), 19 | 3221 (“stay”), 13 | 9219 (“really”), 8 (9)

then the ambidextrous signature of “cool spring” could be the following:

**(“cool spring”):** 2211 (“day”) | 49923 (“very”) | 53342 (“was”) | 3221 (“stay”) | 9219 (“really”) | 3321 (“morning”) (10)

Importantly, the ambidextrous word relationship signature in (10) gives the forward and reverse relationship of the two words “cool spring” *in combination*, as if they were a single word.

Next, at stage (68), the signature database is searched to look for close signature matches for the ambidextrous “word group” signature (10). By comparing the signature in (10) to the word signature database and looking

for close matches, *single* words can be found that are semantically similar to the two word group, “cool spring”. This comparison can be done in various ways. One way is to calculate a matching score between the signature (10) and each of the signatures in the signature database by an algorithm that looks for matches between the fields of the signature (10) and the fields of each of the signatures in the signature database. Decreasing weighting factors can be allocated to each of the fields with the signature so that matches between fields that are further to the right count less than matches between fields that are further left. The algorithm can also allocate a higher weighting factor if the word in the signature database that includes matching fields is not a common word, as these words give more information than common words such as prepositions and conjunctions. At stage (70), the word or words that have the highest weighting factor are then identified as the words that are semantically similar to the two-word group.

As shown at stage (72), stages (64) to (70) are then repeated for each of the other two-word groups in the search string, which in this example is the second two-word group, “spring water”. In this way, one or more other words are identified that are semantically similar to “spring water”. Combining the results of both iterations yields a number of two word strings that are each semantically similar to “cool spring water”. For example, if one of the words identified as semantically similar to “cool spring” was “refreshing” and one of the words identified as semantically similar to “spring water” was “liquid”, then “refreshing liquid” would be identified as semantically similar to “cool spring water”.

Using the substitute word or words for “cool spring” and “spring water”, and repeating the procedure with the substitute two words (e.g. “refreshing liquid”), it is possible to repeat stages (64)-(70) to find individual words that are semantically similar to the three words, “cool spring water”. In this example, the single word “juice” could, for example, be identified as semantically similar to “refreshing liquid”.

It will be appreciated that, by repeating the substitution procedure in stages (64) – (70) a number of times, it is possible to obtain multiple alternate words for the extracted non-popular words, as shown at stage (74). The alternate words can be a string that has any number of words fewer than the extracted non-popular words. For example, if 5 non-popular words were extracted, then alternate word string of 4, 3, 2, or 1 word(s) can be generated. In the case of the three word string, "cool spring water", the following alternatives could perhaps have been generated:

- "refreshing water"
- "cool spring liquid"
- "refreshing liquid"
- "juice" (11)

While the method described above enables the extracted non-popular portion of the search string to be substituted with semantically similar words, it does not necessarily follow that the semantically similar words will be grammatically correct when substituted back into the original search string. For example, in the search string, "Where can I get cool spring water?", if the word "season" is identified as semantically similar to the two words "cool spring", substituting "season" into the original string yields the phrase, "Where can I get season water?" which clearly is not grammatically correct. In this case, the meaning is also not as originally intended because of the multiple meanings of the word "spring". In most cases, the applicant has found that where the substituted words yield a sentence that is grammatically incorrect, the meaning of the alternative string is different from the intended meaning of the original string, but where the substituted words yield a sentence that is grammatically correct, the meaning is generally consistent with the original meaning.

To overcome the problem of grammatically incorrect alternative search strings being generated, the invention includes additional steps by means of which grammatically incorrect alternative strings can be excluded. To do this, the substituted words are first substituted back into the original search string at stage (76). Then, at stage (78), each substituted word is analyzed within the original string to see whether the words preceding it and following it are words that are associated with the substituted word by a predefined degree. This is done by looking up the word in the word relationship database and checking whether the word following it appears within the list of row fields with more than a predetermined frequency. Using the reverse signature of that word, a check is also made to see whether the word preceding it appears within the list of row fields with more than a predetermined frequency. Only if both the preceding and following words appear within the row fields with more than a predetermined frequency is the word regarded as fitting grammatically within the string, otherwise they are rejected at stage (80).

For example, in the case of the alternative string, “Where can I get season water”, it is very unlikely that “get” will appear within the list of words that commonly precede “season” or that “water” will appear within the list of words that commonly follow “season”. This alternative string will therefore be rejected as grammatically incorrect.

If the word “fresh” is identified as semantically similar to “cool spring”, the string, “Where can I get fresh water?” would be checked for grammatical correctness by seeing whether the word “get” commonly precedes “fresh” and whether “water” commonly follows “fresh”. In both cases, the answer will be in the affirmative and, at stage (82), the string “Where can I get fresh water?” will be identified as an alternative string for “Where can I get cool spring water?”.

Once multiple alternative strings have been generated, they can simultaneously or in very rapid succession be input into a search engine as shown in Figure 1 and the results compared. The documents or web pages that are found to be relevant in the results of numerous alternative strings can then be identified as more relevant than those documents or web pages which are only found to be relevant in the results of one or two alternative search strings. The most relevant documents or web pages are then presented to the user first.

It will be appreciated that from the perspective of the user of a search engine the invention described above is completely hidden and is carried out in the background. The user interacts with the search engine in exactly the same way as before – by typing in a search string – and the search engine generates the alternative search strings and identifies the most relevant documents to present to the user.

The applicant has found that the invention leads to a marked improvement in the quality of the results that are presented to a user. Irrelevant search results are excluded far more often than with existing search engines and complex sentence structures can be handled with more precision. Because multiple alternative search strings are generated based on the search string, the applicant has found that it is no longer necessary to substitute different words or attempt to re-write search strings with different sentence structures in an attempt to locate relevant results. This leads to increased user satisfaction and quicker location of relevant search results.

The system of the invention requires no human input to categorize and index content, not does it have to be programmed with complex morphological or grammatical rules or built-in dictionaries. The invention provides a completely autonomous and extremely scalable system that is able to build a contextual language model of any contextual language so that search strings can be interpreted more accurately by search engines, so as to deliver more relevant

and targeted search results without the need to categorize or index existing content.

While it is envisaged that the invention may be applied in web based search engines, it can also be applied in the enterprise search market where companies search their own internal documents and information.

The above description is illustrative and is not restrictive. Many variations of the invention will become apparent to those skilled in the art upon review of the disclosure. The scope of the invention should, therefore, be determined not with reference to the above description, but instead should be determined with reference to the claims along with their full scope or equivalents.

Any of the software components or functions described in this specification may be implemented as software code to be executed by a processor using any suitable computer language such as, for example, Java, C++ or Perl using, for example, conventional or object-oriented techniques. The software code may be stored as a series of instructions, or commands on a computer readable medium, such as a random access memory (RAM), a read only memory (ROM), a magnetic medium such as a hard-drive, or an optical medium such as a CD-ROM. Any such computer readable medium may reside on or within a single computational apparatus, and may be present on or within different computational apparatuses within a system or network.

**CLAIMS:**

1. A method for processing an input search string and building multiple alternative search strings, the method comprising:

extracting text from a multitude of electronically accessible documents or web pages, the text including words;

forming a word relationship database in which each unique word identified in the text is stored in the database and is associated with a number of fields which each represent other words which were found to occur adjacent to that word in the text, each field also including a frequency sub-field which indicates how frequently that other word was found to occur adjacent to the associated word;

processing the word relationship database so as to determine a forward signature and reverse signature for each word, the forward signature including a ranked list of the words that were found to come after that word in the text, and the reverse signature including a ranked list of the words that were found to come before that word in the text;

combining the forward and reverse signatures of each word to form an ambidextrous signature for each word, and storing the ambidextrous signatures in a signature database;

optionally removing popular words from the input search string, being those words with a total frequency higher than a predetermined threshold;

linking the remaining words of the input search string into two-word groups from left to right with the second word of any preceding two-word group forming the first word of the next two-word group;

for each two-word group, carrying out the following steps:

querying the word relationship database to determine the reverse signature of the first word and the forward signature of the second word;

combining the forward and the reverse signatures obtained in the previous step into an ambidextrous signature representative of that two-word group;

comparing the ambidextrous signature of the previous step with the signature database to find the closest matches;

identifying the words in the signature database with the closest signature matches as alternative words that are semantically similar to the two-word group;

substituting one or more of the two-word groups in the input search string with the identified alternative words;

analyzing whether each substituted word fits grammatically into the input search string by querying the word relationship database to see whether the word preceding and the word following the substituted word in the input search string are words that are associated with the substituted word to a predefined extent; and

if the substituted word or words do fit grammatically, identifying the string with the substituted words as an alternative search string that is both semantically similar to the original search string and grammatically correct.

2. The method as claimed in claim 1, wherein the text is extracted from the web pages or documents by autonomous web crawling programs.

3. The method as claimed in claim 1 or 2, wherein the word relationship database is continually updated as more and more text is extracted, and the signature database is periodically rebuilt.

4. The method as claimed in any of the preceding claims, wherein the method includes an additional step of, immediately after extracting text and before forming a word relationship database, parsing the text into sentence portions which start and end with sentence delimiters.

5. The method as claimed in any of the preceding claims, wherein techniques are employed that keep the growth of the word relationship database in check.

6. The method as claimed in claim 5, wherein a technique employed is that the value of the frequency sub-fields are gradually reduced so that only those words that are frequently incremented will develop large frequencies.

7. The method as claimed in claim 5, wherein the technique employed is that words with low frequency fields are periodically discarded.

8. The method as claimed in any of the preceding claims, wherein the ambidextrous signature of the two-word group is matched with ambidextrous signatures in the signature database by calculating a matching score that looks for matches between the fields of the signatures and applies decreasing weighting factors to fields that are associated with lower frequency sub-fields.

9. The method as claimed in any of the preceding claims, wherein the method includes the steps of inputting the multiple alternate search strings into a search engine simultaneously or in rapid succession and comparing the results of each separate search so as to rank the overall results and present those results which were obtained in the greatest number of separate searches as the most relevant search results.

10. The method as claimed in any of the preceding claims, wherein the language of the text is identified so that separate word relationship databases and signature databases can be built for each separate language.

11. A system for processing an input search string and building multiple alternative search strings, comprising:

a processor in the form of a server which is able to access a multitude of web pages or other documents through the Internet and extract text, the text including words;

a word relationship database coupled to the processor and having each unique word identified in the text stored thereon, each unique word in the word relationship database being associated with a number of fields which each represent other words which were found to occur adjacent to that word in the text, each field also including a frequency sub-field which indicates how frequently that other word was found to occur adjacent to the associated word;

a signature database coupled to the server, the signature database being formed by the server processing the word relationship database so as to determine a forward signature and reverse signature for each word, the forward signature including a ranked list of the words that were found to come after that word in the text, and the reverse signature including a ranked list of the words that were found to come before that word in the text, and combining the forward and reverse signatures of each word to form an ambidextrous signature for each word that is stored in the signature database;

and computer software stored on the processor and configured to enable the processor to carry out the following:

optionally removing popular words from the input search string, being those words with a total frequency higher than a predetermined threshold;

linking the remaining words of the input search string into two-word groups from left to right with the second word of any preceding two-word group forming the first word of the next two-word group;

for each two-word group, carrying out the following:

querying the word relationship database to determine the reverse signature of the first word and the forward signature of the second word;

combining the forward and the reverse signatures obtained in the previous step into an ambidextrous signature representative of that two-word group;

comparing the ambidextrous signature of the previous step with the signature database to find the closest matches;

identifying the words in the signature database with the closest signature matches as alternative words that are semantically similar to the two-word group;

substituting one or more of the two-word groups in the input search string with the identified alternative words;

analyzing whether each substituted word fits grammatically into the input search string by querying the word relationship database to see whether the word preceding and the word following the substituted word in the input search string are words that are associated with the substituted word to a predefined extent; and

if the substituted word or words do fit grammatically, identifying the string with the substituted words as an alternative search string that is both semantically similar to the original search string and grammatically correct.

12. The system as claimed in claim 11, wherein the server is configured to input the multiple alternative search strings into a search engine simultaneously or in rapid succession, and to compare the results of each separate search so as to rank the overall results and present those results which are obtained in the greatest number of separate searches as the most relevant search results.

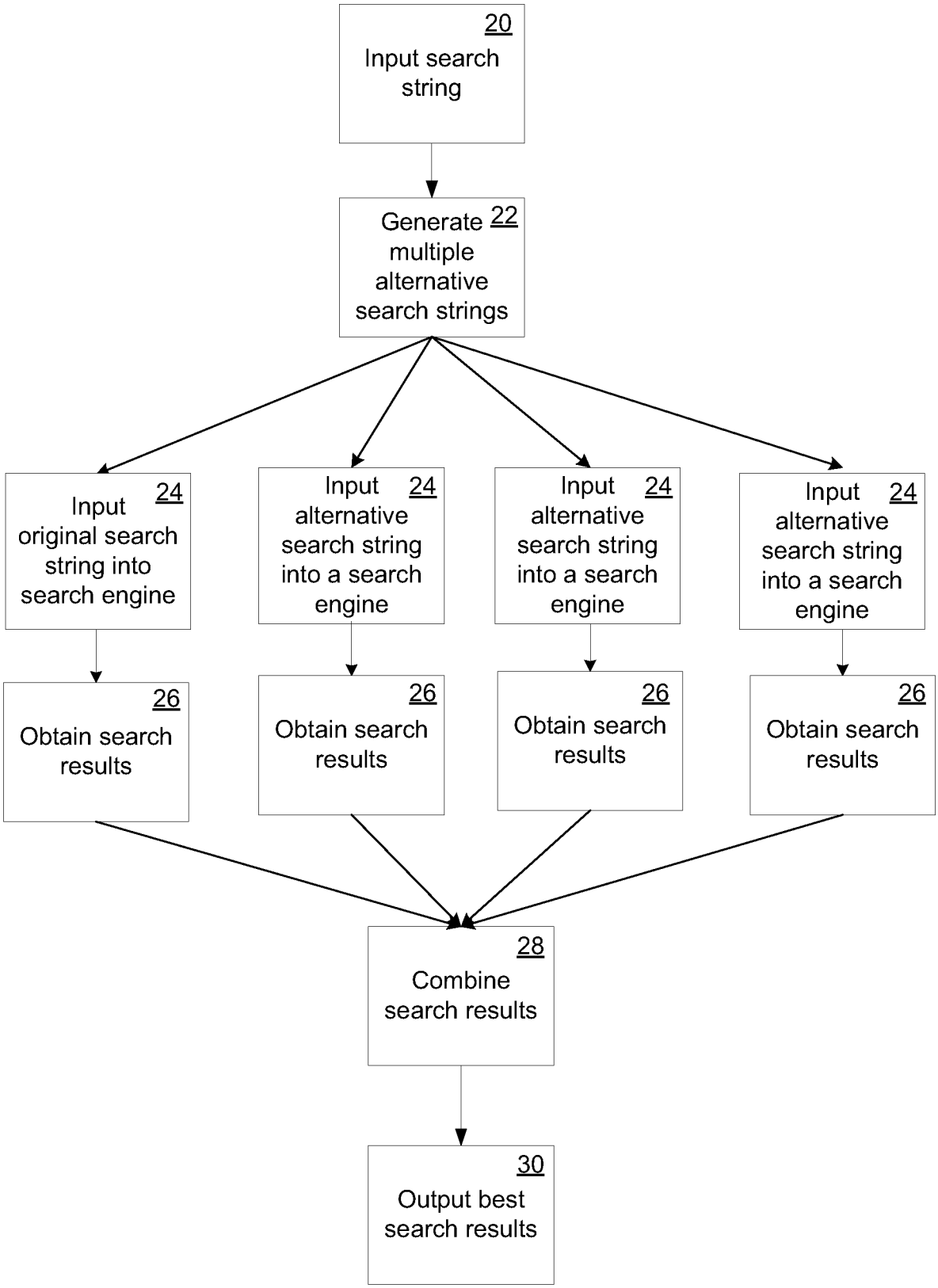
20
Input search string
Generate 22 multiple alternative search strings
Input original search string into search engine 24
Input alternative search string into a search engine 24
Input alternative search string into a search engine 24
Input alternative search string into a search engine 24
Obtain search results 26
Obtain search results 26
Obtain search results 26
Obtain search results 26
28
Combine search results
30
Output best search results

Figure 1

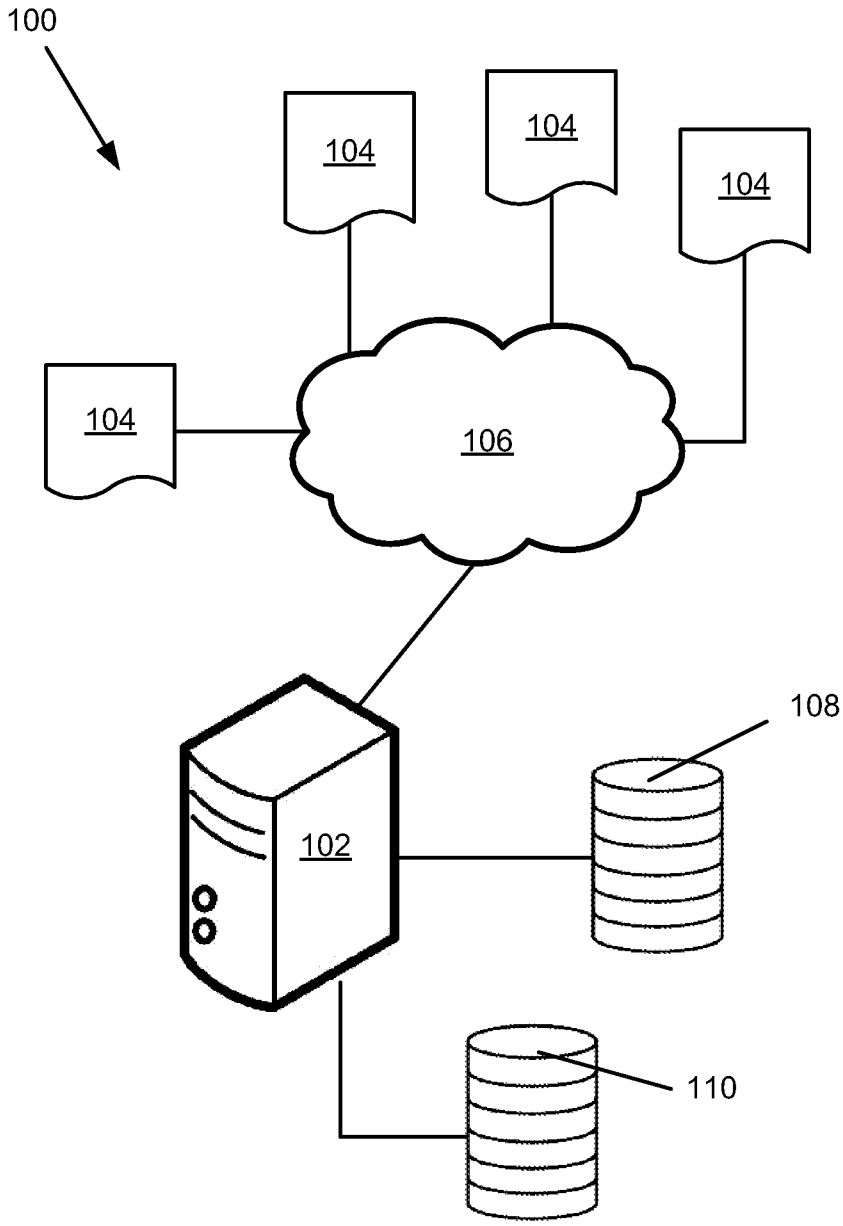


Figure 2

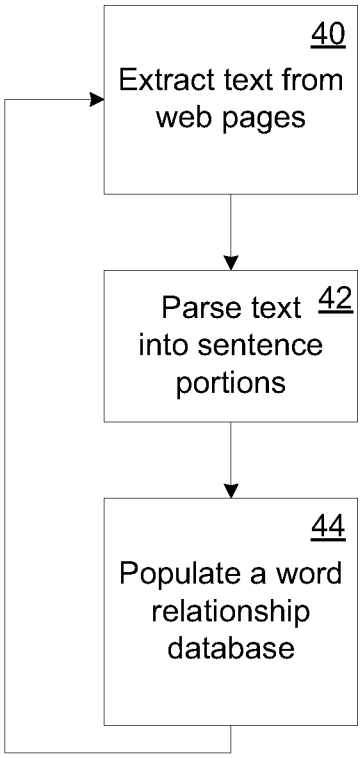


Figure 3

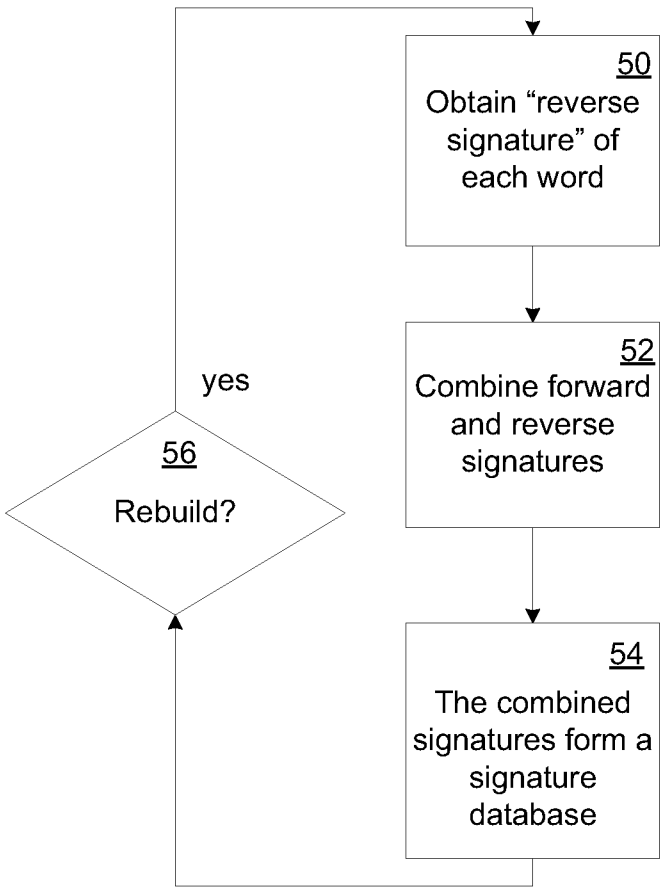


Figure 4

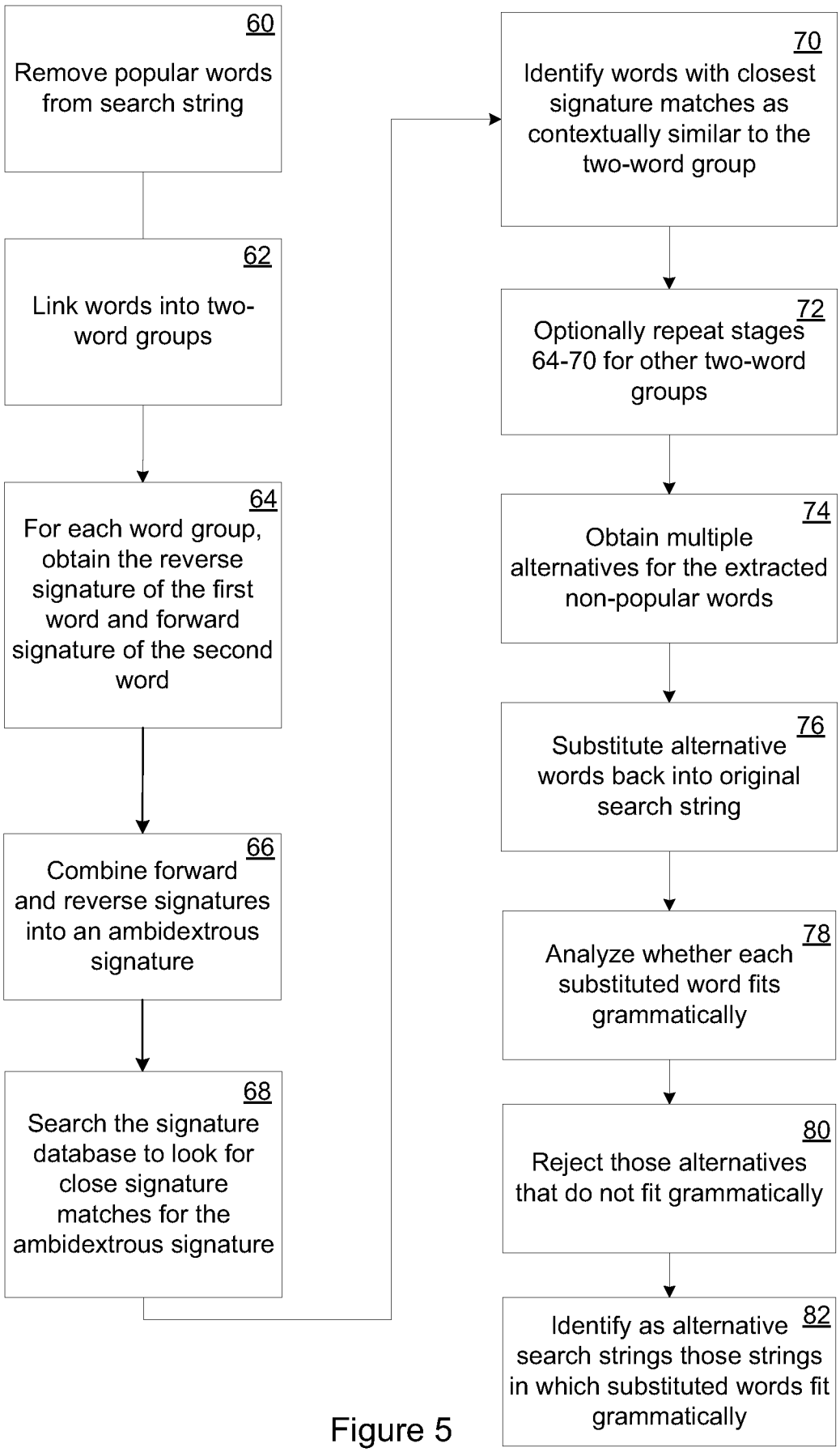
60
Remove popular words from search string
62
Link words into two-word groups
64
For each word group, obtain the reverse signature of the first word and forward signature of the second word
66
Combine forward and reverse signatures into an ambidextrous signature
68
Search the signature database to look for close signature matches for the ambidextrous signature
70
Identify words with closest signature matches as contextually similar to the two-word group
72
Optionally repeat stages 64-70 for other two-word groups
74
Obtain multiple alternatives for the extracted non-popular words
76
Substitute alternative words back into original search string
78
Analyze whether each substituted word fits grammatically
80
Reject those alternatives that do not fit grammatically
82
Identify as alternative search strings those strings in which substituted words fit grammatically

Figure 5

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/IB2012/051870

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 17/27 (2012.01)
USPC - 707/767
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC(8) - G06F 17/20, 17/27, 17/28 (2012.01)
USPC - 707/739, 750, 759, 767

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)
PatBase, Orbit, Google Patents, Google, Google Scholar

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| A | US 2003/0171910 A1 (ABIR) 11 September 2003 (11.09.2003) entire document | 1-3, 11-12 |
| A | US 2006/0253427 A1 (WU et al) 09 November 2006 (09.11.2006) entire document | 1-3, 11-12 |
| A | US 5,675,819 A (SCHUETZE) 07 October 1997 (07.10.1997) entire document | 1-3, 11-12 |
| A | US 6,850,937 B1 (HISAMITSU et al) 01 February 2005 (01.02.2005) entire document | 1-3, 11-12 |

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

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 21 August 2012 | 28 AUG 2012 |
| Name and mailing address of the ISA/US<br>Mail Stop PCT, Attn: ISA/US, Commissioner for Patents<br>P.O. Box 1450, Alexandria, Virginia 22313-1450<br>Facsimile No. 571-273-3201 | Authorized officer:<br>Blaine R. Copenheaver<br>PCT Helpdesk: 571-272-4300<br>PCT OSP: 571-272-7774 |

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/IB2012/051870

**Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)**

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☒ Claims Nos.: 4-10
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

**Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)**

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

**Remark on Protest**

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.