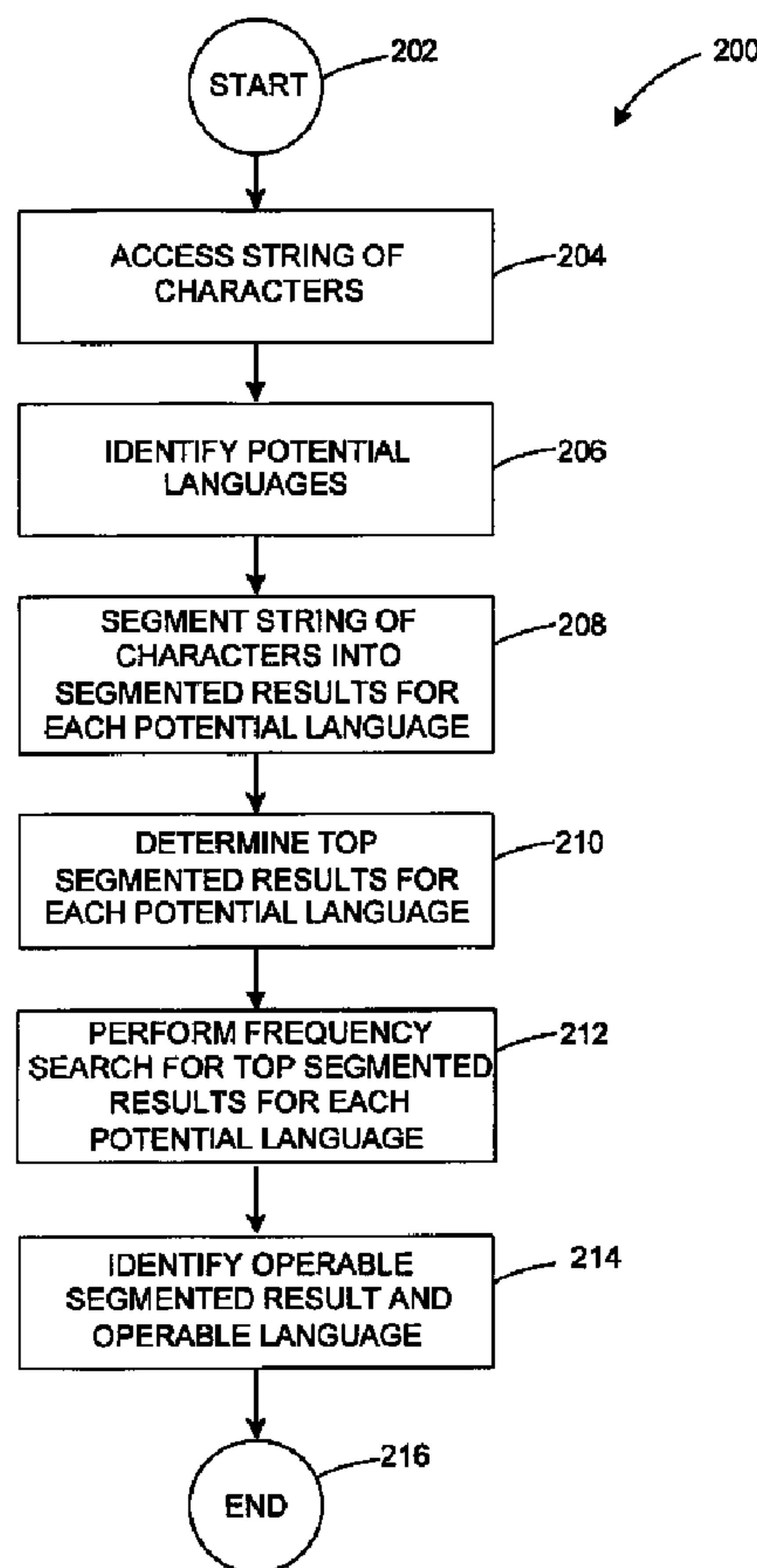




(86) Date de dépôt PCT/PCT Filing Date: 2005/09/28
 (87) Date publication PCT/PCT Publication Date: 2006/04/13
 (45) Date de délivrance/Issue Date: 2013/05/07
 (85) Entrée phase nationale/National Entry: 2007/03/30
 (86) N° demande PCT/PCT Application No.: US 2005/035010
 (87) N° publication PCT/PCT Publication No.: 2006/039398
 (30) Priorité/Priority: 2004/09/30 (US10/955,660)

(51) Cl.Int./Int.Cl. *G06F 17/27* (2006.01)
 (72) Inventeurs/Inventors:
 ELBAZ, GILAD ISRAEL, US;
 MANDELSON, JACOB LEON, US
 (73) Propriétaire/Owner:
 GOOGLE INC., US
 (74) Agent: SMART & BIGGAR

(54) Titre : PROCÉDES ET SYSTEMES DE SELECTION D'UN LANGAGE DE SEGMENTATION DE TEXTE
 (54) Title: METHODS AND SYSTEMS FOR SELECTING A LANGUAGE FOR TEXT SEGMENTATION



(57) Abrégé/Abstract:

Methods and systems for selecting a language for text segmentation are disclosed. In one embodiment, at least a first candidate language and a second candidate language associated with a string of characters are identified, at least a first segmented result

(57) **Abrégé(suite)/Abstract(continued):**

associated with the first candidate language and a second segmented result associated with the second candidate language are determined, a first frequency of occurrence for the first segmented result and a second frequency of occurrence for the second segmented result are determined, and an operable language is identified from the first candidate language and the second candidate language based at least in part on the first frequency of occurrence and the second frequency of occurrence.

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

(19) World Intellectual Property Organization
International Bureau(43) International Publication Date
13 April 2006 (13.04.2006)

PCT

(10) International Publication Number
WO 2006/039398 A3(51) International Patent Classification:
G06F 17/27 (2006.01)

(74) Agents: GARDNER, Steven, J. et al.; Kilpatrick Stockton LLP, 1001 West Fourth Street, Winston-Salem, North Carolina 27101-2400 (US).

(21) International Application Number:

PCT/US2005/035010

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

(22) International Filing Date:

29 September 2005 (29.09.2005)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

10/955,660 30 September 2004 (30.09.2004) US

(71) Applicant (for all designated States except US):
GOOGLE INC. [US/US]; 1600 Amphitheatre Parkway, Mountain View, California 94043 (US).

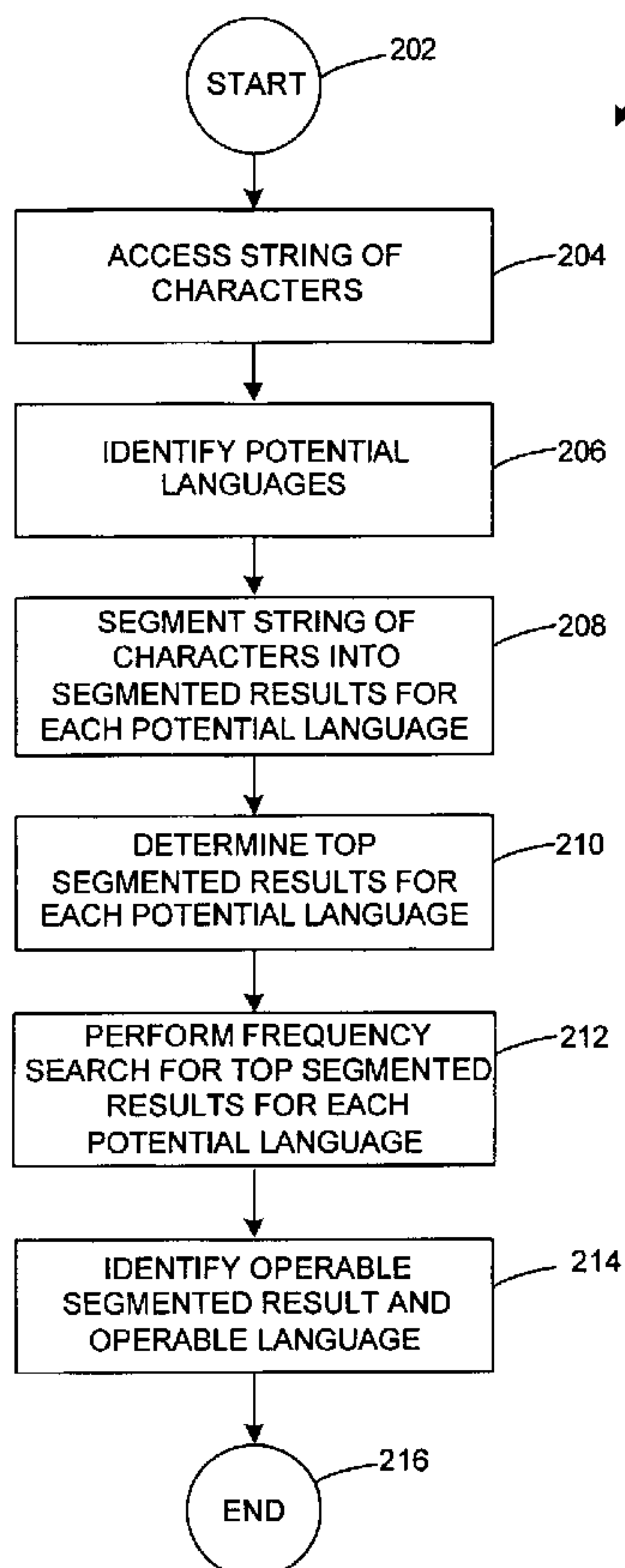
(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

(71) Applicants and

(72) Inventors: ELBAZ, Gilad, Israel [US/US]; 2800 Neilson Way, #810, Santa Monica, California 90405 (US). MANDELSON, Jacob, Leon [US/US]; 400 E. California Boulevard, Apt. #26, Pasadena, California 91106 (US).

[Continued on next page]

(54) Title: METHODS AND SYSTEMS FOR SELECTING A LANGUAGE FOR TEXT SEGMENTATION



(57) Abstract: Methods and systems for selecting a language for text segmentation are disclosed. In one embodiment, at least a first candidate language and a second candidate language associated with a string of characters are identified, at least a first segmented result associated with the first candidate language and a second segmented result associated with the second candidate language are determined, a first frequency of occurrence for the first segmented result and a second frequency of occurrence for the second segmented result are determined, and an operable language is identified from the first candidate language and the second candidate language based at least in part on the first frequency of occurrence and the second frequency of occurrence.

WO 2006/039398 A3

WO 2006/039398 A3



Published:

- *with international search report*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

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

(88) Date of publication of the international search report:

6 July 2006

60412-3762

METHODS AND SYSTEMS FOR SELECTING A LANGUAGE FOR TEXT SEGMENTATION

FIELD OF THE INVENTION

The present invention relates generally to text segmentation and,
5 more particularly, to selecting a language for text segmentation.

BACKGROUND OF THE INVENTION

Text processing methods and systems exist that attempt to interpret
data representing text. Text processing is made more difficult when text
comprising a string of characters is received that has no breaks indicating words
10 or other tokens. When processing such strings of characters using existing
methods and systems, the characters can be segmented into tokens in order to
interpret the string. Tokens can be words, acronyms, abbreviations, proper
names, geographical names, stock market ticker symbols, or other tokens.
Generally, a string of characters may be segmented into multiple combinations of
15 segmented strings of characters using existing methods and systems. Selecting
the correct language to use when segmenting the text can produce more
meaningful results.

SUMMARY

According to one aspect of the present invention, there is provided a
20 computer-implemented method, comprising: receiving from a user of a computing
device, at a computer server system, a request for information about one or more
internet-accessible documents, the request having a string of characters;
identifying, using the computer server system, at least a first candidate language
and a second candidate language associated with the request; determining at
25 least a first segmented result associated with the first candidate language from the
string of characters and a second segmented result associated with the second
candidate language from the string of characters; determining a first frequency of
occurrence for the first segmented result in a group of articles that are associated
by the system with the first language and a second frequency of occurrence for
30 the second segmented result in a group of articles that are associated by the

60412-3762

system with the second language; identifying, with the computer server system, an operable language from the first candidate language and the second candidate language based at least in part on the first frequency of occurrence and the second frequency of occurrence; selecting, for use by the user of the computing device, electronic content in the identified operable language from among available content in multiple languages; and providing the selected content to the computing device so that the selected content is arranged to be displayed to the user in the identified operable language and accompanying the requested one or more internet-accessible documents.

10 According to another aspect of the present invention, there is provided a tangible and non-transitory computer-readable medium containing program code executable on a computer, comprising: program code for receiving from a user of a computing device, at a computer server system, a request to receive one or more internet-accessible documents, the request having a string of characters; program code for identifying at least a first candidate language and a second candidate language associated with a string of characters received in the request; program code for determining at least a first segmented result associated with the first candidate language from the string of characters and a second segmented result associated with the second candidate language from the string of characters; program code for determining a first frequency of occurrence for the first segmented result in a group of articles that are associated by the computer server system with the first language and a second frequency of occurrence for the second segmented result in a group of articles that are associated by the computer server system with the second language; program code for identifying an operable language from the first candidate language and the second candidate language based at least in part on the first frequency of occurrence and the second frequency of occurrence; program code for selecting, for use by the user of the computing device, electronic content in the identified operable language from among available content in multiple languages; and program code for providing the selected content to the computing device so that the selected content is arranged to be displayed to the user in the identified operable language with the requested one or more internet-accessible documents.

60412-3762

Embodiments of the present invention comprise methods and systems for selecting a language for text segmentation. One embodiment of the present invention comprises identifying at least a first candidate language and a second candidate language associated with a string of characters, determining at least a first segmented result associated with the first candidate language from the string of characters and a second segmented result associated with the second candidate language from the string of characters, determining a first frequency of occurrence for the first segmented result and a second frequency of occurrence for the second segmented result, and identifying an operable language from the first candidate language and the second candidate language based at least in part on the first frequency of occurrence and the second frequency of occurrence.

This exemplary embodiment is mentioned not to limit or define the invention, but to provide an example of an embodiment of the invention to aid understanding thereof. Exemplary embodiments are discussed in the Detailed Description, and further description of the invention is provided there. Advantages offered by the various embodiments of the present invention may
5 be further understood by examining this specification.

BRIEF DESCRIPTION OF THE DRAWINGS

These and other features, aspects, and advantages of the present invention are better understood when the following Detailed Description is read with reference to the accompanying drawings, wherein:

10 Figure 1 illustrates a diagram of a system in accordance with one embodiment of the present invention; and

Figure 2 illustrates a flow diagram of one embodiment of a method carried out by the present invention.

DETAILED DESCRIPTION

15 *Introduction*

Embodiments of the present invention comprise methods and systems for selecting a language for text segmentation. There are multiple embodiments of the present invention. By way of introduction and example, one exemplary embodiment of the present invention provides a method for improving segmenting a string of characters, such as a domain name, into multiple
20 tokens or words by selecting the correct language for the string of characters. A number of potential or candidate languages for the string of characters may be selected based on a variety of signals, such as linguistics associated with the string of characters, an IP address associated with the user, a character set used for the string of characters, browser settings of a browser application program associated with the user, and any top-level domain associated with the string
25 of characters. The string of characters may be segmented into many segmented results using each candidate language. Each segmented result can be a particular combination of words or other tokens. For example, the string of characters “usedrugs” can be segmented into the following segmented results for the English language: “used rugs”, “use drugs”, “us ed rugs”,

etc. From this number of segmented results for each candidate language, an operable segmented result and an operable language can be identified based on the number of documents or search queries in the operable language that contain the operable segmented result.

For example, segmented results with the highest probability of being the best operable segmented result can be selected for each candidate language. A search engine can determine the number of documents or search queries containing a selected segmented result and can do this for each selected segmented result in each candidate language. In one embodiment, the segmented result that occurs with the greatest frequency in documents or search queries in the particular language may be identified as the best operable segmented result. The language associated with the best operable segmented result may be identified as the best operable language. Language signals used to determine the candidate languages can also be used to select the operable language. The operable segmented result and operable language can be used for a variety of functions, including selecting advertisements based on the language and result.

This introduction is given to introduce the reader to the general subject matter of the application. By no means is the invention limited to such subject matter. Exemplary embodiments are described below.

System Architecture

Various systems in accordance with the present invention may be constructed. Figure 1 is a diagram illustrating an exemplary system in which exemplary embodiments of the present invention may operate. The present invention may operate, and be embodied in, other systems as well.

Referring now to the drawings in which like numerals indicate like elements throughout the several figures, Figure 1 is a diagram illustrating an exemplary environment for implementation of an embodiment of the present invention. The system 100 shown in Figure 1 comprises multiple client devices 102a-n in communication with a server device 104 and a server device 150 over a network 106. In one embodiment, the network 106 shown comprises the Internet. In other embodiments, other networks, such as an intranet, WAN, or LAN may be used. Moreover, methods according to the present invention may operate within a single computer.

The client devices 102a-n shown in Figure 1 each comprise a computer-readable medium, such as a random access memory (RAM) 108 coupled to a processor 110. The processor 110 executes computer-executable program instructions stored in memory 108. Such processors may comprise a microprocessor, an ASIC, and state machines. Such processors
5 comprise, or may be in communication with, media, for example computer-readable media, which stores instructions that, when executed by the processor, cause the processor to perform the steps described herein. Embodiments of computer-readable media include, but are not limited to, an electronic, optical, magnetic, or other storage or transmission device capable of providing a processor, such as the processor 110 of client 102a, with computer-readable
10 instructions. Other examples of suitable media include, but are not limited to, a floppy disk, CD-ROM, DVD, magnetic disk, memory chip, ROM, RAM, an ASIC, a configured processor, all optical media, all magnetic tape or other magnetic media, or any other suitable medium from which a computer processor can read instructions. Also, various other forms of computer-readable media may transmit or carry instructions to a computer, including a router, private or
15 public network, or other transmission device or channel, both wired and wireless. The instructions may comprise code from any suitable computer-programming language, including, for example, C, C++, C#, Visual Basic, Java, Python, Perl, and JavaScript.

Client devices 102a-n may also comprise a number of external or internal devices such as a mouse, a CD-ROM, DVD, a keyboard, a display, or other input or output devices. Examples
20 of client devices 102a-n are personal computers, digital assistants, personal digital assistants, cellular phones, mobile phones, smart phones, pagers, digital tablets, laptop computers, Internet appliances, and other processor-based devices. In general, a client device 102a may be any suitable type of processor-based platform that is connected to a network 106 and that interacts with one or more application programs. Client devices 102a-n may operate on any operating
25 system capable of supporting a browser or browser-enabled application, such as Microsoft® Windows® or Linux. The client devices 102a-n shown include, for example, personal computers executing a browser application program such as Microsoft Corporation's Internet Explorer™, Netscape Communication Corporation's Netscape Navigator™, and Apple Computer, Inc.'s Safari™.

30 Through the client devices 102a-n, users 112a-n can communicate over the network 106 with each other and with other systems and devices coupled to the network 106. As shown in

Figure 1, a server device 104 and a server device 150 are also coupled to the network 106.

The server device 104 can comprise a server executing a segmentation engine application program and server device 150 can comprise a server executing a search engine application program. Similar to the client devices 102a-n, the server device 104 and server device 150 shown in Figure 1 comprise a processor 116 coupled to a computer-readable memory 118 and a processor 152 coupled to a computer-readable memory 154, respectively. Server devices 104 and 150, depicted as single computer systems, may be implemented as a network of computer processors. Examples of server devices 104, 150 are servers, mainframe computers, networked computers, a processor-based device, and similar types of systems and devices. Client processor 110 and the server processors 116, 152 can be any of a number of computer processors, as described above, such as processors from Intel Corporation of Santa Clara, California and Motorola Corporation of Schaumburg, Illinois.

Memory 118 contains a segmentation application program, also known as a segmentation engine 120. The server device 104, or related device, can access the network 106 to receive strings of characters from other devices or systems connected to the network 106. Characters can include, for example, marks or symbols used in a writing system, including data representing a character, such as ASCII, Unicode, ISO 8859-1, Shift-JIS, and EBCDIC or any other suitable character set. In one embodiment, the segmentation engine 120 can receive a string of characters, such as a domain name, from a server device on the network 106 when a user 112a attempts to direct a web browser application to a domain name that is not active.

In one embodiment, the segmentation engine 120 identifies candidate languages for the string of characters, segments the string of characters into potential combinations of tokens for each candidate language, and selects a particular language and combination to associate with the string of characters. A token can comprise a word, a proper name, a geographic name, an abbreviation, an acronym, a stock market ticker symbol, or other tokens. The segmentation engine 120 can include a segmentation processor 122, a frequency processor 124, and a language processor 126. In the embodiment shown in Figure 1, each comprises computer code residing in the memory 118.

The language processor 126 can identify a candidate language or languages for the string of characters. In one embodiment, the language processor 126 can use signals to identify a

number of candidate languages for the string of characters. For example, the language processor can use linguistics, the IP address of the user, a character set used for the string of characters, browser settings of a browser application program associated with the user, and a top-level domain associated with the string of characters to determine the candidate languages for the
5 string of characters.

The segmentation processor 122 can determine a list of potential combinations of tokens or segmented results from the string of characters for each candidate language. In one embodiment, the token processor 124 determines a probability for each segmented result in the list and selects the top segmented results for each language based on the probability. The
10 probability for a segmented result can be based on frequency values associated with the individual tokens in the result. In one embodiment, the unsegmented string of characters may be included as a segmented result.

The frequency processor 124 can perform a frequency search or cause one to be performed on the top-selected segmented results of each candidate language. The frequency
15 processor 124 can include a spell-checking functionality or can call a spell-checking functionality residing elsewhere to perform a spell check on the selected segmented results. Any spell corrected results can be included in the frequency search. In one embodiment, the frequency processor sends the selected segmented results to the server device 150 to perform a frequency search on the selected segmented results. A frequency search can determine the
20 frequency of occurrence for each particular segmented result as described below. Based on the frequency search a best or operable segmented result can be identified by the segmentation processor 122. The language associated with the operable result can be identified by the segmentation processor 122 as the operable language for the string of characters. In one embodiment, the operable segmented result and the operable language can be sent to an
25 advertising server that can select advertisements targeted based on one or both of the operable language and the segmented result. Other functions and characteristics of the segmentation processor 122, the frequency processor 124, and the language processor 126 are further described below.

Server device 104 also provides access to other storage elements, such as a token storage
30 element, in the example shown a token database 120. The token database 120 can be used to

store tokens and frequency information associated with each token. The token database 120 can also store the language or languages associated with each token. Data storage elements may include any one or combination of methods for storing data, including without limitation, arrays, hash tables, lists, and pairs. The server device 104 can access other similar types of data storage
5 devices.

The server device 150 can include a server executing a search engine application program, such as the Google™ search engine. In other embodiments, the server device 150 can comprise a related information server or an advertising server. In another embodiment, there can be multiple server devices 150.

10 Memory 154 contains the search engine application program, also known as a search engine 156. The search engine 156 can locate relevant information from the network 106 in response to a search query from a user 112a and can maintain a search log of search queries. The search engine 156 can also perform a frequency search in response to a frequency search request from the frequency processor 124. The search engine 156 can provide a search result set to a
15 user 112a or frequency information to the segmentation engine 120 via the network 106.

In one embodiment, the server device 150, or related device, has previously performed a crawl of the network 106 to locate articles, such as web pages, stored at other devices or systems coupled to the network 106. Articles include, for example, documents, emails, instant messenger messages, database entries, web pages of various formats, such as HTML, XML, XHTML,
20 Portable Document Format (PDF) files, and media files, such as image files, audio files, and video files, or any other documents or groups of documents or information of any suitable type whatsoever. An indexer 158 can be used to index the articles in memory 154 or on another data storage device, such as an index 160. The index may also include the language or languages associated with each article. In one embodiment, there are multiple indexes each containing a
25 portion of the total articles indexed. It should be appreciated that other suitable methods for indexing articles in lieu of or in combination with crawling may be used, such as manual submission.

The search engine 156 can perform a frequency search in a number of suitable ways. In one embodiment, the search engine 156 can perform a web search using each top selected
30 segmented result as a search query and can search for articles containing the search query in the

candidate language of the segmented result. In this embodiment, a frequency search result set can be generated and can comprise one or more article identifiers. An article identifier may be, for example, a Uniform Resource Locator (URL), a file name, a link, an icon, a path for a local file, or anything else that identifies an article. In one embodiment, an article identifier can
5 comprise a URL associated with an article. The frequency processor 124 can use the number of article identifiers in each frequency search result set as a representation of the number of occurrences of the respective segmented result.

In another embodiment, the frequency processor 124 can interface directly with the indexer 158. The indexer 158 can determine, for each top selected segmented result, the number
10 of articles in the associated candidate language in which the segmented result appears. This information can be sent to the frequency processor 124. In still another embodiment, the search engine 156 and/or the frequency processor 124 can determine, for each selected segmented result, the number of occurrences in search queries in the associated candidate language from the search log and the frequency processor 124 can determine a frequency of occurrence based on
15 this search log information. In one embodiment, the number of articles or search queries in a frequency search associated with a segmented result may be normalized based on the total number of articles or search queries in the associated language.

It should be noted that the present invention may comprise systems having different architecture than that which is shown in Figure 1. For example, in some systems according to
20 the present invention, server device 104 may comprise a single physical or logical server. The system 100 shown in Figure 1 is merely exemplary, and is used to help explain method illustrated in Figure 2.

Process

Various methods in accordance with embodiments of the present invention may be
25 carried out. One exemplary method according to the present invention comprises identifying at least a first candidate language and a second candidate language associated with a string of characters, determining at least a first segmented result associated with the first candidate language from the string of characters and a second segmented result associated with the second candidate language from the string of characters, determining a first frequency of occurrence for
30 the first segmented result and a second frequency of occurrence for the second segmented result,

and identifying an operable language from the first candidate language and the second candidate language based at least in part on the first frequency of occurrence and the second frequency of occurrence. More than two candidate languages can be identified and more than two segmented results can be determined. For example, three candidate languages can be identified and four
5 segmented results can be determined for each candidate language.

The operable language can be identified based in part on identifying an operable segmented result from the first segmented result and the second segmented result based at least in part on the first frequency of occurrence and the second frequency of occurrence. A first candidate language and a second candidate language may be identified based in part on one or
10 more language signals. The language signals can comprise at least one of linguistics associated with the string of characters, an IP address of a user associated with the string of characters, a character set used for the string of characters, browser settings of a browser application program associated with a user, and a top-level domain associated with the string of characters. In one embodiment, identifying the operable language may be based at least in part on language signals.

15 In one embodiment, identifying the operable language from the first candidate language and the second candidate language based at least in part on the first frequency of occurrence and the second frequency of occurrence can comprise selecting the first candidate language if the first frequency of occurrence is greater than the second frequency of occurrence. The string of characters can comprise a domain name. The first segmented result can comprise a first
20 combination of tokens and the second segmented result comprises a second combination of tokens.

In one embodiment, determining the first frequency of occurrence for the first segmented result can comprise determining a number of articles in the first candidate language containing the first segmented result and normalizing the number of articles based on a total number of
25 articles in the first candidate language and determining the number of articles in the first language containing the first segmented result can comprise determining a number of article identifiers in a search result set generated in response to a search query comprising the first segmented result.

In one embodiment, determining the number of articles in the first language containing
30 the first segmented result can comprise accessing an index of articles. In another embodiment,

determining the first frequency of occurrence can comprise determining a number of occurrences of the first segmented result in a plurality of search queries in the first candidate language and normalizing the number of occurrences based on a total number of search queries in the first candidate language.

5 The method may also comprise selecting an article based at least in part on the operable language or the operable segmented result (or both) and the article can comprise an advertisement. In one embodiment, determining the first segmented result can comprise determining a plurality of segmented results in the first candidate language from the string of characters, and identifying the first segmented result from the plurality of segmented results in
10 the first candidate language. Identifying the first segmented result can comprise calculating a probability value for each of the plurality of segmented results. A first probability value associated with the first segmented result can be based at least in part on a frequency of each token within the first segmented result.

 Another exemplary method comprises determining a first segmented result in a first
15 candidate language and a second segmented result in a second candidate language from a domain name, determining a first frequency of occurrence for the first segmented result in at least one of an article index, a text index, and a search result set, determining a second frequency of occurrence for the second segmented result, if the first frequency of occurrence is greater than the second frequency of occurrence, then selecting the first candidate language as an operable
20 language, if the second frequency of occurrence is greater than the first frequency of occurrence, then selecting the second candidate language as the operable language, selecting an advertisement based at least in part on the operable language, wherein the advertisement includes text in the operable language, and causing a display of the advertisement in association with a web page associated with the domain name.

25 Figure 2 illustrates an exemplary method 200 for selecting a language for text segmentation, in accordance with one embodiment of the invention. This exemplary method is provided by way of example, as there are a variety of ways to carry out methods according to the present invention. The method 200 shown in Figure 2 can be executed or otherwise performed by one or a combination of various systems. The method 200 is described below as carried out
30 by the system 100 shown in Figure 1 by way of example, and various elements of the system 100

are referenced in explaining the example method of Figure 2.

Referring to Fig. 2, in block 202, the example method begins. Block 202 is followed by block 204, in which a string of characters can be accessed by the segmentation engine 120. A string of characters can be received or accessed from a device connected to network 106, for example, or from another device. In one embodiment, the string of characters can be a domain name associated with an inactive or non-existent website received from an advertising server associated with the domain name.

Block 204 is followed by block 206, in which candidate languages for the string of characters are identified. In one embodiment, the language processor 126 can use one or more language signals to determine a number of candidate languages for the string of characters. For example, the language processor can identify, based on language signals, English, French and Spanish as the three candidate languages for the string of characters.

Some of the language signals used can be, for example, linguistics associated with the string of characters, the IP address of a user associated with the string of characters, the character set used for the string of characters, browser settings of a browser application program associated with the user associated with the string of characters, and a top-level domain associated with the string of characters. Linguistics can be used, for example, to determine if the structure or nature of the string of characters indicates that it is in a particular language. For example, certain languages have a tendency to start or end with a certain group of characters and use general patterns. The IP address of the user can indicate the location and country of the user. From the country information a language or languages associated with the country can be used as candidate languages. The character set of the string of characters can indicate a language or languages associated with the string of characters. For example, a Cyrillic character set can indicate Russian or some other Slavic language. The browser settings for a browser application program of a user associated with the character string can indicate a language and/or a character set associated with the character string. For example, the language and character set settings of the user's browser application program can be passed in an HTTP header along with the string of characters. A top-level domain associated with the string of characters can indicate a country. A top-level domain can be the highest level of hierarchy after the root. In a domain name, the top-level domain is the portion of the domain name that appears farthest to the right. For example,

for the domain name "usedrugs.co.uk", the top-level domain is ".uk" and can indicate the United Kingdom. The top-level domain ".ru" can indicate Russia. The country associated with the top-level domain can be used in determining a candidate language, such as "ru" indicates Russia, which indicates that the associated string of characters may be in the Russian language. Some
5 top-level domains may indicate more than one language. For example, ".ch" can indicate Switzerland and can indicate that the string of characters can be associated with French, German, or Italian. Other suitable signals and methods of identifying candidate languages for the string of characters may be used.

Block 206 is followed by block 208, in which a plurality of segmented results is
10 generated from the string of characters by segmenting the string of characters for each of the candidate languages. Segmenting the string of characters may include parsing the characters in the string into a plurality of combinations of tokens and can be performed by the segmentation processor 122. The segmentation processor 122 can develop a list of segmented results for each candidate language. Each segmented result can be a particular combination of tokens or a single
15 token. For example, the string of characters "assocomunicazioni" can be segmented in Italian into "asso comunicazioni" and other segmented results and can be segmented in French into "asso com uni cazioni" and other segmented results. In another example, the string of characters "maisonblanche" can be segmented in French into "maison blanche" and other segmented results and can be segmented in English into "mai son blanc he" and other segmented results. In
20 another example, the string of characters "usedrugs" can be segmented in English into the segmented results including "used rugs", "use drugs", "us ed rugs", "u sed rugs", "usedrugs", etc. Segmented results can also be generated for the other candidate languages, such as, in the example above, French and Spanish. The unsegmented string of characters may be included as a segmented result.

25 The segmentation processor 122 can utilize tokens from the token database 126 in the segmentation process. Various methods can be used to segment the string of characters, such as the segmentation techniques described in PCT International Patent Application No. PCT/US03/41609 entitled "Methods and Systems for Text Segmentation" filed December 30, 2003, which is hereby incorporated in its entirety by this reference.

30 Block 208 is followed by block 210, in which the top segmented results are determined

for each candidate language. The top segmented results can be determined by the segmentation processor 122 and can be the results with the highest probability of being the best or operable segmented result. In one embodiment, the segmented results can be ranked based on a probability value determined for each segmented result. In one embodiment, a probability value can be determined by summing frequency values associated with the individual tokens within each individual segmented result. In another embodiment, a probability value can be determined by a complex function involving summing the logarithms of the frequency values associated with the individual tokens within each individual segmented result. A number of the top ranking segmented results can then be selected. For example, the segmented results for each candidate language can be ranked and the top three results of each candidate language can be selected.

Block 210 is followed by block 212, in which a frequency search is performed for the top, selected segmented results for each candidate language. The frequency search can be performed by the frequency processor 124 in conjunction with the search engine 156. In one embodiment, the segmentation processor 122 can pass the selected segmented results to the frequency processor 124, which can determine the frequency of occurrence for each of the segmented results in a corpus of articles or search queries.

In one embodiment, the frequency processor 124 can determine the frequency of occurrence for the segmented results based on articles indexed by a search engine 156. In one embodiment, the frequency processor 124 can send the top selected segmented results to the search engine 156 via the network 106. The search engine 156 can perform a search for each of the segmented results in the indexed articles by using each segmented result as a search query. For example, the frequency processor 124 can send each segmented result for each candidate language surrounded by quotation marks to the search engine 156 as a search query, so that the search engine 156 performs the search on the exact segmented phrase in articles in the particular language. In one embodiment, for each segmented result, the search engine 156 can generate a search result set containing a number of article identifiers responsive to the search query. The search engine 156 can send the search result set for each of the segmented results back to the frequency processor 124 via the network 106. The frequency processor 124 can determine from each search result set, based on the number of article identifiers, the frequency with which each segmented result occurs.

In another embodiment, the frequency processor 124 can send the top, selected segmented results to the indexer 158 via the network 106. The indexer 158 can access the index 160 to determine the number of articles in the particular language in which a segmented result occurs and can do this for each of the selected segmented results. In one embodiment, the index 5 160 may be multiple indexes and the indexer 158 can check a fraction of the total index for each segmented result. The indexer 158 can then pass the number of occurrences associated with each segmented result to the frequency processor 124 via the network 106.

In still another embodiment, the frequency processor 124 can send the top, selected segmented results to the search engine 156 via the network 106 to determine the number of 10 occurrences of the segmented results in search queries. For example, the search engine 156 can, for each segmented result in the associated language, determine the number of times that the segmented result was used as a search query or part of a search query. The number of occurrences in search queries for each segmented result can be sent by the search engine 156 to the frequency processor 124 via the network 106.

15 For example, if the segmentation processor 122 determines that the selected segmented results for the string of characters "usedrugs" in English are "used rugs", "use drugs", and "us ed rugs", the frequency processor 124 can send these segmented results and the segmented results associated with other candidate languages to the search engine 156. The search engine 156 can, for example, use these results as search queries and generate search result sets for each 20 segmented result. For example, the search engine 156 can use "used rugs" as a search query and determine a search result set for the search query containing article identifiers associated with articles in English containing the phrase "used rugs". The search engine 156 can do the same for the segmented results associated with other candidate languages. In another embodiment, the search engine 156 can determine, from associated search logs containing prior received search 25 queries, the number of times search queries containing the segmented results was received. For example, the search engine 156 can search its search logs for the number of times a search query containing the phrase "used rugs" was received. In still another embodiment, the indexer 158 of the search engine 156 can receive the search results and determine the number of articles in the index 160 or a portion of the index 160 containing the segmented results. For example, the 30 indexer 158 can search through the index 160 or a portion of the index 160 for the number of

English language articles containing “used rugs”.

A spell-checking function can also be included in the frequency search. For example, the frequency processor 124 can include or can call a spell-checking function, so that the top, selected segmented results can be spell checked. The spell-checking function can determine correct or preferred spellings for the individual tokens in each segmented result. The frequency processor 124 can perform a frequency search on the top segmented results as well as any spelling-corrected segmented results to determine a frequency of occurrence for both results. For example, if a segmented result is “basebal game” and the spelling-corrected result is “baseball game”, a frequency search can be performed for both of these results.

In one embodiment, each frequency of occurrence for the segmented results is a normalized value based on the number of total articles or search queries in the particular language. For example, if a segmented result in English language occurs in 70 English language articles or search queries and there are a total number of 1000 English articles or search queries, the frequency of occurrence for this English segmented result is 0.07 (70/1000). Similarly, if a segmented result in French occurs in 60 French language articles or search queries and there are a total number of 400 French language articles or search queries, the frequency of occurrence for this French segmented result is 0.15 (60/400). This way the frequency of occurrence takes into consideration the prevalence of the particular language in the corpus of articles or search results and is not inherently weighted to more prevalent languages.

Block 212 is followed by block 214, in which the operable language and operable segmented result are identified. In one embodiment, the frequency processor 124 can identify the operable language and operable segmented result. For example, the frequency processor 124 can select the segmented result that has the highest associated frequency of occurrence. As explained above, the frequency of occurrence can be a normalized value based on the number of articles or search queries containing the segmented result and the total number of articles or search queries in the particular language. Additional signals can also be used to determine the operable segmented result. For example, the frequency processor 124 can take into account an objective ranking (such as the PageRank™ ranking algorithm for web articles) of the articles containing each segmented result and use the objective ranking to weigh the articles containing each segmented result. The number of times the segmented result occurs in an article and the

location of the segmented result in the articles can also be used to weigh the articles containing a segmented result. The candidate language associated with the operable segmented result can be selected as the operable language.

In one embodiment, the language signals used to identify the candidate languages in block 206 may be used in the determination of the operable language. If the language signals indicate that the string of characters is most likely a particular language, these signals can be used to more heavily weight this language. For example, the language signals, such as linguistics, the IP address of an associated user, the character set used for the string of characters, browser settings of browser application program associated with a user, and top-level domain associated with the string of characters, may indicate that the language associated with the string of characters is a particular language, such as French, for example. The frequency of occurrence information for a segmented result in another language, such as English, for example, may be close to or exceed the frequency of occurrence information for another segmented result in French. The language signals can be used to weight the French language in order to cause the selection of French as the operative language in this example. In 216, the method 200 ends.

The operable language and operable segmented result can be used in a variety of ways. The operable language and/or operable segmented result can be used in the selection of advertisements. For example, a user 112a may attempt to navigate his browser application to the website "usedrugs.com" by entering this string of characters into the browser application. If no such website exists at the domain name "usedrugs.com", the user's browser application may be redirected to a third-party website. The third-party website may desire to place advertisements and/or links relevant to the domain name entered by the user on a web page being viewed by the user. The third-party website can send the domain name "usedrugs.com" to the segmentation engine 120. The segmentation engine 120 can use the methods and systems described above to return an operable language and an operable segmented result to the third party website or an advertising server associated with the website. For example, the operable segmented result can be "used rugs" and the operable language can be English. The third-party website or advertising server can cause the display of advertisements and/or links relevant to the phrase "used rugs" in English on the web page being viewed by the user and can ensure that the language used on the website is English. The operable language can also be used in the selection of the language used

in status messages displayed to the user.

General

While the above description contains many specifics, these specifics should not be construed as limitations on the scope of the invention, but merely as exemplifications of the disclosed embodiments. Those skilled in the art will envision any other possible variations that are within the scope of the invention. The terms first and second are used herein merely to differentiate one item from another item. The terms first and second are not used to indicate first or second in time, first or second in a list, or other order, unless explicitly noted. For example, the “second” may come in time or in a list before the “first”, unless it is otherwise explicitly indicated.

60412-3762

CLAIMS:

1. A computer-implemented method, comprising:

receiving from a user of a computing device, at a computer server system, a request for information about one or more internet-accessible documents, the request having a string of characters;

identifying, using the computer server system, at least a first candidate language and a second candidate language associated with the request;

determining at least a first segmented result associated with the first candidate language from the string of characters and a second segmented result associated with the second candidate language from the string of characters;

determining a first frequency of occurrence for the first segmented result in a group of articles that are associated by the system with the first language and a second frequency of occurrence for the second segmented result in a group of articles that are associated by the system with the second language;

identifying, with the computer server system, an operable language from the first candidate language and the second candidate language based at least in part on the first frequency of occurrence and the second frequency of occurrence;

selecting, for use by the user of the computing device, electronic content in the identified operable language from among available content in multiple languages; and

providing the selected content to the computing device so that the selected content is arranged to be displayed to the user in the identified operable language and accompanying the requested one or more internet-accessible documents.

2. The method of claim 1, wherein identifying the operable language comprises identifying an operable segmented result from the first segmented

60412-3762

result and the second segmented result based at least in part on the first frequency of occurrence and the second frequency of occurrence.

3. The method of claim 1 or 2, wherein identifying a first candidate language and a second candidate language is based in part on at least one language signal determined from an electronic communication received from the user that submitted the string of characters.

4. The method of claim 3, wherein the at least one language signal comprises at least one of linguistics associated with the string of characters, an IP address of the user, a character set associated with the string of characters, a browser setting of a browser application program associated with the user, and a top-level domain associated with the string of characters.

5. The method of any one of claims 1 to 4, wherein the string of characters comprises a domain name submitted by the user.

6. The method of any one of claims 1 to 5, wherein the first segmented result comprises a first combination of tokens and the second segmented result comprises a second combination of tokens, wherein the first combination of tokens and the second combination of tokens are generated by separating the string of characters at a position between two characters that are immediately adjacent each other in the string of characters.

7. The method of any one of claims 1 to 6, wherein determining the first frequency of occurrence for the first segmented result comprises identifying a normalized value of a number of times the first segmented result appears in the first group of articles and identifying a normalized value of a number of times the second segmented result appears in the second group of articles.

8. The method of any one of claims 1 to 7, wherein determining the first frequency of occurrence for the first segmented result comprises determining a number of article identifiers in a search result set that has been generated in response to a search query comprising the first segmented result.

60412-3762

9. The method of claim 7, further comprising determining a number of articles that contain the first segmented result by accessing an index of articles.

10. The method of any one of claims 1 to 6, wherein determining the first frequency of occurrence for the first segmented result comprises determining a
5 number of occurrences of the first segmented result in a plurality of search queries in the first candidate language and normalizing the number of occurrences based on a total number of search queries in the first candidate language.

11. The method of any one of claims 1 to 4, wherein the string of characters comprises a domain name and determining a first segmented result
10 and a second segmented result comprises parsing the domain name into a pair of tokens at a plurality of locations along the string of characters.

12. The method of any one of claims 1 to 11, wherein providing the electronic content comprises identifying one or more keywords from the first segmented result or the second segmented result, and selecting electronic
15 content that matches the identified one or more keywords.

13. The method of claim 12, wherein the electronic content comprises at least one of an advertisement, a web page, and a status message.

14. The method of any one of claims 1 to 13, wherein determining the first segmented result comprises:

20 determining a plurality of segmented results in the first candidate language from the string of characters; and

identifying the first segmented result from the plurality of segmented results.

15. The method of claim 14, wherein identifying the first segmented
25 result comprises calculating a probability value for each of the plurality of segmented results.

60412-3762

16. The method of claim 15, wherein a first probability value associated with the first segmented result is based at least in part on a frequency of each token within the first segmented result.

17. The method of claim 1, wherein identifying the operable language
5 comprises identifying a number of hits of first segmented result against a database of content in the first candidate language and a number of hits of the second segmented result against a database of content in the second candidate language.

18. The computer-implemented method of any one of claims 1 to 17,
10 wherein each article in the group of articles that are associated by the system with the first language includes content in the first language;

wherein each article in the group of articles that are associated by the system with the second language includes content in the second language;
and

15 wherein the first language is different than the second language.

19. A tangible and non-transitory computer-readable medium containing program code executable on a computer, comprising:

program code for receiving from a user of a computing device, at a computer server system, a request to receive one or more internet-accessible
20 documents, the request having a string of characters;

program code for identifying at least a first candidate language and a second candidate language associated with a string of characters received in the request;

program code for determining at least a first segmented result
25 associated with the first candidate language from the string of characters and a second segmented result associated with the second candidate language from the string of characters;

60412-3762

program code for determining a first frequency of occurrence for the first segmented result in a group of articles that are associated by the computer server system with the first language and a second frequency of occurrence for the second segmented result in a group of articles that are associated by the
5 computer server system with the second language;

program code for identifying an operable language from the first candidate language and the second candidate language based at least in part on the first frequency of occurrence and the second frequency of occurrence;

program code for selecting, for use by the user of the computing
10 device, electronic content in the identified operable language from among available content in multiple languages; and

program code for providing the selected content to the computing device so that the selected content is arranged to be displayed to the user in the identified operable language with the requested one or more internet-accessible
15 documents.

20. The computer-readable medium of claim 19, wherein program code for identifying the operable language comprises program code for identifying an operable segmented result from the first segmented result and the second segmented result based at least in part on the first frequency of occurrence and
20 the second frequency of occurrence.

21. The computer-readable medium of claim 19 or 20, wherein identifying a first candidate language and a second candidate language is based in part on at least one language signal determined from an electronic communication received from the computing device that submitted the string of
25 characters.

22. The computer-readable medium of claim 21, wherein the at least one language signal comprise at least one of linguistics associated with the string of characters, an IP address of the user, a character set associated with the string of characters, a browser setting of a browser application program associated with
30 the user, and a top-level domain associated with the string of characters.

60412-3762

23. The computer-readable medium of any one of claims 19 to 22, wherein the string of characters comprises a domain name submitted by the user.

24. The computer-readable medium of any one of claims 19 to 23, wherein the first segmented result comprises a first combination of tokens and the
5 second segmented result comprises a second combination of tokens, wherein the first combination of tokens and the second combination of tokens are generated by separating the string of characters at a position between two characters that are immediately adjacent each other in the string of characters.

25. The computer-readable medium of any one of claims 19 to 24,
10 wherein program code for determining the first frequency of occurrence for the first segmented result comprises program code for identifying a normalized value of a number of times the first segmented result appears in the first group of articles and identifying a normalized value of a number of times the second segmented result appears in the second group of articles.

15 26. The computer-readable medium of any one of claims 19 to 25, wherein program code for determining the first frequency of occurrence for the first segmented result comprises program code for determining a number of article identifiers in a search result set generated in response to a search query comprising the first segmented result.

20 27. The computer-readable medium of claim 26, further comprises program code for determining a number of articles containing the first segmented result by accessing an index of articles.

28. The computer-readable medium of any one of claims 19 to 24, wherein program code for determining the first frequency of occurrence for the first
25 segmented result comprises program code for determining a number of occurrences of the first segmented result in a plurality of search queries in the first candidate language and normalizing the number of occurrences based on a total number of search queries in the first candidate language.

29. The computer-readable medium of any one of claims 19 to 22,
30 wherein the string of characters comprises a domain name and determining a first

60412-3762

segmented result and a second segmented result comprises parsing the domain name into a pair of tokens at a plurality of locations along the string of characters.

30. The computer-readable medium of any one of claims 19 to 29, wherein providing the electronic content comprises identifying one or more
5 keywords from the first segmented result or the second segmented result, and selecting electronic content that matches the identified one or more keywords.

31. The computer-readable medium of claim 30, wherein the electronic content comprises at least one of an advertisement, a web page, and a status message.

10 32. The computer-readable medium of any one of claims 19 to 31, wherein program code for determining the first segmented result comprises:

program code for determining a plurality of segmented results from the string of characters; and

15 program code for identifying the first segmented result from the plurality of segmented results.

33. The computer-readable medium of claim 32, wherein program code for identifying the first segmented result comprises program code for calculating a probability value for each of the plurality of segmented results.

20 34. The computer-readable medium of claim 33, wherein a first probability value associated with the first segmented result is based at least in part on a frequency of each token within the first segmented result.

35. The computer-readable medium of claim 19, wherein program code for identifying the operable language comprises identifying a number of hits of first segmented result against a database of content in the first candidate language
25 and a number of hits of the second segmented result against a database of content in the second candidate language.

36. A computer-implemented method, comprising:

60412-3762

receiving from a user of a computing device, at a computer server system, a request to receive one or more internet-accessible documents, the request having a string of characters that include a domain name;

5 determining at least a first segmented result in a first candidate language and at least a second segmented result in a second candidate language from the domain name;

10 determining at least a first frequency of occurrence for the first segmented result in a group of articles that are associated with the first language and based at least in part on at least one of an article index, a text index, and a search result set;

determining a second frequency of occurrence for the second segmented result in a group of articles that are associated with the second language and;

15 if the first frequency of occurrence is greater than the second frequency of occurrence, then selecting the first candidate language as an operable language;

if the second frequency of occurrence is greater than the first frequency of occurrence, then selecting the second candidate language as the operable language;

20 selecting an advertisement from among available advertisements in multiple languages based at least in part on the operable language, wherein the advertisement includes text in the operable language; and

25 providing the selected advertisement to the computing device arranged to be displayed to the user with the requested one or more internet-accessible documents associated with the domain name.

1/2

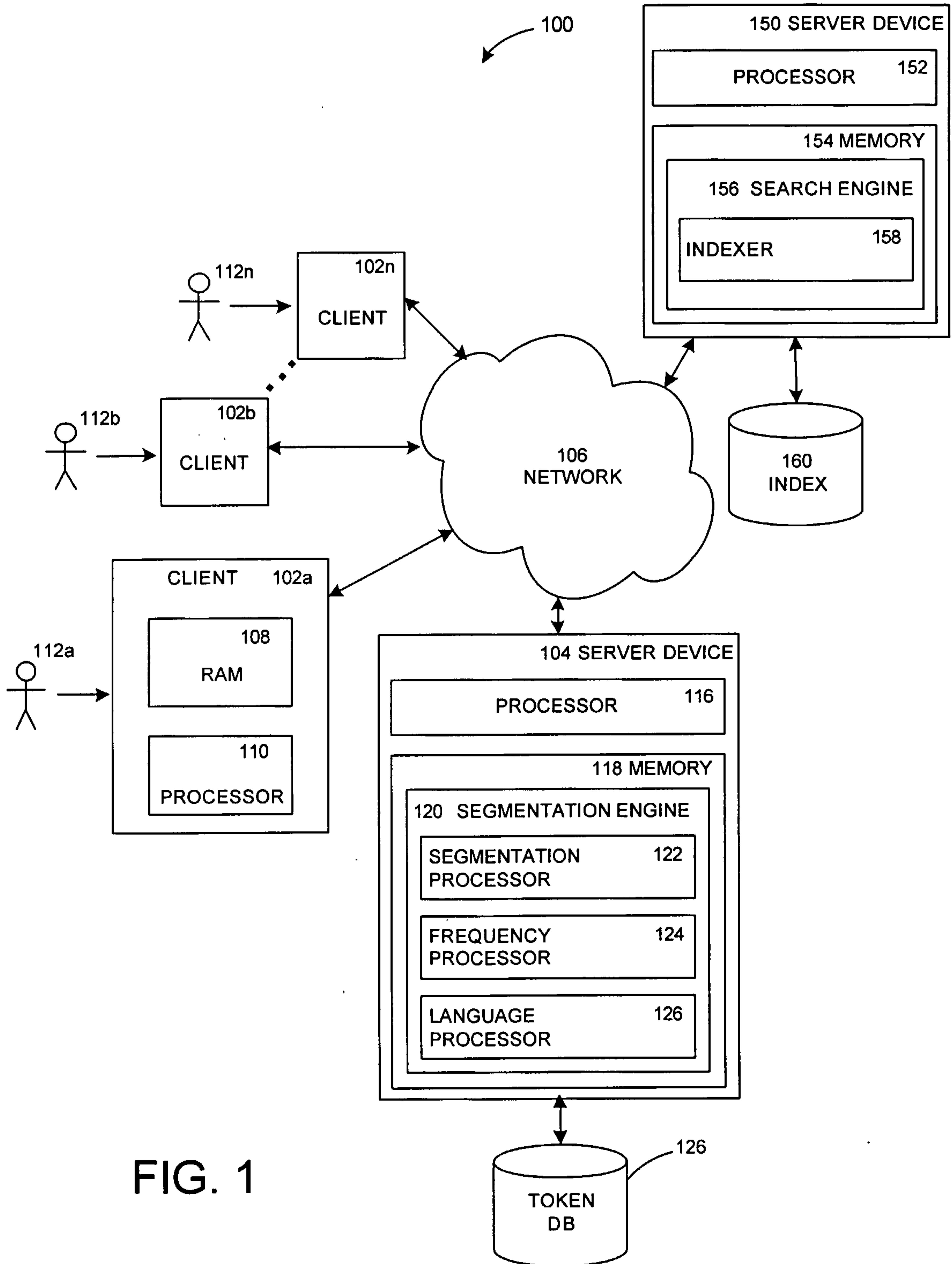


FIG. 1

2/2

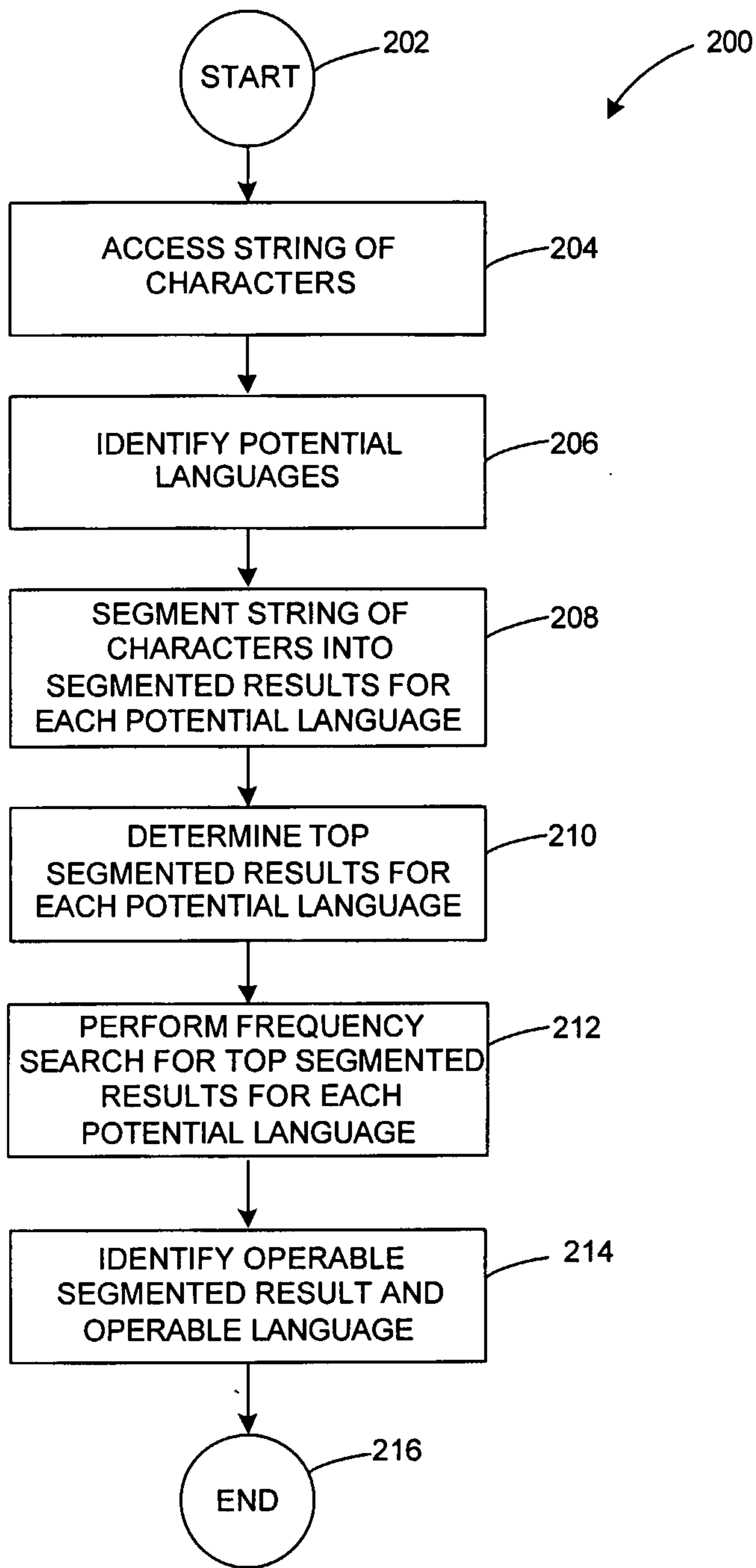


FIG. 2

