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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 December 2019 (19.12.2019)

WIPO | PCT

(10) International Publication Number
WO 2019/241425 A1

(51) International Patent Classification:
G06F 17/22 (2006.01) *G06F 16/33* (2019.01)
G06F 16/332 (2019.01)

(21) International Application Number:
PCT/US2019/036829

(22) International Filing Date:
12 June 2019 (12.06.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

62/684,498	13 June 2018 (13.06.2018)	US
62/749,001	22 October 2018 (22.10.2018)	US
16/438,325	11 June 2019 (11.06.2019)	US

(71) Applicant: **ORACLE INTERNATIONAL CORPORATION** [US/US]; 500 Oracle Parkway M/S 5OP7, Redwood Shores, California 94065 (US).

(72) Inventors: **MALAK, Michael**; 573 South Washington Street, Denver, Colorado 80209 (US). **RIVAS, Luis E.**; 8717 E. Frontier Place, Denver, Colorado 80237 (US). **KREIDER, Mark L.**; 17868 W. 78th Drive, Arvada, Colorado 8007 (US).

(74) Agent: **BRISNEHAN, Brian** et al.; KILPATRICK TOWNSEND & STOCKTON LLP, 1100 Peachtree Street NE, Suite 2800, Mailstop: IP Docketing - 22, Atlanta, Georgia 30309 (US).

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

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

(54) Title: REGULAR EXPRESSION GENERATION BASED ON POSITIVE AND NEGATIVE PATTERN MATCHING EXAMPLES

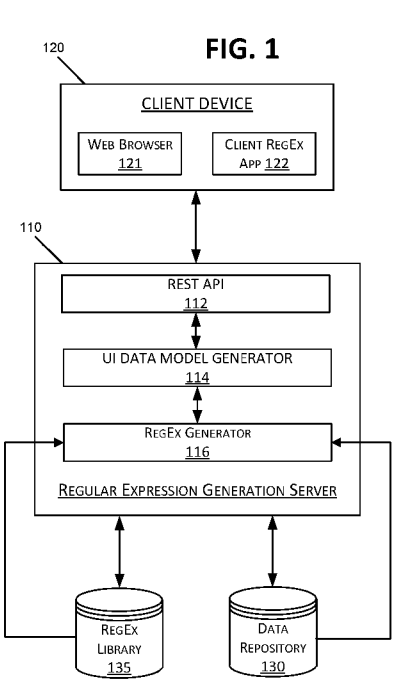


FIG. 1

(57) Abstract: Disclosed herein are techniques related to automated generation of regular expressions. In some embodiments, a regular expression generator may receive input data comprising one or more character sequences. The regular expression generator may convert character sequences into a sets of regular expression codes and/or span data structures. The regular expression generator may identify a longest common subsequence shared by the sets of regular expression codes and/or spans, and may generate a regular expression based upon the longest common subsequence.

WO 2019/241425 A1

[Continued on next page]

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

REGULAR EXPRESSION GENERATION BASED ON POSITIVE AND NEGATIVE PATTERN MATCHING EXAMPLES

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] The present application claims priority under 35 U.S.C. § 119(e) to U.S. Provisional Patent Application No. 62/684,498, filed June 13, 2018, entitled "AUTOMATED GENERATION OF REGULAR EXPRESSIONS," and also claims priority under 35 U.S.C. § 119(e) to U.S. Provisional Patent Application No. 62/749,001, filed October 22, 2018, entitled "AUTOMATED GENERATION OF REGULAR EXPRESSIONS." The entire contents of U.S. Provisional Patent Application Nos. 62/684,498 and 62/749,001 are incorporated herein by reference for all purposes.

BACKGROUND

[0002] Big data analytics systems can be used for predictive analytics, user behavior analytics, and other advanced data analytics. However, before any data analysis may be performed effectively to provide useful results, the initial data set may need to be formatted into clean and curated data sets. This data onboarding often presents challenges for cloud-based data repositories and other big data systems, where data from various different data sources and/or data streams may be compiled into a single data repository. Such data may include structured data in multiple different formats, semi-structured data in accordance with different data models, and even unstructured data. Repositories of such data often include data representations within various different formats and structures, and also may include duplicate data and erroneous data. When these data repositories are analyzed for reporting, predictive modeling, and other analytics tasks, a poor signal-to-noise ratio of the initial data set may lead to results that are inaccurate or not useful.

[0003] Many current solutions to the problems of data formatting and preprocessing include manual and ad hoc processes to clean and curate the data, in order to manipulate the data into a common format before performing a data analysis. While these manual processes can be effective for certain smaller data sets, such processes may be inefficient and impractical when attempting to preprocess and format large-scale data sets.

BRIEF SUMMARY

[0004] Aspects described herein provide various techniques for generating regular expressions. As used herein, a "regular expression" may refer to a sequence of characters defining a pattern, which may be used to search for matches within longer input text strings. In some embodiments, regular expressions may be composed using a symbolic wildcard-matching language, and the patterns defined by regular expressions may be used to match character strings and/or extract information from character strings provided as input. In various embodiments described herein, a regular expression generator implemented as data processing system may be used to receive and display input text data, receive selections via a client user interface of specific character subsets of the input text, and then generate one or more regular expressions based on the selected character subsets. After generating one or more regular expressions, a regular expression engine may be used to match the pattern of the regular expression against one or more data sets. In various embodiments, data matching the regular expression may be extracted, reformatted, or modified, etc. In some cases, additional columns, tables, or other data sets may be created based on the data matching the regular expression.

[0005] According to certain aspects described herein, a regular expression generator implemented via a data processing system may generate regular expressions based upon a determined longest common subsequence (LCS) that is shared by different sets of one or more regular expression codes. Regular expression codes (which also may be referred to as category codes) may include, for example, *L* for letters of the English alphabet, *N* for numbers, *Z* for white spaces, *P* for punctuation marks, and *S* for other symbols. Each set of one or more regular expression codes may be converted from a different sequence of one or more characters received as input data through a user interface. Regular expression codes excluded from the LCS may be represented as optional and/or alternatives. In some embodiments, a regular expression code may be associated with a minimum number of occurrences of the regular expression code. Additionally or alternatively, the regular expression code may be associated with a maximum number of occurrences of the regular expression code. For example, a set of category codes may comprise $L<0,1>$ to indicate that a particular portion of an LCS includes a letter at most once if at all. As discussed in more detail below, generalizing the input data as intermediate regular expression codes (IRECs) may provide various technical advantages, including, using very little

input data, enabling near-instantaneous generation of regular expressions that do not succumb to false positive matches or false negative matches in yet-to-be-seen data.

[0006] According to additional aspects described herein, a regular expression may be generated based on input data comprising three or more character sequences. When three or more character sequences are identified as input data, a regular expression generator that identifies the LCS of the character sequences may result in an exponential increase in runtime. In order to identify the LCS of all character sequences in a performant manner, the regular expression generator may perform an LCS algorithm on each distinct combination of two character sequences. A fully-connected graph may be generated based on the results of the LCS algorithms, where each graph node represents a different character sequence and the length of each graph edge corresponds to the LCS of the nodes defining the graph edge. The order for selecting character sequences then may be determined by performing a depth-first traversal of a minimum spanning tree for the fully-connected graph.

[0007] Further aspects described herein relate to generating regular expressions based on input including both positive character sequence examples and negative character sequence examples. A positive example may refer to sequence of characters that are to match the regular expression to be generated, while a negative example may refer to a sequence of characters that are not to match the regular expression to be generated. In some embodiments, when both positive and negative examples are received, the regular expression generator may identify a discriminator, or shortest subsequence of one or more characters that distinguish the positive example(s) from the negative example(s). The selected discriminator may be a shortest sequence (e.g., expressed in category codes), and may either be positive or negative, so that the positive examples will match and the negative examples will not. The discriminator then may be hard-coded into the regular expression that is generated by the regular expression generator. In some cases, the shortest subsequence may be included in a prefix or suffix portion of the negative example(s).

[0008] Additional aspects described herein relate to one or more user interfaces through which input data may be provided to generate regular expressions. In some embodiments, a user interface may be displayed at a client device communicatively coupled to the regular expression generator server. The user interface may be generated programmatically by the server, by the client device, or by a combination of software components executing at the server and the client.

Input data received via the user interface may correspond to user selections of one or more character sequences, which may represent positive or negative examples. In some cases, the user interface may support input data that includes a selection of a first character sequence within a second character sequence. For instance, a user may highlight one or more characters within a larger previously highlighted character sequence, and the second user selection may provide context for the larger first user selection. This enables input data to be provided to the regular expression generator with greater specificity, and to provide the regular expression generator with "context" so that it can generate regular expressions that avoid false positives. In response to a user selection of a character sequence via the user interface, the regular expression generator may generate and display a regular expression. For example, when a user highlights a first sequence of characters, the regular expression generator may generate and display a regular expression matching the first sequence of characters, as well as other similar character sequences (e.g., aligning with the intentions of the user for matching sequences). When the user highlights a second sequence of characters, the regular expression generator may generate an updated regular expression which encompasses both the first and second sequences of characters. Then, when the user highlights a third sequence of characters (e.g., within either the first or second sequence) the regular expression generator may update the regular expression again, and so on.

[0009] In accordance with additional aspects described herein, regular expressions may be generated based on the longest common subsequence from one or more input sequence examples, but also may handle characters that are present in only some of the examples. To handle characters that are present in only some input examples, spans may be defined in which both a minimum and maximum number of occurrences of a regular expression code are tracked. In cases when a span might not present at all of the given input examples, the minimum number of occurrences may be set to zero. These minimum and maximum numbers can then be mapped to the regular expression multiplicity syntax. A longest common subsequence (LCS) algorithm may be run on the spans of characters derived from the input examples, including "optional" spans (e.g., minimum length of zero) which do not appear in every input example. As discussed below, consecutive spans may be merged during the execution of the LCS algorithm. In such cases, when extra optional spans that are being carried along end up appearing consecutively, the LCS algorithm may be run recursively on those optional spans as well.

[0010] Further aspects described herein relate to a combinatoric search, in which the LCS algorithm executed by the regular expression generator may be run multiple times to generate a "correct" regular expression (e.g., a regular expression that properly matches all given positive examples and properly excludes all given negative examples), and/or to generate multiple correct regular expressions from which a most desirable or optimal regular expression may be selected. In some embodiments, an LCS algorithm may generally be executed right-to-left on the input examples to generate a regular expression. However, for comparison purposes and to find alternative regular expressions, the LCS algorithm may be separately executed backward (e.g., in the left-to-right direction) on the input examples. For example, the example character sequences received as user input may be reversed before they are run through the LCS algorithm, and the results from the LCS algorithm then may be reversed back (including the original text fragments). Further, in some embodiments, the LCS algorithm may be run multiple times by the regular expression generator, both in the usual character sequence order and the reverse order, with anchoring at the beginning of the line, anchoring at the end of the line, and no anchoring at the beginning or end the line. Thus, in some cases, the LCS algorithm may be execute at least these six times, and the shortest successful regular expression may be selected from these executions.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] FIG. 1 is a block diagram illustrating components of an exemplary distributed system for generating regular expressions, in which various embodiments may be implemented.

[0012] FIG. 2 is a flowchart illustrating a process for generating regular expressions based on input received via a user interface, according to one or more embodiments described herein.

[0013] FIG. 3 is a flowchart illustrating a process for generating regular expressions using a longest common subsequence (LCS) algorithm on sets of regular expression codes, according to one or more embodiments described herein.

[0014] FIG. 4 is an example diagram for generating a regular expression based on two character sequence examples, using a longest common subsequence (LCS) algorithm on sets of regular expression codes, according to one or more embodiments described herein.

[0015] FIG. 5 is a flowchart illustrating a process for generating regular expressions using a longest common subsequence (LCS) algorithm on larger sets of regular expression codes, according to one or more embodiments described herein.

[0016] FIG. 6 is an example diagram for generating a regular expression based on five character sequence examples, using a longest common subsequence (LCS) algorithm on sets of regular expression codes, according to one or more embodiments described herein.

[0017] FIG. 7 is a flowchart illustrating a process for determining an order of execution for a longest common subsequence (LCS) algorithm on larger sets of regular expression codes, according to one or more embodiments described herein.

[0018] FIGS. 8A and 8B show a fully-connected graph and a minimum spanning tree representation of the fully-connected graph, used for determining an order of execution for a longest common subsequence (LCS) algorithm on larger sets of regular expression codes, according to one or more embodiments described herein.

[0019] FIG. 9 is a flowchart illustrating a process for generating a regular expression based on positive and negative character sequence examples, according to one or more embodiments described herein.

[0020] FIGS. 10A and 10B are example user interface screens showing generation of regular expressions based on positive and negative character sequence examples, according to one or more embodiments described herein.

[0021] FIG. 11 is a flowchart illustrating a process for generating regular expressions based on user data selections received within a user interface, according to one or more embodiments described herein.

[0022] FIG. 12 is a flowchart illustrating a process for generating regular expressions and extracting data based on a capture group, via user data selections received within a user interface, according to one or more embodiments described herein.

[0023] FIG. 13 is an example user interface screen showing a tabular data display, according to one or more embodiments described herein.

[0024] FIGS. 14 and 15 are example user interface screens illustrating the generation of regular expressions and capture groups based on selection of data from a tabular display, according to one or more embodiments described herein.

[0025] FIGS. 16A and 16B are example user interface screens illustrating the generation of regular expressions based on selection of positive and negative examples from a tabular display, according to one or more embodiments described herein.

[0026] FIG. 17 is another example user interface screen illustrating the generation of a regular expression and capture group based on selection of data from a tabular display, according to one or more embodiments described herein.

[0027] FIG. 18 is a flowchart illustrating a process for generating regular expressions, including optional spans, using a longest common subsequence (LCS) algorithm, according to one or more embodiments described herein.

[0028] FIG. 19 is an example diagram for generating regular expressions, including optional spans, using a longest common subsequence (LCS) algorithm, according to one or more embodiments described herein.

[0029] FIG. 20 is a flowchart illustrating a process for generating regular expressions based on combinatoric executions of a longest common subsequence (LCS) algorithm, according to one or more embodiments described herein.

[0030] FIG. 21 is a block diagram illustrating components of an exemplary distributed system in which various embodiments of the present invention may be implemented.

[0031] FIG. 22 is a block diagram illustrating components of a system environment by which services provided by embodiments of the present invention may be offered as cloud services.

[0032] FIG. 23 is a block diagram illustrating an exemplary computer system in which embodiments of the present invention may be implemented.

DETAILED DESCRIPTION

[0033] In the following description, for the purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of various embodiments of the present

invention. It will be apparent, however, to one skilled in the art that embodiments of the present invention may be practiced without some of these specific details. In other instances, well-known structures and devices are shown in block diagram form.

[0034] The ensuing description provides exemplary embodiments only, and is not intended to limit the scope, applicability, or configuration of the disclosure. Rather, the ensuing description of the exemplary embodiments will provide those skilled in the art with an enabling description for implementing an exemplary embodiment. It should be understood that various changes may be made in the function and arrangement of elements without departing from the spirit and scope of the invention as set forth in the appended claims.

[0035] Specific details are given in the following description to provide a thorough understanding of the embodiments. However, it will be understood by one of ordinary skill in the art that the embodiments may be practiced without these specific details. For example, circuits, systems, networks, processes, and other components may be shown as components in block diagram form in order not to obscure the embodiments in unnecessary detail. In other instances, well-known circuits, processes, algorithms, structures, and techniques may be shown without unnecessary detail in order to avoid obscuring the embodiments.

[0036] Also, it is noted that individual embodiments may be described as a process which is depicted as a flowchart, a flow diagram, a data flow diagram, a structure diagram, or a block diagram. Although a flowchart may describe the operations as a sequential process, many of the operations can be performed in parallel or concurrently. In addition, the order of the operations may be re-arranged. A process is terminated when its operations are completed, but could have additional steps not included in a figure. A process may correspond to a method, a function, a procedure, a subroutine, a subprogram, etc. When a process corresponds to a function, its termination can correspond to a return of the function to the calling function or the main function.

[0037] The term “computer-readable medium” includes, but is not limited to non-transitory media such as portable or fixed storage devices, optical storage devices, and various other mediums capable of storing, containing or carrying instruction(s) and/or data. A code segment or computer-executable instructions may represent a procedure, a function, a subprogram, a program, a routine, a subroutine, a module, a software package, a class, or any combination of

instructions, data structures, or program statements. A code segment may be coupled to another code segment or a hardware circuit by passing and/or receiving information, data, arguments, parameters, or memory contents. Information, arguments, parameters, data, etc. may be passed, forwarded, or transmitted via any suitable means including memory sharing, message passing, token passing, network transmission, etc.

[0038] Furthermore, embodiments may be implemented by hardware, software, firmware, middleware, microcode, hardware description languages, or any combination thereof. When implemented in software, firmware, middleware or microcode, the program code or code segments to perform the necessary tasks may be stored in a machine readable medium. A processor(s) may perform the necessary tasks.

[0039] Various techniques (e.g., methods, systems, non-transitory computer-readable storage memory storing a plurality of instructions executable by one or more processors, etc.) are described herein for generating regular expressions corresponding to patterns identified within one or more input data examples. In certain embodiments, in response to receiving selections of input data, one or more patterns in the input data are automatically identified and a regular expression (or "regex" for short) may be automatically and efficiently generated to represent the identified patterns. Such patterns may be based upon sequences of characters (e.g., sequences of letters, numbers, spaces, punctuation marks, symbols, etc.). Various embodiments are described herein, including methods, systems, non-transitory computer-readable storage media storing programs, code, or instructions executable by one or more processors, and the like.

[0040] In some embodiments, regular expressions may be composed using a symbolic wildcard-matching language, in order to match character strings and/or extract information from character strings provided as input. For instance, a first example regular expression `[A-Za-z]{3} \d?\d, \d\d\d\d` may match certain dates (e.g., April 3, 2018), and a second example regular expression `[A-Za-z]{3} \d?\d, (\d\d\d\d)` may be used to extract the year from matching dates. Input data received by a regular expression generator system may include, for example, one more "positive" data examples, and/or one or more "negative" data examples. As used herein, a positive example may refer to a character sequence received as input that is to be matched by a regular expression generated based on the input. In

contrast, a negative example may refer to an input character sequence that is not to be matched by a regular expression generated based on the input.

[0041] A number of technical advantages may be realized within the various embodiments and examples described herein. For example, certain techniques described in this disclosure may improve speed and efficiency of regular expression generation processes (e.g., regex solutions may be generated in less than a second, and user interfaces may be suitable for interactive real-time use). Various techniques described herein also may be deterministic, may require no training data, may produce a solution without requiring any initial regular expression input, and may be completely automated (e.g., generating regular expressions within requiring any human intervention). Furthermore, various techniques described herein need not be limited regarding the types of data inputs that may be handled effectively, and such techniques may improve human readability of the resulting regular expressions.

[0042] Certain embodiments described herein include one or more executions of a Longest Common Subsequence (LCS) algorithm. LCS algorithms may be used in some contexts as difference engines (e.g., the engine behind the Unix "diff" utility) which are configured to determine and show differences between two text files. In some embodiments, input data (e.g., strings and other character sequences) may be converted into abstract tokens, which then may be provided as inputs to an LCS algorithm. Such abstract tokens may be for example, tokens based upon regular expression codes (e.g., Loogle codes or other character class codes) representing regular expression character classes. Various different examples of such codes are possible, and may be referred to herein as "regular expression codes" or "intermediate regular expression codes" (IRECs). For example, an input character sequence "May 3" may be converted to the IREC code "LLLZN," after which the tokenized string may be provided with other tokenized strings to the LCS algorithm. In some embodiments, an IREC (e.g., regular expression codes) that the input character sequences do not have in common, may appear in the final generated regular expression as optional (e.g., an optional span). In certain embodiments, regular expression codes may be category codes based upon the Unicode category codes shown at https://www.regular-expressions.info/unicode.html#category. For instance, the code L may represent letters , the code N may represent numbers, the code Z may represent spaces, the code S may represent symbols, the code P may represent punctuation, and so on. For example, the

code L may correspond to Unicode \p{L}and the code N may correspond to Unicode \p{N}. This allows for working one-to-one mappings from the LCS output to regular expressions (e.g. \pN\pN\pZ\pL\pL can match "10 am"), which may provide advantages for human readability. Additionally, these different categories may be disjoint, or mutually exclusive. That is, in this example, the categories L, N, Z, P, and S may be disjointed so that there may be no overlap between members of the categories.

[0043] Additional technical advantages may be realized in various embodiments, including more efficient generation of regular expressions based on the use of regular expressions codes (e.g., category codes), spans, etc. By using such codes, computing resources need not be wasted when the LCS algorithm successfully identifies all or substantially all of the characters in the input strings as being different. Further technical advantages provided by the various embodiments herein include improved readability of the generated regular expressions, as well as supporting both positive and negative examples as input data, and providing various advantageous user interface features (e.g., allowing the user to highlight text fragments within a larger character sequence or data cell for extraction).

I. GENERAL OVERVIEW

[0044] Various embodiments disclosed herein are related to generation of regular expressions. In some embodiments, a data processing system configured as a regular expression generator may generate a regular expression, by identifying a longest common subsequence (LCS) that is shared by different sets of regular expression codes (e.g., category codes). Each set of regular expression codes may be converted from sequence of characters received as input data through a user interface. Among the technical advantages described herein, abstracting input data as intermediate codes (e.g., regular expression codes, spans, etc.) may enable efficient generation of regular expressions using very little input data.

[0045] FIG. 1 is a block diagram illustrating components of an exemplary distributed system for generating regular expressions, in which various embodiments may be implemented. As shown in this example, a client device 120 may communicate with a regular expression generator server 110 (or regular expression generator) and interact with a user interface to retrieve and display tabular data, and generate regular expressions based on the selection of input data (e.g.,

examples) via the user interface. In some embodiments, a client device 120 may communicate with a regular expression generator 110 via a client web browser 121 and/or a client-side regular expression application 122 (e.g., client-side application that receives/consumes regular expressions generated by a server 110). Within the regular expression generator 110, requests from client devices 120 may be received over various communication networks at an network interface and processed by an application programming interface (API), such as a REST API 112. A user interface data model generator 114 component with the regular expression generator 110 may provide the server-side programming components and logic to generate and render the various user interface features described herein. Such feature may include the functionality to allow users to retrieve and display tabular data from data repositories 130, select input data examples to initiate the generation of regular expressions, and modify and/or extract data based the regular expressions generated. In this example, a regular expression generator component 116 may be implemented to generate regular expressions, including converting input character sequences into regular expression codes and/or spans, executing algorithms (e.g., LCS algorithms) on input data, and generating/simplifying regular expressions. The regular expressions generated by the regular expression generator 116, may be transmitted by the REST service 112 to the client device 120, where Javascript code on the client browser 121 (or corresponding client-side application components 122) may then apply the regular expression against every cell in the spreadsheet column rendered in the browser. In other cases, a separate regular expression engine component may be implemented on the server-side to compare the generated regular expressions with the tabular data displayed on the user interface and/or within other data stored in data repositories 130, in order to identify matching data/non-matching data on the server-side. In various embodiments, the matching/non-matching data may be automatically selected (e.g., highlighted) within the user interface, and may be selected for extraction, modification, deletion, etc. Any data extracted or modified via the user interface, based on the generation of the regular expressions, may be stored in one or more data repositories 130. Additionally, in some embodiments, the regular expressions generated (and/or corresponding inputs to the LCS algorithm) may be stored in a regular expression library 135 for future retrieval and use. In some embodiments, the generated regular expressions need not actually be stored in a "library," buy may be incorporated into a "transform script". For examples, as described in more detail in U.S. Patent No. 10,210, 246 (which is incorporated

herein by reference for all purposes), such transform scripts may include programs, code, or instructions that may be executable by one or more processing units to transform received data. Other possible examples of transform script actions may include "rename column", "uppercase column data", or "infer gender from first name and create a new column with gender", etc.

[0046] FIG. 2 is a flowchart illustrating a process for generating regular expressions based on input received via a user interface, according to one or more embodiments described herein. In step 201, the regular expression generator 110 may receive a request from a client device 120 to access a regular expression generator user interface, and to view particular data via the user interface. The request in step 201 may be received via the REST API 112, and/or a web server, authentication server, or the like, and the user's request may be parsed and authenticated. For instance, a user within an business or organization may access the regular expression generator 110 to analyze and/or modify transaction data, customer data, performance data, forecast data, and/or any other categories of data that may be stored in the data repositories 130 of the organization. In step 202, the regular expression generator 110 may retrieve and display the requested data via a user interface that supports generation of regular expressions based on selected input data. Various embodiments and examples of such user interfaces are described in detail below.

[0047] In step 203, a user may select one or more input character sequences, from the data displayed in the user interface provided by the regular expression generator 110. In some embodiments, the data may be displayed in tabular form within the user interface, including labeled columns with specific data types and/or categories of data. In such cases, the selection of input data in step 203 may correspond to a user selecting a data cell, or selecting (e.g., highlighting) an individual text fragment within a data cell. However, in other embodiments, the regular expression generator 110 may support retrieval and display of semi-structured and unstructured data via the user interface, and users may select input data for regular expression generation by selecting character sequences from the semi-structured or unstructured data. As described below in examples, the user selecting input character sequences from the tabular data displayed is just one example use case. In other examples, a user (e.g., a software developer or power user perhaps trying to compose a regular expression for the Linux command line tools

grep, sed, or awk, etc.) may type in examples from scratch rather than picking them off a spreadsheet.

[0048] In step 204, the regular expression generator 110 may generate one or more regular expressions based on the input data selected by the user in step 203. In step 205, the regular expression generator 110 may update the user interface, for example, to display the generated regular expression and/or to highlight matching/non-matching data within the displayed data. In step 206, which may be optional in some embodiments, the user interface may support functionality to allow the user to modify the underlying data based on the generated regular expression. For example, the user interface may support features to allow the user to filter, modify, delete, or extract particular data fields from the tabular data, based on whether those fields match or do not match the regular expression. Filtering or modifying data may include modifying the underlying data stored in the repositories 130, and in some cases, extracted data may be stored in a repository 130 as new columns and/or new tables.

[0049] Although these steps illustrate a general and high-level overview of an example user interaction with the user interface of the regular expression generator 110, various additional features and functionalities may be supported in other embodiments. For example, in some embodiments, a regular expression code (or category code) may be associated with a minimum number of occurrences of the code. Additionally or alternatively, the regular expression code may be associated with a maximum number of occurrences of the code. As an example, a set of regular expression codes may include the code $L<0,1>$ to indicate that a particular portion of an LCS includes a letter either at least zero times, and at most once.

[0050] Additionally, in some embodiments, the input data may include three or more character sequences. In such embodiments, techniques may be used to determine order for performing the LCS algorithms on the three or more character sequences, so that the resulting regular expression may be generated in a performant manner to avoid the exponential increase in runtime caused by the three or more input character sequences. The regular expression generator 110 may instead perform an LCS algorithm on two character sequences at a time, and may determine an order for selecting the pair of character sequences based on a graph. For example, a fully-connected graph may indicate that a first execution of the LCS algorithm (e.g., LCS1) should be performed for Sequence1 and Sequence3, and then a second execution of the LCS algorithm (e.g., LCS2)

should be perform for LCS1 and Sequence2, and so on. The graph may be a fully-connected graph, with nodes representing the character sequences, and edges connecting the nodes to represent the length of an LCS shared by the connected nodes. Each node in the graph may be connected to every other node in the graph, and the order for selecting the character sequences may be determined by a performing a depth-first traversal of a minimum spanning tree for the graph.

[0051] In further embodiments, input data may be provided via the user interface in a number of different ways. For example, the input data may indicate a first user selection of one or more characters within a second user selection of a set of characters. For instance, a user may highlight a character within a set of previously highlighted characters. Thus, a second user selection may provide context for the first user selection, which may enable input data to be provided to the regular expression generator 110 with greater specificity. In some embodiments, the regular expression generator 110 may generate and display, in near-real-time, a regular expression in response to each user selection. For example, when a user highlights a first range of characters, the regular expression generator 110 may display a regular expression representing the first range of characters. Then, when the user highlights a second range of characters within the first range of characters, the regular expression generator 110 may update the regular expression that is displayed.

[0052] Additionally, in some embodiments, the regular expression generator 110 may generate regular expressions based on input comprising both positive and negative examples. As noted above, a positive example may refer to a sequence of characters that are to be encompassed by a regular expression, and a negative example may refer to a sequence of characters that are not to be encompassed by the regular expression. In such cases, the regular expression generator 110 may identify a shortest subsequence of one or more characters, at a particular location, that distinguish the positive example(s) from the negative example(s). The shortest subsequence then may be hard-coded within the regular expression that is generated by the regular expression generator 110. In various examples, the shortest subsequence may be included in a prefix/suffix portion, or mid-span within the negative example(s).

[0053] Further examples for automatically generating regular expressions according to certain embodiments are described below. These examples may correspond to various specific possible

implementations of the general technique in FIG. 2, and be implemented in software (e.g., code, instructions, programs, etc.) executed by one or more processing units (e.g., processors, cores) of the respective systems, hardware, or combinations thereof. The software may be stored on a non-transitory storage medium (e.g., on a memory device). The further examples described below are intended to be illustrative and non-limiting. Although these examples depict the various processing steps occurring in a particular sequence or order, this is not intended to be limiting. In certain alternative embodiments, the steps may be performed in some different order or some steps may also be performed in parallel.

[0054] In some examples, the user inputs received via the user interface (e.g., step 203) may include one or more "positive examples" to be matched by the regular expression output, and zero or more "negative examples" that are not to be matched by the regular expression output. Optionally, one or more of the positive examples may be highlighted to select a particular range (or subsequence) of characters. In some cases, in step 204, the positive examples received via the user interface may be converted to spans of regular expression codes (e.g., character category codes such as Unicode category codes). For each positive example, a sequence of spans may be generated. A graph may be created in some embodiments, where each vertex corresponds to one of the sequences of spans, and the edge weight equals the length of the output from the LCS algorithm executed on those two sequences of spans corresponding to the endpoints of the edge. A minimum spanning tree may be determined for the graph. For example, Prim's algorithm may be used in some embodiments to obtain a minimum spanning tree. A depth-first traversal may be performed on the minimum spanning tree to determine a traversal order, after which the LCS algorithm may be executed on the first two elements of the traversal. Then, one by one, each additional element of the traversal may be merged in order into the current LCS output, by executing the LCS algorithm again on the output of the previous LCS iteration and the next current traversal element. The final output of the LCS algorithm, which may be a sequence of spans, then may be converted into a regular expression. The conversion may be a one-to-one conversion in some embodiments, while certain optional embellishments described herein might not correspond to one-to-one conversions. Finally, the resulting regular expression may be tested against all positive and negative examples received via the user interface in step 203. If any of the tests fail, then the aforementioned process may be repeated using all the positive examples and any negative examples that failed.

II. REGULAR EXPRESSION GENERATION USING LONGEST COMMON SUBSEQUENCE ALGORITHM ON REGULAR EXPRESSION CODES

[0055] As noted above, certain aspects described herein relate to generation of regular expressions based upon the calculation of a longest common subsequence (LCS) shared by different sets of regular expression codes corresponding to input data.

[0056] FIG. 3 is a flowchart illustrating a process for generating regular expressions using an LCS algorithm on sets of regular expression codes, according to one or more embodiments described herein. In step 301, the regular expression generator 110 may receive one or more character sequences as input data. As noted above, in some examples, the input data may correspond to positive example data selected from within the tabular data displayed in the user interface, although it should be understood that the user interface is optional in some embodiments, and the input data may correspond to any character sequence received any other communication channel (e.g., non-user interface) in various examples.

[0057] In step 302, each character sequence received in step 301 may be converted into a corresponding regular expression code. In various embodiments, the regular expression codes may be Loogle codes, Unicode category codes, or any other character class codes representing regular expression character classes. For example, an input character sequence "May 3" may be converted to the Loogle code "LLLZN." In some embodiments, regular expression codes may be category codes based upon the Unicode category codes shown at https://www.regular-expressions.info/unicode.html#category. For instance, the code L may represent letters , the code N may represent numbers, the code Z may represent spaces, the code S may represent symbols, the code P may represent punctuation, and so on. For example, the code L may correspond to Unicode \p{L}and the code N may correspond to Unicode \p{N}.

[0058] In step 303, a longest common subsequence may be determined from among the sets of regular expression codes generated in step 302. In some embodiments, an LCS algorithm may be executed using two sets of regular expression codes as input. Various different characteristics of the execution of the LCS algorithm (e.g., direction of processing, anchoring, pushing spaces, coalescing low cardinality spans, aligning on common tokens, etc.), may be used in different embodiments. In step 304, a regular expression may be generated based on the output of the

LCS algorithm. In some cases, step 304 may include capturing the output of the LCS algorithm in regular expression codes, and converting the regular expression codes into a regular expression. In step 305, the regular expression may be simplified and output, for example, by displaying the regular expression for the user via the user interface.

[0059] FIG. 4 is an example diagram for generating a regular expression based on two character sequence examples, using a longest common subsequence (LCS) algorithm on sets of regular expression codes. Thus, FIG. 4 shows an example of applying the process discussed above in FIG. 3. As shown in FIG. 4, the regular expression in this example is generated based on the two input strings: "iPhone 5" and "iPhone X." Each sequence in this example may be converted into a respective set of regular expression codes. Thus, iPhone 5 may be converted into "LLLLLLZN," and iPhone X may be converted into "LLLLLLZL." As shown in FIG. 4, these category codes are then provided as input to an LCS algorithm, which determines that both sets of IRECs (or category codes) comprise six Ls and one Z. Category codes excluded from the LCS may be represented as optional and/or alternatives. Thus, a regular expression that encompasses both character sequences may be represented as the following: `\pL{6}\pZ\pN?\pL?` In this example, the regular expression includes Unicode category codes (e.g., `\pL` for letters, `\pZ` for spaces, and `\pN` for numbers). The curly braces containing the number 6 indicates six instances of a letter, and the question marks indicate that a number/letter at the end are optional. Finally, a simplification process may be executed by the regular expression generator, during which the regular expression is simplified by inserting the common text fragment "iPhone" back into the final regular expression, replacing the broader "`\pL{6}\`" portion of the regular expression.

[0060] As shown in this example, the input strings received by the regular expression generator 110 may be converted into "regular expression codes" representing regular expression broad categories (which also may be referred to as "category codes"), and the LCS algorithm may be run on those regular expression codes. In some embodiments, the Unicode category codes may be used for the regular expression codes. For example, an input text string may be converted into codes representing regex Unicode broad categories (e.g., \pL for letters, \pP for punctuation, etc.). This approach, illustrated by FIGS. 3 and 4 may be referred to as the indirect

approach. However, in other embodiments, a direct approach may be used, in which the LCS algorithm is run directly on the character sequences received as input.

[0061] In some embodiments, the indirect approach may provide additional technical advantages, in that it need not require large amounts of training data, and may generate an effective regular expression with a relatively lower number of input examples. This is because the indirect approach employs heuristics to reduce the uncertainty in the regular expression generation, and to eliminate potential false positives and false negatives. For example, in generating a regular expression based on the input strings "May 3" and "Apr 11," the direct approach may need at least one example for every month to generate an effective regular expression matching date patterns. Relying on only those two examples, the direct approach may generate a regex of "`[AM][ap][yr] [13]1?`" In contrast, the indirect approach, based on Unicode broad categories, may generate a more effective regular expression of "`\pL{3} \d{1,2}`". Additionally, as noted above, one of the technical advantages described herein includes efficient generation of regular expressions using very little input data, even potentially from a single example. For instance, regarding generation of a regular expression from the single example "am", a heuristic may determine whether to generate "am" or "\pL\pL" for the regular expression. Either is arguably correct, but so a programmed heuristic may implement user preferences and/or criteria to determine how to generate an optimal regular expression (e.g., whether or not it should match "pm" as well).

[0062] Additionally, the indirect approach may further simplify the generated regular expression "`\pL{3} \d{1,2}`" to "`[A-Za-z]{3} \d{1,2}`" to make it more human-readable. This may be beneficial in some embodiments, such as when outputting to non-sophisticated regular expression users who might not be familiar with the Unicode expressions for regular expressions.

[0063] Further, in some embodiments, instead of treating each character independently when executing the LCS algorithm, sequential and equal regular expression codes may be converted into span data structures (which also may be referred to as spans). In some cases, a span may include a representation of single regular expression code (e.g., Unicode broad category code), along with a repetition count range (e.g., a minimum number and/or a maximum number). Conversion from regular expression codes into spans may facilitates some various additional

features described below, such as recognizing alternations (e.g., disjunctions), and also may facilitate merging of adjacent optional spans to further simplify the generated regular expressions.

[0064] As noted above, the LCS algorithm may be configured to store and retain the underlying text fragments within the input character sequences, which may potentially be inserted back into the final regular expression, such as the string "iPhone" in FIG. 4. By keeping track of the text fragments that originally gave rise to the category code assigned to that span, such embodiments may allow for literal text (e.g., `am` and `pm`) to be included directly in the generated regular expression, which may reduce false positives and make the regular expression output more human readable.

III. REGULAR EXPRESSION GENERATION USING LONGEST COMMON SUBSEQUENCE ALGORITHM ON COMBINATIONS OF REGULAR EXPRESSION CODES

[0065] Additional aspects described herein relate to the generation of regular expressions based on input data comprising three or more strings (e.g., three or more separate character sequences). When three or more strings are identified as input data, the regular expression generator 110 may use a performance optimization feature in which an optimal order is determined for the sequence of LCS algorithm executions. As discussed below, the performance optimization feature for more than two strings may involve building a graph with a vertex corresponding to each string, and edge lengths/weights which may be based on the size of the LCS output between each string and every other string. A minimum spanning tree then may be derived using those edge weights, and a depth-first traversal may be performed to determine an order of the input strings. Finally, the series of LCS algorithms may be done using the determined order of input strings.

[0066] FIG. 5 is a flowchart illustrating a process for generating regular expressions using a longest common subsequence (LCS) algorithm on larger sets (e.g., three or more character sequences) of regular expression codes. Thus, steps 502-505 in this example may correspond to step 303 discussed above in FIG. 3. However, because this example relates to generating regular expressions based on three or more input character sequences, the LCS algorithm may be

performed multiple times. For example, in order to avoid an exponential increase in runtime for three or more input strings, the LCS algorithm may be executed multiple times, wherein each execution is performed on only two input strings. For example, the regular expression generator 110 may perform an initial execution of the LCS algorithm on two strings (e.g., two input character sequences or two converted regular expression codes), then may perform a second execution of the LCS algorithm on the output of the first LCS algorithm and a third string, and then may perform a third execution of the LCS algorithm on the output of the second LCS algorithm and a fourth string, and so on.

[0067] In order to improve and/or optimize the performance of such embodiments, it may be desirable to determine an optimal order for the input strings (e.g., input character sequences or regular expression codes) to perform the sequence of LCS algorithms. For example, a good order for taking the input strings may affect the readability of the generated regular expression, such as by minimizing the number of optional spans. To keep the generated regex concise, additional strings that are LCS'd into the current regex should preferably already be somewhat similar to the current regex (the intermediate result from LCS'ing the already-seen strings).

[0068] Thus, in step 501, the plurality (e.g., 3 or more) input character sequences are converted into regular expression codes. In step 502, an order is determined for processing the regular expression codes using the LCS algorithm. The determination of the order in step 502 is discussed more below in reference to FIG. 7. In step 503, either the first two regular expression codes in the determined order are selected (for the first iteration of step 503), or the next regular expression codes in the determined order is selected (for subsequent iterations of step 503). In step 504, the LCS algorithm is executed on two input strings corresponding to the format of regular expression codes. For the first iteration of step 504, the LCS algorithm is executed on the first two regular expression codes in the determined order, and for subsequent iterations of step 504, the LCS algorithm is executed on the next regular expression code in the determined order and the output of the previous LCS algorithm (which also may be in same format of regular expression codes). In step 505, the regular expression generator 110 determines whether or not there are additional regular expression codes in the determined order that have not yet been provided as input to the LCS algorithm. If so, the process returns to step 503 for another

execution of the LCS algorithm. If not, in step 506, a regular expression is generated based on the output of the last execution of the LCS algorithm.

[0069] FIG. 6 is an example diagram for generating a regular expression based on five input character sequence examples. In this example, each input character sequence is converted to a regular expression code, and then an LCS algorithm is executed repeatedly based on a determined order of the regular expression codes. Thus, FIG. 6 shows one example of applying the process discussed above in FIG. 5. In this example, the determined order for the five regular expression codes is Code #1 to Code #5, and each codes is input to the LCS algorithm in the determined order to generate a regular expression output. The final regular expression output (Reg Ex #4) corresponds to the final regular expression generated based on all five of the input character sequences.

[0070] FIG. 7 is a flowchart illustrating a process for determining an order of execution for a longest common subsequence (LCS) algorithm on larger sets (e.g., three or more) of regular expression codes. Thus, as shown in this example, steps 701-704 may correspond to the order determination in step 502, discussed above. In step 701, the LCS algorithm may be run on each unique pair of regular expression codes corresponding to the input data, and the resulting output LCS may be stored for each execution. Thus, for k number of input data, this may represent all (k(k-1)) / 2 possible pairings of strings to be run through the LCS algorithm, or k(k-1) in some embodiments. For example, if k=3 input character sequences are received, LCS algorithm may be run three times in step 701; if k=4 input character sequences are received, the LCS algorithm may be run six times in step 701; if k=5 input character sequences are received, the LCS algorithm may be run ten times in step 701, and so on. In step 702, a fully-connected graph may be constructed of k nodes representing the strings with the edge weight of the (k(k-1)) / 2 edges being the length of the raw LCS output between the two nodes. In step 703, a minimum spanning tree may be derived from the fully-connected graph in step 702. In step 704, a depth-first traversal may be performed on the minimum spanning tree. The output of this traversal may correspond to the order in which regular expression codes will be input into the sequence of LCS algorithm executions.

[0071] Referring briefly to FIGS. 8A and 8B, an example of a fully-connected graph is shown in FIG. 5, generated based on k=5 input character sequences received, and in FIG. 8B a minimum spanning tree representation is shown for the fully-connected graph.

[0072] In some embodiments, the approach described in FIGS. 5-8B may provide additional technical advantages with respect to performance. For example, certain conventional implementations of the LCS algorithm may exhibit a run-time performance of $O(n^2)$ where n is the length of the strings. Extending such implementations to k strings instead of only 2, may results in an exponential run-time performance $O(n^k)$, because the LCS algorithm may be required to search a k-dimensional space. Such conventional implementations of the LCS algorithm might not be performant or sufficiently suitable for real-time on-line user experiences.

[0073] As noted above, the LCS algorithm may be executed (k(k-1)) / 2 times, where sometimes the duplicates are the very same as have been seen before, because the LCS algorithm may when the raw input examples from the user have been converted to regex category codes. Thus, memoization may be implemented in some cases, in which a cache can be used to map previously-seen LCS problems to the previously worked LCS solution.

IV. REGULAR EXPRESSION GENERATION BASED ON POSITIVE AND NEGATIVE PATTERN MATCHING EXAMPLES

[0074] Additional aspects described herein relate to generating regular expressions based on input data corresponding to both positive and negative examples. As noted above, a positive example may refer to an input data character sequence that is designated as an example string that should match the regular expression that will be generated by the regular expression generator. In contrast, a negative example may refer to an input data character sequence that is designated as an example string that should not match the regular expression that will be generated by the regular expression generator. As discussed below, in some embodiments, the regular expression generator 110 may be configured to identify a location and a shortest subsequence of characters at the location that distinguish the positive examples from the negative examples. The shortest subsequence then may be hard-coded into the generated regular expression, so that the positive examples will match the regular expression and the negative examples will be excluded by (e.g., will not match) the regular expression .

[0075] FIG. 9 is a flowchart illustrating a process for generating a regular expression based on positive and negative character sequence examples. In step 901, the regular expression generator 110 may receive one or more input data character sequences corresponding to positive examples. In step 902, the regular expression generator 110 may generate a regular expression based on the received positive examples. Thus, steps 901-902 may include some or all of the steps performed in FIG. 3 or FIG. 5, discussed above, to generate a regular expression based on input data character sequences.

[0076] In step 903, the regular expression generator 110 may receive one additional input data character sequences corresponding to negative examples. Thus, the negative examples are specifically designated so as not the match the regular expression generated in step 902. In some embodiments, the negative examples received in step 903 may be initially tested against the regular expression generated in step 902, and if it is determined that the negative examples do not match the regular expression, then no further action is taken. However, in this example it may be assumed that at least one of the negative examples received in step 903 matches the regular expression generated in step 902. Thus, in step 904, a disambiguation location may be determined within the regular expression generated in step 902. In some embodiments, the disambiguation location may be selected as either the prefix location (e.g., at the beginning of the regular expression) or the suffix location (e.g., at the end of the regular expression). For instance, the regular expression generator 110 may determine a first number of characters that would be needed at the prefix to distinguish the positive examples from the negative examples, and second number of characters that would be needed at the suffix to distinguish the positive examples from the negative examples. The regular expression generator 110 may then select the suffix or prefix based on the shortest number of replacement characters needed. In some cases, using the prefix as the disambiguation location may be preferred (e.g., weighted) for readability purposes. In still other examples, the disambiguation location may be a mid-span location that does not correspond to the prefix or suffix of the regular expression.

[0077] In step 905, the regular expression generator 110 may determine a replacement sequence of custom character classes which, when inserted into the regular expression at the determined location, may distinguish the positive examples from the negative examples. In some embodiments, the regular expression generator 110 in step 905 may retrieve text fragments

from each of the positive and negative examples, corresponding to the disambiguation location (or replacement location), and then use the text fragments to determine a discriminator to be used as a replacement sequence that distinguishes the positive examples from the negative examples. Additionally, the discriminator replacement sequence determined in step 905 may include multiple different replacement sequences of custom character classes, which may be replaced either at the same location or at different locations within the regular expression.

[0078] As noted above, in some cases, the determination of the replacement sequence in step 905 may be performed in conjunction with the determination of the disambiguation location (or replacement location) in step 904. For example, the regular expression generator 110 may determine one or more replacement sequences which, at a first possible replacement location, may distinguish the positive from the negative examples. The regular expression generator 110 also may determine one or more other replacement sequences which, at a second different possible replacement location, may distinguish the positive from the negative examples. In this example, when selecting between the different possible replacement locations and corresponding replacement sequences, the regular expression generator 110 may apply a heuristic formula to perform the selection based on one or more of the sizes in characters of the replacement locations, and the numbers and/or sizes of the corresponding replacement sequences. Finally, in step 906, the regular expression may be modified by inserting the one or more determined replacement sequences into the determined location to replace the previous portion of the regular expression. In some cases, following the modification of the regular expression in step 906, the positive and/or negative examples may be tested against the modified regular expression to confirm that the positive examples match and that the negative examples do not match the regular expression.

[0079] FIGS. 10A and 10B are example user interface screens showing generation of regular expressions based on positive and negative character sequence examples. Thus, the example shown in FIGS. 10A and 10B may correspond to the user interfaces displayed during the execution of the process of FIG. 9 discussed above. In FIG. 10A, the user provides three positive examples of data input character sequences 1001, and the regular expression generator 110 generates a regular expression 1002 that matches each of the positive examples. Then, in FIG. 4B, the user provides one negative example 1004, and the regular expression generator 110

generates a modified regular expression 1005, which is based on both the current sets of positive examples 1003 and negative examples 1004.

[0080] As noted above, in some embodiments, when both positive and negative examples are received, the regular expression generator 110 may identify a discriminator, or the shortest subsequence of one or more characters that distinguish the positive example(s) from the negative example(s). The selected discriminator may be a shortest sequence (e.g., expressed in category codes), and may either be positive or negative, so that the positive examples will match and the negative examples will not. In some cases, the discriminator may correspond to a replacement subsequence which then may be hardcoded into the regular expression in step 905. As an example, in "[AL][a-z]+" the [AL] is a positive discriminator that, assuming it is applied to street suffixes, would match with (or allow) "Alley", "Avenue", and "Lane" but would not match with (or disallow) everything else. As another example, in "[BC][o][a-z]+" the [BC][o] is a positive dscrimnator consisting of a sequence of two character classes that would match with "Boulevard" and "Court". As yet another example, in "[^A][a-z]+" the [^A] may be a negative discriminator that would disallow "Alley" and "Avenue". In some cases, the algorithm may make generate a negative-look-behind to discriminate correctly. For example, (?<!Av)[A-Za-z]+ would exclude "Avenue" but would allow "Alley".

[0081] As another example, if the user supplies the positive examples "202-456-7800" and "313-678-8900" and negative examples "404-765-9876" and "515-987-6570", then in certain embodiments, the regular expression generator 110 may generate the regular expression "`\d\d\d-\d\d\d-\d\d00`". That is, the replacement character subsequence may be identified for the suffix of the regular expression, based on the determination that phone numbers that end in 00 distinguish the positive examples from the negative examples (e.g., assuming that the goal is a regular expression the matches business phone numbers). This is an example of negative example by suffix (or more specifically, an example of accomodating negative examples by using a positive suffix), but various other embodiments may support either replacements at prefixes, suffixes, or mid-span locations. In examples of replacement at mid-span locations, a character offset into the span may be kept track of, and may be split at the mid-span point.

[0082] To decide between whether to use a prefix or suffix, in some embodiments, a heuristic is employed where the minimum score is chosen over all combinations of k_a and prefix/suffix:

$$score = k_a \min^2 \left\{ \frac{|\mathbf{F}_p|}{1 + |\mathbf{E}_p|}, \frac{|\mathbf{F}_n|}{1 + |\mathbf{E}_n|} \right\} + \begin{cases} 0.0 \text{ if prefix} \\ 0.1 \text{ if suffix} \end{cases}$$

Where:

k_a = number of characters being considered to disambiguate the affix (prefix or suffix)

$|\mathbf{F}_p|$ = number of unique text fragments from the positive examples required to disambiguate the affix

$|\mathbf{F}_n|$ = number of unique text fragments from the negative examples required to disambiguate the affix

$|\mathbf{E}_p|$ = number of (complete) positive examples provided by the user

$|\mathbf{E}_n|$ = number of (complete) negative examples provided by the user

[0083] In the above example, the heuristic is designed to favor shorter disambiguation text fragments over longer ones (e.g., thus the multiplication by k_a). The heuristic is also designed to favor the prefix over the suffix (e.g., thus the penalty of 0.1 for suffix), to improve readability. Finally, the heuristic is designed to favor disambiguating (e.g., replacing) a longer prefix or suffix, over disambiguating by using a larger number of string fragments (e.g., thus the squaring of the number of string fragments to be replaced.

[0084] As noted above, some embodiments also may support negative mid-span examples as well as negative look-behind examples and negative look-ahead examples.

[0085] Once a prefix/suffix and k (the number of characters to disambiguate) have been determined, the regular expression generator 110 still may determine how to represent that disambiguation in the generated regular expression. The generated regular expression may be either permissive for affixes (e.g., prefixes or suffixes) that look like the positive examples, or may exclude affixes that look like the negative examples.

$$usePermissive = \frac{|\mathbf{E}_p|}{|\mathbf{F}_p|} - \frac{|\mathbf{E}_n|}{|\mathbf{F}_n|}$$

If usePermissive is greater than zero, then things that look like the positive examples are allowed through by generating regular expressions that allows characters, one by one for (each character position), taken from the positive examples. In other cases, the regular expression generator 110 may take the approach of disallowing things that look like the negative examples by generating a regular expression that disallows characters, one by one (for each character position), taken from the negative examples.

[0086] As another example, a generated regular expression for the positive example 8am and negative example 9pm might be `\d[^p]m`. This uses the caret syntax. In some cases, the regular expression generator 110 may be configured to favor a shorter regular expression, which may be not only more readable to the user, but also may be more likely to be correct. The rationale is that a frequently appearing character is more likely to appear again in the future, and so an emphasis should be placed upon frequently appearing characters. If there are fewer unique characters $|\mathbf{F}_p|$ (fewer unique because the ones that do appear do so more frequently) then this is rewarded in the heuristic by having it in the denominator.

[0087] Referring again to the usePermissive example heuristic above, determining one unique positive affix is no big feat if there was only one positive example from the user. Thus, in this heuristic low $|\mathbf{E}_p|$ is penalized by having it in the numerator (i.e. high $|\mathbf{E}_p|$ is rewarded in this heuristic).

[0088] Additionally, in some embodiments, negative examples may be based on look-behind and/or look-ahead. For example, the user may provide a positive example of "323-1234" and a negative example of "202-754-9876" then that involves use of the regex look-behind syntax `(?<!)` to exclude phone numbers with area codes.

[0089] Negative examples also may be based on optional spans in some cases. For example, the user may provide positive examples of "ab" and "a2b" and a negative example of "a3b". In this case, an example implementation may fail, because it may attempt to discriminate based only on required spans and the "2" digit is in an optional span. In this example, failure may refer to a situation in which the generated regular expression matches all of the positive examples

(correctly) and also matches one or more of the negative examples (erroneously). In such cases, the user may alerted to the failure and may be provided the options, via the user interface, to manually repair the generated regular expression and/or to remove some of the negative examples.

V. USER INTERFACE FOR REGULAR EXPRESSION GENERATION

[0090] Additional aspect described herein include several different features and functionality within a graphical user interface related to generation of regular expressions. As discussed below, certain of these features may including various options for user selection and highlighting for positive and negative examples, color-coding for positive and negative examples, and multiple overlapping/nested highlighting within a data cell.

[0091] FIG. 11 is a flowchart illustrating a process for generating regular expressions based on user data selections received within a user interface. The example process in FIG. 11 may correspond to any of the previously discussed examples of generating regular expressions based on input data character sequences. However, FIG. 11 describes the process with respect to the user interface that may be generated and displayed on a client device 120. In step 1101, in response to a request from a user via the user interface, the regular expression generator 110 may retrieve data (e.g., from a data repository 130) and render/display the data in tabular form within a graphical user interface. Although tabular data is used in this example, it should be understood that tabular data need not be used or displayed in other examples. For instance, in some cases a user may type in raw data directly (rather than selecting data from the user interface). Additionally, when data is presented on via the user interface, the data need not be in tabular form, but may be unstructured data (e.g., a document) or semi-structured (e.g., a spreadsheet of unformatted/unstructured data items such as tweets or posts). In various examples, the tabular data may correspond transaction data, customer data, performance data, forecast data, and/or any other categories of data that may be stored in the data repositories 130 for a business or other organization. In step 1102, a user selection of input data may be received via the user interface. The selected input data may, for example, correspond to an entire data cell selected by the user, or a subsequence of characters within a data cell. In step 1103, the regular expression generator 110 may generate a regular expression based on the input data received in step 1102 (e.g., the data cell or portions thereof). In step 1104, the user interface may be updated in response to the

generation of the regular expression. In some cases, the user interface may be updated simply to display the generated regular expression to the user, while in other cases the user interface may be updated in various other ways as discussed below. As shown in this example, the user may select multiple different input data character sequences via the user interface, and in response to each new input data received, the regular expression generator 110 may generate an updated regular expression which encompasses both the first and second (positive) examples of character sequences. Then, when the user highlights a third sequence of characters (e.g., outside of both character sequences, or within the first or second character sequence) the regular expression generator 110 may update the regular expression again, and so on. In some embodiments, the regular expression generator 110 may execute the algorithm in real-time (or near real-time) so that an entirely new regular expression may be generated in response to each new keystroke or each new highlighted section made by the user.

[0092] Thus, as shown in FIG. 11, in response to user selections of character sequences via the user interface, the regular expression generator 110 may generate and display a regular expression. For example, when a user highlights a first sequence of characters, the regular expression generator may generate and display a regular expression representing the first sequence of characters. When the user highlights a second sequence of characters, the regular expression generator may generate an updated regular expression which encompasses both the first and second sequences of characters. Then, when the user highlights a third sequence of characters (e.g., within either the first or second sequence) the regular expression generator may update the regular expression again, and so on.

[0093] FIG. 12 is another flowchart illustrating a process for generating regular expressions and extracting data based on a capture group, via user data selections received within a user interface. In step 1201, as discussed above in step 1101, the regular expression generator 110 may retrieve data (e.g., from a data repository 130) and render/display the data in tabular form within the graphical user interface. In step 1202, the regular expression generator 110 may receive selection of user highlighting of a text fragment within a particular data cell. In step 1203, the regular expression generator 110 may generate a regular expression based on the positive example of the selected data cell, and in step 1204 may create a regular expression capture group based on the text fragment highlighted within the cell. In step 1205, the regular

expression generator 110 may determine one or more additional cells within the displayed tabular data that match the generated regular expression, and in step 1206 the corresponding text fragments within the additional cells that match the generated regular expression may be extracted.

[0094] Thus, in addition to supplying the positive examples, the user also may select (e.g., via mouse text highlighting) a text fragment within any of the selected positive examples. In response, the regular expression generator 110 may create a regular expression capture group to extract that text fragment from the example as well as the corresponding fragment from all other matches in the text the regular expression is applied to. Extracting the text fragments from matching data cells also may include deleting and modifying, and may be used in some cases to create a new column of data out of an existing column of semi-structured or unstructured text.

[0095] Using an example of a user selecting a positive data example, and if the user highlighted the year, then the regular expression generator 110 may generate the regular expression (?:[A-Z]{3}\s+\d\d,\s+|\d\d/\d\d)(\d\d\d\d). As shown in this example, the regular expression generator 110 has put parentheses around the year, and also converted the old parentheses around the month and day (used for alternation) into a "non-capturing" group by use of the ?: regex syntax. In some embodiments, an extraction/capture group may be required to fall on span boundaries, and in such embodiments the regular expression generator 110 may take the highlighted character range as input and expands it to encompass the nearest anchor span boundaries. However, in other examples, the mid-span extraction/capture may be supported by the user interface.

[0096] In some embodiments, the user interface may support input data from uses that includes a selection of a first character sequence within a second character sequence. For instance, a user may highlight one or more characters within a larger previously highlighted character sequence, and the second user selection may provide context for the larger first user selection. Such embodiments may enable input data to be provided to the regular expression generator 110 with greater specificity.

[0097] Additionally, in some examples, an operation may be initiated and a dialog may be opened in response to a user selecting (e.g., highlighting text) within the user interface. In some cases, the dialog may be a non-model dialog, such as floating toolbox window that does not

prevent user interaction with the main screen. The dialog also may change in appearance and/or functionality depending on what major operation the user is performing. Thus, in such cases, the user need not search for a further menu item after highlighting the selected text, in order to initiate the modification, extracting, etc., of the capture group text fragments. Additionally, in certain embodiments, the user interface provided for generating regular expressions may include three highlight modes: nested-auto, nested-manual, and single-level. In certain cases, the default mode of operation may be that the entire cell is identified as the highlighted region, and the user may further highlight one or more additional subsequences within the highlighted cell. In other modes, the user may be allowed to manually specify both highlights within a data cell of the tabular data display. In still other modes, the user may be allowed to manually specify an outer highlight with no inner highlight. These other modes may be better suited to "semi-structured" data, for example, a column of data consisting of tweets or other long strings such as browser "user agent" strings. "Semi-structured" data refers to data that may be displayed in tabular form within the user interface, but where a column within the table consists of unstructured text.

[0098] In some such embodiments, inner and outer selection (e.g., highlighting) by the user via the user interface may be distinguished by color coding. For example, the outer highlights of a positive example may be shown in a first text/background color combination, and the inner highlight of a positive example may be shown in a different contrasting text/background color combination.

[0099] As indicated above, a user may specify a selection of a capture group via selection of a character subsequence. The GUI may be used to facilitate user selection via highlighting (or other indications). An example is shown in FIG. 13, in which an example user interface screen is shown with a tabular data display. In this example, FIG. 13 depicts highlighting within a column value, for example, caused by a user dragging a mouse across one or more desired elements of the column value. Note that the "cell" in which the user highlighting is performed may exhibit a color change indicating selection of the column value. This color change may be construed as automated highlighting responsive to the user highlighting.

[0100] FIGS. 14 and 15 are example user interface screens illustrating the generation of regular expressions and capture groups based on selection of data from a tabular display. In these examples, FIGS. 14 and 15 show an additional user interface window that be displayed

automatically detection of user highlighting 1401 within the tabular data display. The window comprises a field 1402 for displaying positive examples, a field for displaying negative examples, and a field for displaying the regular expression that is generated dynamically (and near-instantaneously) in response to the selection of positive examples form the tabular data display. In these example, user highlighting within a column value 1401 may be equivalent to user highlighting within automated highlighting. Thus, user highlighting of the area code causes not only the user-highlighted area code 1401, but also the rest of the phone number to be populated in the positive example field 1402.

[0101] However, it should be appreciated that user highlighting is not limited to performance within automated highlighting. For example, user highlighting may alternatively be performed within other user highlighting. As another example, user highlighting may alternatively be performed without any inner highlighting (e.g., further highlighting within highlighted text). These alternative examples are particularly suitable for semi-structured data, such as a column of data comprising "Tweets" or other long strings (e.g., browser "user agent" strings).

[0102] Furthermore, upon generation of the corresponding regular expression, other column values 1402 matching the regular expression may be identified based upon additional automated highlighting. In the examples shown in FIGS. 14 and 15, the additional automated highlighting indicates the elements of these other column values that match the capture group of the generated regular expression. The additional automated highlighting may be performed using a color that is different from the one used for the user highlighting.

[0103] As shown in FIG. 15, additional user highlighting is shown to indicate user selection of other examples. The additional user highlighting may be performed in a manner similar to that described above. Thus, the user interface in FIG. 15 shows the population of other examples in the field 1502 for displaying positive examples. This may occur responsive to detection of the additional user highlighting. Additionally, the generated regular expression 1503 may be updated dynamically and near-instantaneously, such that it matches all of the positive examples 1502. Responsive to generation of the updated regular expression , automated highlighting of other column values 1504 matching the updated regular expression may also be updated. In some implementations, dynamic color-coding also may be used. For instance, matches may be color-coded using a first color (e.g., blue), while positive examples are color-coded using a

second color (e.g., green), and negative examples may be color-coded using a third color (e.g., red).

[0104] FIGS. 16A and 16B are example user interface screens illustrating the generation of regular expressions based on selection of positive and negative examples from the tabular display. In FIGS. 16A-16B, individual examples from the positive examples field 1602 may be removed from the positive examples field 1603, and/or moved to the negative examples field 1603. Within the user interface, this may be performed, for example, by the user clicking (e.g., right-clicking) on one of the examples to selecting it. The selection may cause the user interface to display a menu 1602 comprising a delete option and a change option. Thereafter, clicking on an option may cause performance of the corresponding function.

[0105] In the example shown in FIGS. 16A and 16B, the result of the user selection of the change option, is to move the selected example is moved to the negative examples field 1603, causing the regular expression 1601 to be updated to regular expression 1604, which may be generated dynamically and near-instantaneously (e.g., between 30ms and 9000 ms in certain embodiments). Responsive to generation of the updated regular expression 1604, the automated highlighting of other column values matching the updated regular expression may also be updated within the tabular data display. Furthermore, automated highlighting may be performed on some or all of the negative examples, including any column values corresponding to the negative example, which may be highlighted using a color that is different from any of the colors used above, or otherwise distinguished within the user interface using other visual techniques.

[0106] In some embodiments, specifying a negative example via the user interface need not require first specifying the example as a positive example, and then converting it into a negative example as shown in FIGS. 16A and 16B. Rather, a negative example may be specified in a variety of ways. For example, a user may select (e.g., right click) a column value via the user interface (e.g., one of the other column values on which automated highlighting was performed to indicate that it matches the generated regular expression), which may thereby cause display of a menu comprising an option (e.g., “Make New Counterexample”) to designate the selected column value as a negative example.

[0107] Thus, using the examples shown in FIGS. 16A and 16B, responsive to generation of the updated regular expression 1604, automated highlighting of other column values matching the

updated regular expression may also be updated. In these examples, the updated regular expression specifies telephone numbers that end "9".

[0108] Returning briefly to FIGS. 14 and 15, when the "Extract" button is clicked or otherwise selected by the user, an operation may be initiated to extract the highlighted text fragments within all of the cells matching the current regular expression 1403 or 1503. Although not shown in FIGS. 14 and 15, in some embodiments the user interface may provide other selectable buttons in addition to or instead of the "Extract" button. For example, a "Replace" button may be presented as an option to replace user-highlighted elements with user-specified elements. Additionally or alternatively, one or more "Delete" buttons may be presented as an option to, in effect, replace user-highlighted elements with nothing. For instance, one or both of a "Delete Fragment" operation and/or a "Delete Row" operation may be implemented, which will delete either the user-highlighted text fragment or the either row, respectively. Additional operations that may be implemented in various embodiments may include a "Keep Row" operation, a "Split" operation (e.g., highlight comma, then extract the comma-separated components into separate multiple new columns), and an "Obfuscate" operation (e.g., replace highlighted text/capture group with a sequence of "#" or other symbols).

In this example, in response to the selection of the "Extract" button, an extraction operation may be added to a list of transform scripts to be performed by a downstream operation. In some embodiments, the list of transform scripts may be displayed in a portion of the user interface for review/modification by the user. Alternatively, the extraction operation may be performed *in situ* to generate a new column comprising the contents of the regex capture group (e.g., the elements corresponding to the user-highlighted portions of a positive example). In the examples shown in FIGS. 14 and 15, a new column and/or a new table of area codes may be generated in response to a selection of the "Extract" button.

[0109] FIG. 17 is another example user interface screen illustrating the generation of a regular expression and capture group based on selection of data from a tabular display, according to one or more embodiments described herein.

VI. REGULAR EXPRESSION GENERATION USING LONGEST COMMON SUBSEQUENCE ALGORITHM ON SPANS

[0110] Additional aspects described herein relate to the generation of regular expressions, based on the LCS algorithm from one or more data input character sequences, but wherein the regular expression generator 110 also may handle characters that are present in only some of the examples. To handle characters that are present in only some input examples, spans may be defined in which both a minimum and maximum number of occurrences of a regular expression code are tracked. For example, for the character sequence inputs of "9pm" and "9 pm" an optional space is present between the number and the "pm" text. In such cases, when a certain span (e.g., the single space between "9" and "pm") might not be present at all of the given input examples, the minimum number of occurrences may be set to zero. These minimum and maximum numbers can then be mapped to the regular expression multiplicity syntax. A longest common subsequence (LCS) algorithm may be run on the spans of characters derived from the input examples, including "optional" spans (e.g., minimum length of zero) which do not appear in every input example. As discussed below, consecutive spans may be merged during the execution of the LCS algorithm. In such cases, when extra optional spans that are being carried along end up appearing consecutively, the LCS algorithm may be run recursively on those optional spans as well. That is, although the running of the LCS algorithm is by its nature recursive, in these cases the entire LCS algorithm may be run recursively (e.g., recursively running the recursive LCS algorithm). Among other technical advantages, this may allow for a shorter, cleaner, and more readable regular expression generation. For instance, (am| am) (i.e., with optional space before the am) might be generated without recursively running the LCS algorithm, whereas recursively running the LCS algorithm may result in the regular expression generated as (?am), which is shorter and cleaner.

[0111] FIG. 18 is a flowchart illustrating a process for generating regular expressions, including optional spans, using a longest common subsequence (LCS) algorithm, according to one or more embodiments described herein. In step 1801, the regular expression generator 110 may receive one or more character sequences as input data, corresponding to positive regular expression examples. In step 1802, the regular expression generator 110 may convert the character sequences into regular expression codes. Thus, steps 1801 and 1802 may be similar or

identical to previous corresponding examples discussed above. Then, in step 1802, the regular expression codes may further by converted into span data structures (or spans). As noted above, each span may include a data structure storing a character class code (e.g., a regex code) and a repetition count range (e.g., a minimum count and/or a maximum count). In step 1804, the regular expression generator 110 may execute an LCS algorithm, providing the sets of spans as input to the algorithm. The output of the LCS algorithm in this example may include an output set of spans, including at least one span having a minimum repetition count range equal to zero, which corresponds to an optional span within the output of the LCS algorithm. Finally, in step 1805, the regular expression generator 110 may generate a regular expression based on the output of the output of the LCS algorithm, including the optional span.

[0112] FIG. 19 is an example diagram illustrating the generation of a regular expression using a longest common subsequence (LCS) algorithm, wherein the generated regular expression includes an optional span. In this example, the two input data character sequences are "8am" and "9 pm". The input data character sequences are first converted to regular expression codes (step 1802) and then to spans (step 1803), as discussed above. The spans may be provided as input to an LCS algorithm (step 1804), and the LCS output includes the optional span $Z^{<0,1>}$, indicating that an optional single space may be number and the two-letter text sequence. That is, the superscript notation in this example may include the two numbers, the minimum repetition count range (e.g., 0), and the maximum repetition count range (e.g., 1) which apply to the preceding code (e.g., Z = spaces). Finally, the regular expression may be generated based on the output span of the LCS algorithm, and the optional span may be converted to the corresponding regular expression code "pZ^{*}".

[0113] In some embodiments, the rendition and use of optional space by the regular expression generator 110, during the execution of the LCS algorithm, may provide additional technical advantages with respect to performance and readability. For example, when generating regular expressions, it is desirable in some cases to be able to handle both the characters that are in common amongst all the given examples, and the characters that are present in only some of the examples.

[0114] In certain embodiments, for each span data structure, both the minimum number of occurrences of a category code and a maximum number of occurrences of the category code may

be tracked. In the case where a span is not present at all in one or more of the given examples, the minimum is set to zero. As another example, to generate a regular expression to handle months of the year spelled out, minimum and maximum numbers may then be mapped to the regular expression multiplicity syntax involving curly braces (e.g., `[A-Za-z]{3,9}`).

[0115] In some embodiments, the regular expression generator 110 may keep track of minimum and maximum number of occurrences for each span, but also may handle additional implementation details. For example, as a result of the combination of handling optional spans and running LCS on spans of characters, the regular expression generator 110 may be configured to detect and merge consecutive spans throughout the execution of the LCS algorithm. Additionally, the any extra optional spans being carried along sometimes appearing consecutively , and it may be desirable for the LCS algorithm to be run on those recursively as well. For example, in some cases, the regular expression generator 110 modify and/or extend the LCS algorithm to favor (or weight) fewer transitions between optional and required sequence elements (e.g., spans). For example, grouping optional spans together may minimize the number of grouping parentheses that have to be used within the regular expression, which may thus improve the human readability of the generated regular expression. In some cases, if the resultant lengths are equal even after considering optional spans, then the regular expression generator 110 may exhibit a preference for the alternative with fewer transitions between optional and required spans. For example, in some cases a standard LCS algorithm may be implemented to prefer the choice of longer sequences at its decision points. However, at decision points where the options are of equal length, a configuration preference may be programmed into the regular expression generator 110. One such configuration preferecne may be, for example, is to prefer shorter sequences (once optional spans are considered). Thus, the customized LCS within this configuration may simultaneously optimize for longer sequences (of required spans) and shorter sequences (of total required and optional spans).

[0116] In some embodiments, generated regular expressions may be more readable if they begin with a required span (which may also serve as a mental anchor to a human reader), rather than starting the regular expressions with optional spans. Thus, in some cases, if the resultant options have equal numbers of transitions, then the option with earlier non-optional spans may be chosen. Additionally, the LCS algorithm executed by the regular expression generator 110

may be configured in some embodiments to push all spaces (including optional spans corresponding to spaces) to the right within the regular expression. By pushing all the spaces to the right, there may be an increased chance that spans of spaces may be merged together, which may simplify the resulting regular expression as well as improving readability. Thus, during the execution of the LCS algorithm, when a determination is made that two sets of substrings have the same LCS, instead of arbitrarily selecting one of the two sets, the set that facilitates improved readability may be selected. Further, in some embodiments, the LCS algorithm may be configured to favor a greater number of required spans, and/or fewer optional spans, in order to improve readability.

[0117] As noted above, negative examples also may be based on optional spans in some cases. For example, the user may provide positive examples of "ab" and "a2b" and a negative example of "a3b". In this case, an example implementation may fail, because it may attempt to discriminate based only on required spans and the "2" digit is in an optional span. In such cases, the user may alerted to the failure and may be provided the options, via the user interface, to manually repair the generated regular expression and/or to remove some of the negative examples.

[0118] In some embodiments, there may be an isSuccess returned as part of the JSON coming back from the REST service. In some embodiments, the generated regex may become a different color (e.g., red) when isSuccess=false.

VII. REGULAR EXPRESSION GENERATION USING COMBINATORIC LONGEST COMMON SUBSEQUENCE ALGORITHMS

[0119] Further aspects described herein relate to a combinatoric search, in which the LCS algorithm executed by the regular expression generator 110 may be run multiple times to generate a "correct" regular expression (e.g., a regular expression that properly matches all given positive examples and properly excludes all given negative examples), and/or to generate multiple correct regular expressions from which a most desirable or optimal regular expression may be selected. For example, during a combinatoric search, the full LCS algorithm and regular expression generation process may be run multiple times, including different

combinations/permutations of text processing directions, different anchoring, and other different characteristics of the LCS algorithm.

[0120] FIG. 20 is a flowchart illustrating a process for generating regular expressions based on combinatoric executions of a longest common subsequence (LCS) algorithm. In step 2001, the regular expression generator 110 may receive input data character sequences corresponding to positive examples. In step 2002, the regular expression generator 110 may iterate over various different combinations of execution techniques for the LCS algorithm. As shown in this examples, during each iteration of steps 2002, the regular expression generator 110 may select a different combination of the following LCS algorithm execution parameters (or characteristics): anchor (i.e., no anchoring, anchoring to the beginning of the line, anchoring to the end of the line), processing direction (i.e., right-to-left order, left-to-right order), push space (i.e., do or do not push spaces), and collapse spans (i.e., do or not collapse spans). In step 2003, the LCS algorithm is run on the input data character sequences (or on regular expression codes if the input character sequences were converted first), wherein the LCS algorithm is configured based on the parameters/characteristics selected in step 2002. In step 2004, the output of the LCS algorithm of may be stored by the regular expression generator 110, include data such as whether or not an LCS was successfully identified by the algorithm, and the length of the corresponding regular expression. In step 2005, the process may iterate until the LCS algorithm has been run with all possible combinations of the parameters/characteristics of the combinatoric search. Finally, in step 2006, a particular output from one of the LCS is selected as an optimal output (e.g., based on success and regular expression length), and a regular expression may be generated based on the selected LCS algorithm output.

[0121] In various embodiments, a combinatoric search such as that described above in reference to FIG. 20, may be performed for various different combinations of parameters/characteristics. For example, in some embodiments an LCS algorithm may use the caret symbol ^ to anchor the regular expression to the beginning of the text, and/or the dollar symbol $ to anchor the regular expression to the end of the text. In some cases, such anchoring may result in generating a shorter regular expression. Anchors may be particularly useful when a user wishes to find a particular pattern at the beginning and/or at the end of a string. For example, a user may want a product name at the beginning. To avoid confusing the LCS

algorithm with the varying number of words describing the product name, a caret may be used to anchor the regex to the beginning of a string as depicted in the image below.

[0122] Additionally, in some embodiments, the LCS algorithm may be executed with input data that is either forward or reversed (or similarly the LCS algorithm may be configured to receive the input data in the usual order and then reverse the order before executing the algorithm). Thus, in some embodiments, a combinatoric search of LCS algorithms that may be performed on a pair of input character sequences or codes may be:

1. Usual (right-to-left) order, no anchoring to start or end
2. Usual (right-to-left) order, anchoring to beginning of line using caret ^
3. Usual (right-to-left) order, anchoring to end of line using dollar $
4. Reverse (left-to-right) order, no anchoring to start or end
5. Reverse (left-to-right) order, anchoring to beginning of line using caret ^
6. Reverse (left-to-right) order, anchoring to end of line using dollar $

In this example, out of the six executions of the LCS, the shortest resulting regular expression may be selected (step 2006).

[0123] In some embodiments, the combinatoric search of the LCS algorithm also may iterate over a greedy quantifier "?" and non-greedy quantifier "??" . For example, by default if there is an optional span a single question mark is emitted, e.g., [A-Z]+(?: [A-Z]\.)? [A-Z]+ for first and last name with optional middle initial. If a satisfactory regular expression cannot be found when using the greedy quantifier, then the combinatoric search may attempt to replace all question mark quantifiers with double-question mark quantifiers (e.g., [A-Z]+(?: [A-Z]\.)?? [A-Z]+). The double question mark corresponds to a non-greedy quantifier, which may instruct a downstream regular expression matcher to go into backtracking mode in order to find a match.

[0124] Additionally, in some embodiments, the combinatoric search of the LCS algorithm also may iterate over whether or not to prefer spaces on the right. For example, as noted above, a strategy may be used in some embodiments of pushing spaces to the right, e.g., when the LCS algorithm is faced with an arbitrary choice of otherwise equal options, in the hope that space

spans may get merged together, resulting in a fewer number of overall spans. This feature adds another option to the combinatoric search, that is, to either push spaces to the right or execute in accordance with a traditional LCS approach of leaving the decision to be arbitrary.

[0125] Further, in some embodiments, the combinatoric search of LCS algorithm also may iterative over scanning/not scanning for literals common among all the examples, by running LCS on the original strings. In such embodiments, the LCS algorithm may be configured to identify and align on common words. As used herein, a "common word" may refer to a word that appears in every positive example. Once a common word is identified, its span type may be converted from LETTER to WORD, and the subsequent run through the LCS algorithm may then naturally aligns on it.

[0126] Thus, in the example below, a combinatoric search may iterate over several parameters/characteristics to reach 96 times that the complete LCS algorithm is to be performed. The various parameters/characteristics to be iterated over in this example are:

- Anchor (3) (Values = ^, $, or neither)
- Pushing Spaces (2) (Values = Yes or No)
- Coalescing Low Cardinality Spans to Wildcards (2) (Values = Yes or No)
- Greedy Quantifier ? (2) (Values = Yes or No)
- Aligning the LCS Algorithm on Common Tokens (2) (Values = Yes or No)
- Using "\w" to Represent Alphanumeric, Versus Keeping Letters "\pL" and Numbers "\pN" Treated as Separate Spans (2) (Values = Yes or No)

As noted above, in this example, the complete LCS algorithm is to be performed 96 times (e.g., 3 * 2 * 2 * 2 * 2 * 2 = 96).

[0127] However, in other embodiments, the regular expression generator 110 may provide a performance enhancement, by which only the first three characteristics in the above list (Anchor, Pushing Spaces, and Coalescing Low Cardinality Spans to Wildcard, may participate in the combinatoric search. This may result in a far fewer number of complete LCS algorithm is to be performed (e.g., 3 * 2 * 2 = 12 times). In such embodiments, while the last three characteristics

in the above list (Greedy Quantifier, Aligning the LCS Algorithm on Common Tokens, and Using "\w" to Represent Alphanumeric, Versus Keeping Letters "\pL" and Numbers "\pN" Treated as Separate Spans) do not participate in the combinatoric search, these characteristics may be tested at the end, individually and serially. Technical advantages may be realized in such embodiments, because dividing the search space in this manner may still resulted in a satisfactory regular expression being found, but with approximately an 8x speedup in performance.

[0128] To illustrate, the following example of a combinatoric search may provide a performance advantage over the previous example. In this example, the combinatoric search may be performed based on the following parameters/characteristics to be iterated over:

- Anchoring (3): BEGINNING_OF_LINE_MODE, END_OF_LINE_MODE, NO_EOL_MODE
- Order/Direction (2): Right-to-left (normal) LCS vs. Left-to-right (reverse) LCS
- Push (2): Whether or not to try to push spaces to the right within the LCS algorithm
- Compress to Wildcards (2): Whether or not to try to compress long sequences of only-sometimes occurring spans down to the wildcards .*?

[0129] The combinatoric in this example may result in running the complete algorithm 3 * 2 * 2 * 2 = 24 times). The regular expression generator 110 then may take the best of the 24 results of the LCS algorithm, where "best" may means that (a) the LCS algorithm succeeded, and (b) the shortest regular expression was generated. The regular expression generator 110 then may perform the following three additional tasks:

1. Try condensing sequences of letters and numbers that are unbroken by spaces, punctuation, or symbols, down to a new span type I called ALPHANUMERIC, corresponding to generated regex of \w. This may be useful for hexadecimal numbers as found in IPv6 addresses from clickstream logs (refer to novelty 64 from April 2019).
2. Try using the non-greedy quantifier ?? instead of the greedy quantifier ?
3. Try aligning on literals

VIII. HARDWARE OVERVIEW

[0130] FIG. 21 depicts a simplified diagram of a distributed system 2100 for implementing an embodiment. In the illustrated embodiment, distributed system 2100 includes one or more client computing devices 2102, 2104, 2106, and 2108, coupled to a server 2112 via one or more communication networks 2110. Clients computing devices 2102, 2104, 2106, and 2108 may be configured to execute one or more applications.

[0131] In various embodiments, server 2112 may be adapted to run one or more services or software applications that enable automated generation of regular expressions, as described in this disclosure. For example, in certain embodiments, server 2112 may receive user input data transmitted from a client device, where the user input data is received by the client device through a user interface displayed at the client device. Server 2112 may then convert the user input data into a regular expression that is transmitted to the client device for display through the user interface.

[0132] In certain embodiments, server 2112 may also provide other services or software applications that can include non-virtual and virtual environments. In some embodiments, these services may be offered as web-based or cloud services, such as under a Software as a Service (SaaS) model to the users of client computing devices 2102, 2104, 2106, and/or 2108. Users operating client computing devices 2102, 2104, 2106, and/or 2108 may in turn utilize one or more client applications to interact with server 2112 to utilize the services provided by these components.

[0133] In the configuration depicted in FIG. 21, server 2112 may include one or more components 2118, 2120 and 2122 that implement the functions performed by server 2112. These components may include software components that may be executed by one or more processors, hardware components, or combinations thereof. It should be appreciated that various different system configurations are possible, which may be different from distributed system 2100. The embodiment shown in FIG. 21 is thus one example of a distributed system for implementing an embodiment system and is not intended to be limiting.

[0134] Users may use client computing devices 2102, 2104, 2106, and/or 2108 to execute one or more applications, which may generate regular expressions in accordance with the teachings

of this disclosure. A client device may provide an interface that enables a user of the client device to interact with the client device. The client device may also output information to the user via this interface. Although FIG. 21 depicts only four client computing devices, any number of client computing devices may be supported.

[0135] The client devices may include various types of computing systems such as portable handheld devices, general purpose computers such as personal computers and laptops, workstation computers, wearable devices, gaming systems, thin clients, various messaging devices, sensors or other sensing devices, and the like. These computing devices may run various types and versions of software applications and operating systems (e.g., Microsoft Windows®, Apple Macintosh®, UNIX® or UNIX-like operating systems, Linux or Linux-like operating systems such as Google Chrome™ OS) including various mobile operating systems (e.g., Microsoft Windows Mobile®, iOS®, Windows Phone®, Android™, BlackBerry®, Palm OS®). Portable handheld devices may include cellular phones, smartphones, (e.g., an iPhone®), tablets (e.g., iPad®), personal digital assistants (PDAs), and the like. Wearable devices may include Google Glass® head mounted display, and other devices. Gaming systems may include various handheld gaming devices, Internet-enabled gaming devices (e.g., a Microsoft Xbox® gaming console with or without a Kinect® gesture input device, Sony PlayStation® system, various gaming systems provided by Nintendo®, and others), and the like. The client devices may be capable of executing various different applications such as various Internet-related apps, communication applications (e.g., E-mail applications, short message service (SMS) applications) and may use various communication protocols.

[0136] Network(s) 2110 may be any type of network familiar to those skilled in the art that can support data communications using any of a variety of available protocols, including without limitation TCP/IP (transmission control protocol/Internet protocol), SNA (systems network architecture), IPX (Internet packet exchange), AppleTalk®, and the like. Merely by way of example, network(s) 2110 can be a local area network (LAN), networks based on Ethernet, Token-Ring, a wide-area network (WAN), the Internet, a virtual network, a virtual private network (VPN), an intranet, an extranet, a public switched telephone network (PSTN), an infra-red network, a wireless network (e.g., a network operating under any of the Institute of Electrical

and Electronics (IEEE) 1002.11 suite of protocols, Bluetooth®, and/or any other wireless protocol), and/or any combination of these and/or other networks.

[0137] Server 2112 may be composed of one or more general purpose computers, specialized server computers (including, by way of example, PC (personal computer) servers, UNIX® servers, mid-range servers, mainframe computers, rack-mounted servers, etc.), server farms, server clusters, or any other appropriate arrangement and/or combination. Server 2112 can include one or more virtual machines running virtual operating systems, or other computing architectures involving virtualization such as one or more flexible pools of logical storage devices that can be virtualized to maintain virtual storage devices for the server. In various embodiments, server 2112 may be adapted to run one or more services or software applications that provide the functionality described in the foregoing disclosure.

[0138] The computing systems in server 2112 may run one or more operating systems including any of those discussed above, as well as any commercially available server operating system. Server 2112 may also run any of a variety of additional server applications and/or mid-tier applications, including HTTP (hypertext transport protocol) servers, FTP (file transfer protocol) servers, CGI (common gateway interface) servers, JAVA® servers, database servers, and the like. Exemplary database servers include without limitation those commercially available from Oracle®, Microsoft®, Sybase®, IBM® (International Business Machines), and the like.

[0139] In some implementations, server 2112 may include one or more applications to analyze and consolidate data feeds and/or event updates received from users of client computing devices 2102, 2104, 2106, and 2108. As an example, data feeds and/or event updates may include, but are not limited to, Twitter® feeds, Facebook® updates or real-time updates received from one or more third party information sources and continuous data streams, which may include real-time events related to sensor data applications, financial tickers, network performance measuring tools (e.g., network monitoring and traffic management applications), clickstream analysis tools, automobile traffic monitoring, and the like. Server 2112 may also include one or more applications to display the data feeds and/or real-time events via one or more display devices of client computing devices 2102, 2104, 2106, and 2108.

[0140] Distributed system 2100 may also include one or more data repositories 2114, 2116. These data repositories may be used to store data and other information in certain embodiments. For example, one or more of the data repositories 2114, 2116 may be used to store information such as a new column of data that matches a system-generated regular expression. Data repositories 2114, 2116 may reside in a variety of locations. For example, a data repository used by server 2112 may be local to server 2112 or may be remote from server 2112 and in communication with server 2112 via a network-based or dedicated connection. Data repositories 2114, 2116 may be of different types. In certain embodiments, a data repository used by server 2112 may be a database, for example, a relational database, such as databases provided by Oracle Corporation® and other vendors. One or more of these databases may be adapted to enable storage, update, and retrieval of data to and from the database in response to SQL-formatted commands.

[0141] In certain embodiments, one or more of data repositories 2114, 2116 may also be used by applications to store application data. The data repositories used by applications may be of different types such as, for example, a key-value store repository, an object store repository, or a general storage repository supported by a file system.

[0142] In certain embodiments, the functionalities described in this disclosure may be offered as services via a cloud environment. FIG. 22 is a simplified block diagram of a cloud-based system environment in which various services may be offered as cloud services, in accordance with certain examples. In the example depicted in FIG. 22, cloud infrastructure system 2202 may provide one or more cloud services that may be requested by users using one or more client computing devices 2204, 2206, and 2208. Cloud infrastructure system 2202 may comprise one or more computers and/or servers that may include those described above for server 2112. The computers in cloud infrastructure system 2202 may be organized as general purpose computers, specialized server computers, server farms, server clusters, or any other appropriate arrangement and/or combination.

[0143] Network(s) 2210 may facilitate communication and exchange of data between clients 2204, 2206, and 2208 and cloud infrastructure system 2202. Network(s) 2210 may include one or more networks. The networks may be of the same or different types. Network(s) 2210 may

support one or more communication protocols, including wired and/or wireless protocols, for facilitating the communications.

[0144] The example depicted in FIG. 22 is only one example of a cloud infrastructure system and is not intended to be limiting. It should be appreciated that, in some other examples, cloud infrastructure system 2202 may have more or fewer components than those depicted in FIG. 22, may combine two or more components, or may have a different configuration or arrangement of components. For example, although FIG. 22 depicts three client computing devices, any number of client computing devices may be supported in alternative examples.

[0145] The term cloud service is generally used to refer to a service that is made available to users on demand and via a communication network such as the Internet by systems (e.g., cloud infrastructure system 2202) of a service provider. Typically, in a public cloud environment, servers and systems that make up the cloud service provider's system are different from the customer's own on-premise servers and systems. The cloud service provider's systems are managed by the cloud service provider. Customers may thus avail themselves of cloud services provided by a cloud service provider without having to purchase separate licenses, support, or hardware and software resources for the services. For example, a cloud service provider's system may host an application, and a user may, via the Internet, on demand, order and use the application without the user having to buy infrastructure resources for executing the application. Cloud services are designed to provide easy, scalable access to applications, resources and services. Several providers offer cloud services. For example, several cloud services are offered by Oracle Corporation® of Redwood Shores, California, such as middleware services, database services, Java cloud services, and others.

[0146] In certain embodiments, cloud infrastructure system 2202 may provide one or more cloud services using different models such as under a Software as a Service (SaaS) model, a Platform as a Service (PaaS) model, an Infrastructure as a Service (IaaS) model, and others, including hybrid service models. Cloud infrastructure system 2202 may include a suite of applications, middleware, databases, and other resources that enable provision of the various cloud services.

[0147] A SaaS model enables an application or software to be delivered to a customer over a communication network like the Internet, as a service, without the customer having to buy the

hardware or software for the underlying application. For example, a SaaS model may be used to provide customers access to on-demand applications that are hosted by cloud infrastructure system 2202. Examples of SaaS services provided by Oracle Corporation® include, without limitation, various services for human resources/capital management, customer relationship management (CRM), enterprise resource planning (ERP), supply chain management (SCM), enterprise performance management (EPM), analytics services, social applications, and others.

[0148] An IaaS model is generally used to provide infrastructure resources (e.g., servers, storage, hardware and networking resources) to a customer as a cloud service to provide elastic compute and storage capabilities. Various IaaS services are provided by Oracle Corporation®.

[0149] A PaaS model is generally used to provide, as a service, platform and environment resources that enable customers to develop, run, and manage applications and services without the customer having to procure, build, or maintain such resources. Examples of PaaS services provided by Oracle Corporation® include, without limitation, Oracle Java Cloud Service (JCS), Oracle Database Cloud Service (DBCS), data management cloud service, various application development solutions services, and others.

[0150] Cloud services are generally provided on an on-demand self-service basis, subscription-based, elastically scalable, reliable, highly available, and secure manner. For example, a customer, via a subscription order, may order one or more services provided by cloud infrastructure system 2202. Cloud infrastructure system 2202 then performs processing to provide the services requested in the customer's subscription order. Cloud infrastructure system 2202 may be configured to provide one or more cloud services.

[0151] Cloud infrastructure system 2202 may provide the cloud services via different deployment models. In a public cloud model, cloud infrastructure system 2202 may be owned by a third party cloud services provider and the cloud services are offered to any general public customer, where the customer may be an individual or an enterprise. Under a private cloud model, cloud infrastructure system 2202 may be operated within an organization (e.g., within an enterprise organization) and services provided to customers that are within the organization. For example, the customers may be various departments of an enterprise such as the Human Resources department, the Payroll department, etc. or even individuals within the enterprise. Under a community cloud model, the cloud infrastructure system 2202 and the services provided

may be shared by several organizations in a related community. Various other models such as hybrids of the above mentioned models may also be used.

[0152] Client computing devices 2204, 2206, and 2208 may be of different types (such as devices 2102, 2104, 2106, and 2108 depicted in FIG. 21) and may be capable of operating one or more client applications. A user may use a client device to interact with cloud infrastructure system 2202, such as to request a service provided by cloud infrastructure system 2202.

[0153] In some embodiments, the processing performed by cloud infrastructure system 2202 for providing management-related services may involve big data analysis. This analysis may involve using, analyzing, and manipulating large data sets to detect and visualize various trends, behaviors, relationships, etc. within the data. This analysis may be performed by one or more processors, possibly processing the data in parallel, performing simulations using the data, and the like. For example, big data analysis may be performed by cloud infrastructure system 2202 for determining regular expressions in an automated manner. The data used for this analysis may include structured data (e.g., data stored in a database or structured according to a structured model) and/or unstructured data (e.g., data blobs (binary large objects)).

[0154] As depicted in the example in FIG. 22, cloud infrastructure system 2202 may include infrastructure resources 2230 that are utilized for facilitating the provision of various cloud services offered by cloud infrastructure system 2202. Infrastructure resources 2230 may include, for example, processing resources, storage or memory resources, networking resources, and the like.

[0155] In certain embodiments, to facilitate efficient provisioning of these resources for supporting the various cloud services provided by cloud infrastructure system 2202 for different customers, the resources may be bundled into sets of resources or resource modules (also referred to as "pods"). Each resource module or pod may comprise a pre-integrated and optimized combination of resources of one or more types. In certain embodiments, different pods may be pre-provisioned for different types of cloud services. For example, a first set of pods may be provisioned for a database service, a second set of pods, which may include a different combination of resources than a pod in the first set of pods, may be provisioned for Java service, and the like. For some services, the resources allocated for provisioning the services may be shared between the services.

[0156] Cloud infrastructure system 2202 may itself internally use services 2232 that are shared by different components of cloud infrastructure system 2202 and which facilitate the provisioning of services by cloud infrastructure system 2202. These internal shared services may include, without limitation, a security and identity service, an integration service, an enterprise repository service, an enterprise manager service, a virus scanning and white list service, a high availability, backup and recovery service, service for enabling cloud support, an email service, a notification service, a file transfer service, and the like.

[0157] Cloud infrastructure system 2202 may comprise multiple subsystems. These subsystems may be implemented in software, or hardware, or combinations thereof. As depicted in FIG. 22, the subsystems may include a user interface subsystem 2212 that enables users or customers of cloud infrastructure system 2202 to interact with cloud infrastructure system 2202. User interface subsystem 2212 may include various different interfaces such as a web interface 2214, an online store interface 2216 where cloud services provided by cloud infrastructure system 2202 are advertised and are purchasable by a consumer, and other interfaces 2218. For example, a customer may, using a client device, request (service request 2234) one or more services provided by cloud infrastructure system 2202 using one or more of interfaces 2214, 2216, and 2218. For example, a customer may access the online store, browse cloud services offered by cloud infrastructure system 2202, and place a subscription order for one or more services offered by cloud infrastructure system 2202 that the customer wishes to subscribe to. The service request may include information identifying the customer and one or more services that the customer desires to subscribe to. For example, a customer may place a subscription order for an automated-generation-of-regular-expressions-related service offered by cloud infrastructure system 2202.

[0158] In certain embodiments, such as the example depicted in FIG. 22, cloud infrastructure system 2202 may comprise an order management subsystem (OMS) 2220 that is configured to process the new order. As part of this processing, OMS 2220 may be configured to: create an account for the customer, if not done already; receive billing and/or accounting information from the customer that is to be used for billing the customer for providing the requested service to the customer; verify the customer information; upon verification, book the order for the customer; and orchestrate various workflows to prepare the order for provisioning.

[0159] Once properly validated, OMS 2220 may then invoke the order provisioning subsystem (OPS) 2224 that is configured to provision resources for the order including processing, memory, and networking resources. The provisioning may include allocating resources for the order and configuring the resources to facilitate the service requested by the customer order. The manner in which resources are provisioned for an order and the type of the provisioned resources may depend upon the type of cloud service that has been ordered by the customer. For example, according to one workflow, OPS 2224 may be configured to determine the particular cloud service being requested and identify a number of pods that may have been pre-configured for that particular cloud service. The number of pods that are allocated for an order may depend upon the size/amount/level/scope of the requested service. For example, the number of pods to be allocated may be determined based upon the number of users to be supported by the service, the duration of time for which the service is being requested, and the like. The allocated pods may then be customized for the particular requesting customer for providing the requested service.

[0160] Cloud infrastructure system 2202 may send a response or notification 2244 to the requesting customer to indicate when the requested service is now ready for use. In some instances, information (e.g., a link) may be sent to the customer that enables the customer to start using and availing the benefits of the requested services. In certain embodiments, for a customer requesting the automated-generation-of-regular-expressions-related service, the response may include instructions which, when executed, cause display of a user interface.

[0161] Cloud infrastructure system 2202 may provide services to multiple customers. For each customer, cloud infrastructure system 2202 is responsible for managing information related to one or more subscription orders received from the customer, maintaining customer data related to the orders, and providing the requested services to the customer. Cloud infrastructure system 2202 may also collect usage statistics regarding a customer's use of subscribed services. For example, statistics may be collected for the amount of storage used, the amount of data transferred, the number of users, and the amount of system up time and system down time, and the like. This usage information may be used to bill the customer. Billing may be done, for example, on a monthly cycle.

[0162] Cloud infrastructure system 2202 may provide services to multiple customers in parallel. Cloud infrastructure system 2202 may store information for these customers, including possibly proprietary information. In certain embodiments, cloud infrastructure system 2202 comprises an identity management subsystem (IMS) 2228 that is configured to manage customer information and provide the separation of the managed information such that information related to one customer is not accessible by another customer. IMS 2228 may be configured to provide various security-related services such as identity services; information access management, authentication and authorization services; services for managing customer identities and roles and related capabilities, and the like.

[0163] FIG. 23 illustrates an example of computer system 2300. In some embodiments, computer system 2300 may be used to implement any of the systems described above. As shown in FIG. 23, computer system 2300 includes various subsystems including a processing subsystem 2304 that communicates with a number of other subsystems via a bus subsystem 2302. These other subsystems may include processing acceleration unit 2306, I/O subsystem 2308, storage subsystem 2318, and communications subsystem 2324. Storage subsystem 2318 may include non-transitory computer-readable storage media including storage media 2322 and system memory 2310.

[0164] Bus subsystem 2302 provides a mechanism for letting the various components and subsystems of computer system 2300 communicate with each other as intended. Although bus subsystem 2302 is shown schematically as a single bus, alternative examples of the bus subsystem may utilize multiple buses. Bus subsystem 2302 may be any of several types of bus structures including a memory bus or memory controller, a peripheral bus, a local bus using any of a variety of bus architectures, and the like. For example, such architectures may include an Industry Standard Architecture (ISA) bus, Micro Channel Architecture (MCA) bus, Enhanced ISA (EISA) bus, Video Electronics Standards Association (VESA) local bus, and Peripheral Component Interconnect (PCI) bus, which may be implemented as a Mezzanine bus manufactured to the IEEE P1386.1 standard, and the like.

[0165] Processing subsystem 2304 controls the operation of computer system 2300 and may comprise one or more processors, application specific integrated circuits (ASICs), or field programmable gate arrays (FPGAs). The processors may include be single core or multicore

processors. The processing resources of computer system 2300 may be organized into one or more processing units 2332, 2334, etc. A processing unit may include one or more processors, one or more cores from the same or different processors, a combination of cores and processors, or other combinations of cores and processors. In some embodiments, processing subsystem 2304 may include one or more special purpose co-processors such as graphics processors, digital signal processors (DSPs), or the like. In some embodiments, some or all of the processing units of processing subsystem 2304 may be implemented using customized circuits, such as application specific integrated circuits (ASICs), or field programmable gate arrays (FPGAs).

[0166] In some embodiments, the processing units in processing subsystem 2304 may execute instructions stored in system memory 2310 or on computer readable storage media 2322. In various examples, the processing units may execute a variety of programs or code instructions and may maintain multiple concurrently executing programs or processes. At any given time, some or all of the program code to be executed may be resident in system memory 2310 and/or on computer-readable storage media 2322 including potentially on one or more storage devices. Through suitable programming, processing subsystem 2304 may provide various functionalities described above. In instances where computer system 2300 is executing one or more virtual machines, one or more processing units may be allocated to each virtual machine.

[0167] In certain embodiments, a processing acceleration unit 2306 may optionally be provided for performing customized processing or for off-loading some of the processing performed by processing subsystem 2304 so as to accelerate the overall processing performed by computer system 2300.

[0168] I/O subsystem 2308 may include devices and mechanisms for inputting information to computer system 2300 and/or for outputting information from or via computer system 2300. In general, use of the term input device is intended to include all possible types of devices and mechanisms for inputting information to computer system 2300. User interface input devices may include, for example, a keyboard, pointing devices such as a mouse or trackball, a touchpad or touch screen incorporated into a display, a scroll wheel, a click wheel, a dial, a button, a switch, a keypad, audio input devices with voice command recognition systems, microphones, and other types of input devices. User interface input devices may also include motion sensing and/or gesture recognition devices such as the Microsoft Kinect® motion sensor that enables

users to control and interact with an input device, the Microsoft Xbox® 360 game controller, devices that provide an interface for receiving input using gestures and spoken commands. User interface input devices may also include eye gesture recognition devices such as the Google Glass® blink detector that detects eye activity (e.g., "blinking" while taking pictures and/or making a menu selection) from users and transforms the eye gestures as inputs to an input device (e.g., Google Glass®). Additionally, user interface input devices may include voice recognition sensing devices that enable users to interact with voice recognition systems (e.g., Siri® navigator) through voice commands.

[0169] Other examples of user interface input devices include, without limitation, three dimensional (3D) mice, joysticks or pointing sticks, gamepads and graphic tablets, and audio/visual devices such as speakers, digital cameras, digital camcorders, portable media players, webcams, image scanners, fingerprint scanners, barcode reader 3D scanners, 3D printers, laser rangefinders, and eye gaze tracking devices. Additionally, user interface input devices may include, for example, medical imaging input devices such as computed tomography, magnetic resonance imaging, position emission tomography, and medical ultrasonography devices. User interface input devices may also include, for example, audio input devices such as MIDI keyboards, digital musical instruments and the like.

[0170] In general, use of the term output device is intended to include all possible types of devices and mechanisms for outputting information from computer system 2300 to a user or other computer. User interface output devices may include a display subsystem, indicator lights, or non-visual displays such as audio output devices, etc. The display subsystem may be a cathode ray tube (CRT), a flat-panel device, such as that using a liquid crystal display (LCD) or plasma display, a projection device, a touch screen, and the like. For example, user interface output devices may include, without limitation, a variety of display devices that visually convey text, graphics and audio/video information such as monitors, printers, speakers, headphones, automotive navigation systems, plotters, voice output devices, and modems.

[0171] Storage subsystem 2318 provides a repository or data store for storing information and data that is used by computer system 2300. Storage subsystem 2318 provides a tangible non-transitory computer-readable storage medium for storing the basic programming and data constructs that provide the functionality of some examples. Storage subsystem 2318 may store

software (e.g., programs, code modules, instructions) that when executed by processing subsystem 2304 provides the functionality described above. The software may be executed by one or more processing units of processing subsystem 2304. Storage subsystem 2318 may also provide a repository for storing data used in accordance with the teachings of this disclosure.

[0172] Storage subsystem 2318 may include one or more non-transitory memory devices, including volatile and non-volatile memory devices. As shown in FIG. 23, storage subsystem 2318 includes system memory 2310 and computer-readable storage media 2322. System memory 2310 may include a number of memories including a volatile main random access memory (RAM) for storage of instructions and data during program execution and a non-volatile read only memory (ROM) or flash memory in which fixed instructions are stored. In some implementations, a basic input/output system (BIOS), containing the basic routines that help to transfer information between elements within computer system 2300, such as during start-up, may typically be stored in the ROM. The RAM typically contains data and/or program modules that are presently being operated and executed by processing subsystem 2304. In some implementations, system memory 2310 may include multiple different types of memory, such as static random access memory (SRAM), dynamic random access memory (DRAM), and the like.

[0173] By way of example, and not limitation, as depicted in FIG. 23, system memory 2310 may load application programs 2312 that are being executed, which may include various applications such as Web browsers, mid-tier applications, relational database management systems (RDBMS), etc., program data 2314, and operating system 2316. By way of example, operating system 2316 may include various versions of Microsoft Windows®, Apple Macintosh®, and/or Linux operating systems, a variety of commercially-available UNIX® or UNIX-like operating systems (including without limitation the variety of GNU/Linux operating systems, the Google Chrome® OS, and the like) and/or mobile operating systems such as iOS, Windows® Phone, Android® OS, BlackBerry® OS, Palm® OS operating systems, and others.

[0174] Computer-readable storage media 2322 may store programming and data constructs that provide the functionality of some examples. Computer-readable media 2322 may provide storage of computer-readable instructions, data structures, program modules, and other data for computer system 2300. Software (programs, code modules, instructions) that, when executed by processing subsystem 2304 provides the functionality described above, may be stored in storage

subsystem 2318. By way of example, computer-readable storage media 2322 may include non-volatile memory such as a hard disk drive, a magnetic disk drive, an optical disk drive such as a CD ROM, DVD, a Blu-Ray® disk, or other optical media. Computer-readable storage media 2322 may include, but is not limited to, Zip® drives, flash memory cards, universal serial bus (USB) flash drives, secure digital (SD) cards, DVD disks, digital video tape, and the like. Computer-readable storage media 2322 may also include, solid-state drives (SSD) based on non-volatile memory such as flash-memory based SSDs, enterprise flash drives, solid state ROM, and the like, SSDs based on volatile memory such as solid state RAM, dynamic RAM, static RAM, DRAM-based SSDs, magnetoresistive RAM (MRAM) SSDs, and hybrid SSDs that use a combination of DRAM and flash memory based SSDs.

[0175] In certain embodiments, storage subsystem 2318 may also include computer-readable storage media reader 2320 that may further be connected to computer-readable storage media 2322. Reader 2320 may receive and be configured to read data from a memory device such as a disk, a flash drive, etc.

[0176] In certain embodiments, computer system 2300 may support virtualization technologies, including but not limited to virtualization of processing and memory resources. For example, computer system 2300 may provide support for executing one or more virtual machines. In certain embodiments, computer system 2300 may execute a program such as a hypervisor that facilitated the configuring and managing of the virtual machines. Each virtual machine may be allocated memory, compute (e.g., processors, cores), I/O, and networking resources. Each virtual machine generally runs independently of the other virtual machines. A virtual machine typically runs its own operating system, which may be the same as or different from the operating systems executed by other virtual machines executed by computer system 2300. Accordingly, multiple operating systems may potentially be run concurrently by computer system 2300.

[0177] Communications subsystem 2324 provides an interface to other computer systems and networks. Communications subsystem 2324 serves as an interface for receiving data from and transmitting data to other systems from computer system 2300. For example, communications subsystem 2324 may enable computer system 2300 to establish a communication channel to one

or more client devices via the Internet for receiving and sending information from and to the client devices.

[0178] Communication subsystem 2324 may support both wired and/or wireless communication protocols. In certain embodiments, communications subsystem 2324 may include radio frequency (RF) transceiver components for accessing wireless voice and/or data networks (e.g., using cellular telephone technology, advanced data network technology, such as 3G, 4G or EDGE (enhanced data rates for global evolution), WiFi (IEEE 802.XX family standards, or other mobile communication technologies, or any combination thereof), global positioning system (GPS) receiver components, and/or other components. In some embodiments, communications subsystem 2324 may provide wired network connectivity (e.g., Ethernet) in addition to or instead of a wireless interface.

[0179] Communication subsystem 2324 may receive and transmit data in various forms. In some embodiments, in addition to other forms, communications subsystem 2324 may receive input communications in the form of structured and/or unstructured data feeds 2326, event streams 2328, event updates 2330, and the like. For example, communications subsystem 2324 may be configured to receive (or send) data feeds 2326 in real-time from users of social media networks and/or other communication services such as Twitter® feeds, Facebook® updates, web feeds such as Rich Site Summary (RSS) feeds, and/or real-time updates from one or more third party information sources.

[0180] In certain embodiments, communications subsystem 2324 may be configured to receive data in the form of continuous data streams, which may include event streams 2328 of real-time events and/or event updates 2330, that may be continuous or unbounded in nature with no explicit end. Examples of applications that generate continuous data may include, for example, sensor data applications, financial tickers, network performance measuring tools (e.g. network monitoring and traffic management applications), clickstream analysis tools, automobile traffic monitoring, and the like.

[0181] Communications subsystem 2324 may also be configured to communicate data from computer system 2300 to other computer systems or networks. The data may be communicated in various different forms such as structured and/or unstructured data feeds 2326, event streams

2328, event updates 2330, and the like to one or more databases that may be in communication with one or more streaming data source computers coupled to computer system 2300.

[0182] Computer system 2300 may be one of various types, including a handheld portable device (e.g., an iPhone® cellular phone, an iPad® computing tablet, a PDA), a wearable device (e.g., a Google Glass® head mounted display), a personal computer, a workstation, a mainframe, a kiosk, a server rack, or any other data processing system. Due to the ever-changing nature of computers and networks, the description of computer system 2300 depicted in FIG. 23 is intended only as a specific example. Many other configurations having more or fewer components than the system depicted in FIG. 23 are possible. Based on the disclosure and teachings provided herein, a person of ordinary skill in the art will appreciate other ways and/or methods to implement the various examples.

[0183] Although specific examples have been described, various modifications, alterations, alternative constructions, and equivalents are possible. Examples are not restricted to operation within certain specific data processing environments, but are free to operate within a plurality of data processing environments. Additionally, although certain examples have been described using a particular series of transactions and steps, it should be apparent to those skilled in the art that this is not intended to be limiting. Although some flowcharts describe operations as a sequential process, many of the operations may be performed in parallel or concurrently. In addition, the order of the operations may be rearranged. A process may have additional steps not included in the figure. Various features and aspects of the above-described examples may be used individually or jointly.

[0184] Further, while certain examples have been described using a particular combination of hardware and software, it should be recognized that other combinations of hardware and software are also possible. Certain examples may be implemented only in hardware, or only in software, or using combinations thereof. The various processes described herein may be implemented on the same processor or different processors in any combination.

[0185] Where devices, systems, components or modules are described as being configured to perform certain operations or functions, such configuration may be accomplished, for example, by designing electronic circuits to perform the operation, by programming programmable electronic circuits (such as microprocessors) to perform the operation such as by executing

computer instructions or code, or processors or cores programmed to execute code or instructions stored on a non-transitory memory medium, or any combination thereof. Processes may communicate using a variety of techniques including but not limited to conventional techniques for inter-process communications, and different pairs of processes may use different techniques, or the same pair of processes may use different techniques at different times.

[0186] Specific details are given in this disclosure to provide a thorough understanding of the examples. However, examples may be practiced without these specific details. For example, well-known circuits, processes, algorithms, structures, and techniques have been shown without unnecessary detail in order to avoid obscuring the examples. This description provides example examples only, and is not intended to limit the scope, applicability, or configuration of other examples. Rather, the preceding description of the examples will provide those skilled in the art with an enabling description for implementing various examples. Various changes may be made in the function and arrangement of elements.

[0187] The specification and drawings are, accordingly, to be regarded in an illustrative rather than a restrictive sense. It will, however, be evident that additions, subtractions, deletions, and other modifications and changes may be made thereunto without departing from the broader spirit and scope as set forth in the claims. Thus, although specific examples have been described, these are not intended to be limiting. Various modifications and equivalents are within the scope of the following claims.

[0188] In the foregoing specification, aspects of the disclosure are described with reference to specific examples thereof, but those skilled in the art will recognize that the disclosure is not limited thereto. Various features and aspects of the above-described disclosure may be used individually or jointly. Further, examples may be utilized in any number of environments and applications beyond those described herein without departing from the broader spirit and scope of the specification. The specification and drawings are, accordingly, to be regarded as illustrative rather than restrictive.

[0189] In the foregoing description, for the purposes of illustration, methods were described in a particular order. It should be appreciated that in alternate examples, the methods may be performed in a different order than that described. It should also be appreciated that the methods described above may be performed by hardware components or may be embodied in sequences

of machine-executable instructions, which may be used to cause a machine, such as a general-purpose or special-purpose processor or logic circuits programmed with the instructions to perform the methods. These machine-executable instructions may be stored on one or more machine readable mediums, such as CD-ROMs or other type of optical disks, floppy diskettes, ROMs, RAMs, EPROMs, EEPROMs, magnetic or optical cards, flash memory, or other types of machine-readable mediums suitable for storing electronic instructions. Alternatively, the methods may be performed by a combination of hardware and software.

[0190] Where components are described as being configured to perform certain operations, such configuration may be accomplished, for example, by designing electronic circuits or other hardware to perform the operation, by programming programmable electronic circuits (e.g., microprocessors, or other suitable electronic circuits) to perform the operation, or any combination thereof.

[0191] While illustrative examples of the application have been described in detail herein, it is to be understood that the inventive concepts may be otherwise variously embodied and employed, and that the appended claims are intended to be construed to include such variations, except as limited by the prior art.

[0192] Where components are described as being "configured to" perform certain operations, such configuration may be accomplished, for example, by designing electronic circuits or other hardware to perform the operation, by programming programmable electronic circuits (e.g., microprocessors, or other suitable electronic circuits) to perform the operation, or any combination thereof.

WHAT IS CLAIMED IS:

1. A method of generating regular expressions, comprising:

receiving, by a regular expression generator comprising one or more processors, first input data comprising one or more positive character sequences, each of the one or more positive character sequences corresponding to a positive example that is to be matched by a regular expression generated by the regular expression generator;

generating, by the regular expression generator, a first regular expression, wherein the first regular expression matches each of the one or more positive examples;

receiving, by the regular expression generator, second input data comprising one or more negative character sequences, each of the one or more negative character sequences corresponding to a negative example that is not to be matched by the regular expression generated by the regular expression generator;

in response to receiving the second input data, determining whether each of the one or more negative examples matches the first regular expression; and

in response to determining that at least one of the negative examples matches the first regular expression:

(a) determining a subsequence of characters at a location within the first regular expression;

(b) determining a replacement character sequence, wherein the replacement character sequence at the location within the regular expression distinguishes the one or more positive examples from the one more negative examples; and

(c) updating the first regular expression by replacing the determined subsequence of characters within the first regular expression with the replacement character sequence.

2. The method of claim 1, wherein determining the subsequence of characters at the location within the first regular expression comprises:

determining the location within the first regular expression;

retrieving a text fragment from each of the one or more positive examples and from each of the one or more negative examples corresponding to the location within the first regular expression; and

determining the subsequence of characters as one or more characters at the location within the first regular expression, from which the one or more positive examples are distinguishable from the one more negative examples.

3. The method of claim 2, wherein determining the location within the first regular expression comprises:

determining a first number of characters at a prefix portion of the first regular expression from which the one or more positive examples are distinguishable from the one more negative examples;

determining a second number of characters at a suffix portion of the first regular expression from which the one or more positive examples are distinguishable from the one more negative examples; and

selecting either the prefix portion or the suffix portion as the location within the first regular expression, based at least in part on whether the first number of characters or the second number of characters is shorter.

4. The method of claim 3, wherein determining the location within the first regular expression further comprises:

executing a formula to determine the location within the first regular expression, wherein the formula weights the prefix portion over the suffix portion of the first regular expression.

5. The method of claim 2, wherein the determined location within the first regular expression is a mid-span location that does not correspond to a prefix portion of the first regular expression or a suffix portion of the first regular expression.

6. The method of claim 2, wherein determining the replacement character sequence comprises determining a plurality of replacement character sequences, and wherein updating the first regular expression comprises replacing the determined subsequence of

characters within the first regular expression with the plurality of replacement character sequences.

7. The method of claim 1, wherein determining the replacement character sequence comprises:

determining a first number of characters at the location within the first regular expression, and a corresponding first number of replacement character sequences each having the first number of characters, from which the one or more positive examples are distinguishable from the one more negative examples;

determining a second number of characters at the location within the first regular expression, and a corresponding second number of replacement character sequences each having the second number of characters, from which the one or more positive examples are distinguishable from the one more negative examples; and

selecting either the first number of characters or the second number of characters for the replacement character sequence within the first regular expression, based on (a) the sizes of the first and second numbers of characters, and (b) the sizes of the corresponding first and second numbers of replacement character sequences.

8. A system for generating regular expressions, the system comprising:

a processing unit comprising one or more processors; and

memory storing instructions that, when executed by the processing unit, cause the system to:

receive first input data comprising one or more positive character sequences, each of the one or more positive character sequences corresponding to a positive example that is to be matched by a regular expression generated by the regular expression generator;

generate a first regular expression, wherein the first regular expression matches each of the one or more positive examples;

receive second input data comprising one or more negative character sequences, each of the one or more negative character sequences corresponding to a negative example that is not to be matched by the regular expression generated by the regular expression generator;

in response to receiving the second input data, determine whether each of the one or more negative examples matches the first regular expression; and

in response to determining that at least one of the negative examples matches the first regular expression:

(a) determine a subsequence of characters at a location within the first regular expression;

(b) determine a replacement character sequence, wherein the replacement character sequence at the location within the regular expression distinguishes the one or more positive examples from the one more negative examples; and

(c) update the first regular expression by replacing the determined subsequence of characters within the first regular expression with the replacement character sequence.

9. The system of claim 8, wherein determining the subsequence of characters at the location within the first regular expression comprises:

determining the location within the first regular expression;

retrieving a text fragment from each of the one or more positive examples and from each of the one or more negative examples corresponding to the location within the first regular expression; and

determining the subsequence of characters as one or more characters at the location within the first regular expression, from which the one or more positive examples are distinguishable from the one more negative examples.

10. The system of claim 9, wherein determining the location within the first regular expression comprises:

determining a first number of characters at a prefix portion of the first regular expression from which the one or more positive examples are distinguishable from the one more negative examples;

determining a second number of characters at a suffix portion of the first regular expression from which the one or more positive examples are distinguishable from the one more negative examples; and

selecting either the prefix portion or the suffix portion as the location within the first regular expression, based at least in part on whether the first number of characters or the second number of characters is shorter.

11. The system of claim 10, wherein determining the location within the first regular expression further comprises:

executing a formula to determine the location within the first regular expression, wherein the formula weights the prefix portion over the suffix portion of the first regular expression.

12. The system of claim 9, wherein the determined location within the first regular expression is a mid-span location that does not correspond to a prefix portion of the first regular expression or a suffix portion of the first regular expression.

13. The system of claim 9, wherein determining the replacement character sequence comprises determining a plurality of replacement character sequences, and wherein updating the first regular expression comprises replacing the determined subsequence of characters within the first regular expression with the plurality of replacement character sequences.

14. The system of claim 8, wherein determining the replacement character sequence comprises:

determining a first number of characters at the location within the first regular expression, and a corresponding first number of replacement character sequences each having the first number of characters, from which the one or more positive examples are distinguishable from the one more negative examples;

determining a second number of characters at the location within the first regular expression, and a corresponding second number of replacement character sequences each having the second number of characters, from which the one or more positive examples are distinguishable from the one more negative examples; and

selecting either the first number of characters or the second number of characters for the replacement character sequence within the first regular expression, based on (a) the sizes

of the first and second numbers of characters, and (b) the sizes of the corresponding first and second numbers of replacement character sequences.

15. A non-transitory computer-readable media for generating regular expressions, the computer-readable media comprising computer-executable instructions which when executed on a computer system, cause the computer system to:

receive first input data comprising one or more positive character sequences, each of the one or more positive character sequences corresponding to a positive example that is to be matched by a regular expression generated by the regular expression generator;

generate a first regular expression, wherein the first regular expression matches each of the one or more positive examples;

receive second input data comprising one or more negative character sequences, each of the one or more negative character sequences corresponding to a negative example that is not to be matched by the regular expression generated by the regular expression generator;

in response to receiving the second input data, determine whether each of the one or more negative examples matches the first regular expression; and

in response to determining that at least one of the negative examples matches the first regular expression:

(a) determine a subsequence of characters at a location within the first regular expression;

(b) determine a replacement character sequence, wherein the replacement character sequence at the location within the regular expression distinguishes the one or more positive examples from the one more negative examples; and

(c) update the first regular expression by replacing the determined subsequence of characters within the first regular expression with the replacement character sequence.

16. The computer-readable media of claim 15, wherein determining the subsequence of characters at the location within the first regular expression comprises:

determining the location within the first regular expression;

retrieving a text fragment from each of the one or more positive examples and from each of the one or more negative examples corresponding to the location within the first regular expression; and

determining the subsequence of characters as one or more characters at the location within the first regular expression, from which the one or more positive examples are distinguishable from the one more negative examples.

17. The computer-readable media of claim 16, wherein determining the location within the first regular expression comprises:

determining a first number of characters at a prefix portion of the first regular expression from which the one or more positive examples are distinguishable from the one more negative examples;

determining a second number of characters at a suffix portion of the first regular expression from which the one or more positive examples are distinguishable from the one more negative examples; and

selecting either the prefix portion or the suffix portion as the location within the first regular expression, based at least in part on whether the first number of characters or the second number of characters is shorter.

18. The computer-readable media of claim 17, wherein determining the location within the first regular expression further comprises:

executing a formula to determine the location within the first regular expression, wherein the formula weights the prefix portion over the suffix portion of the first regular expression.

19. The computer-readable media of claim 16, wherein the determined location within the first regular expression is a mid-span location that does not correspond to a prefix portion of the first regular expression or a suffix portion of the first regular expression.

20. The computer-readable media of claim 16, wherein determining the replacement character sequence comprises determining a plurality of replacement character sequences, and wherein updating the first regular expression comprises replacing the determined

subsequence of characters within the first regular expression with the plurality of replacement character sequences.

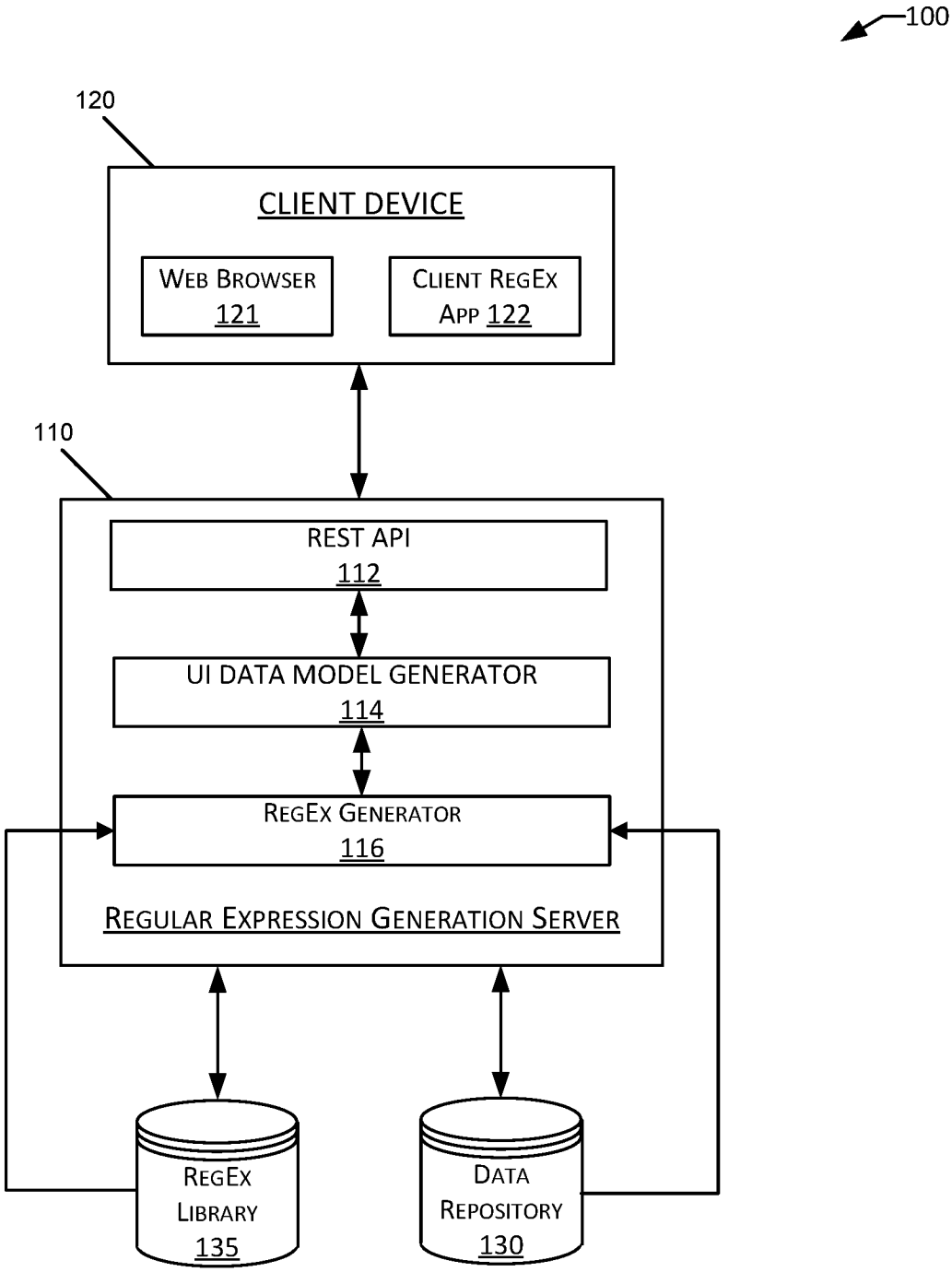

100
120
CLIENT DEVICE
WEB BROWSER 121
CLIENT REGEX APP 122
110
REST API 112
UI DATA MODEL GENERATOR 114
REGEX GENERATOR 116
REGULAR EXPRESSION GENERATION SERVER
REGEX LIBRARY 135
DATA REPOSITORY 130

FIG. 1

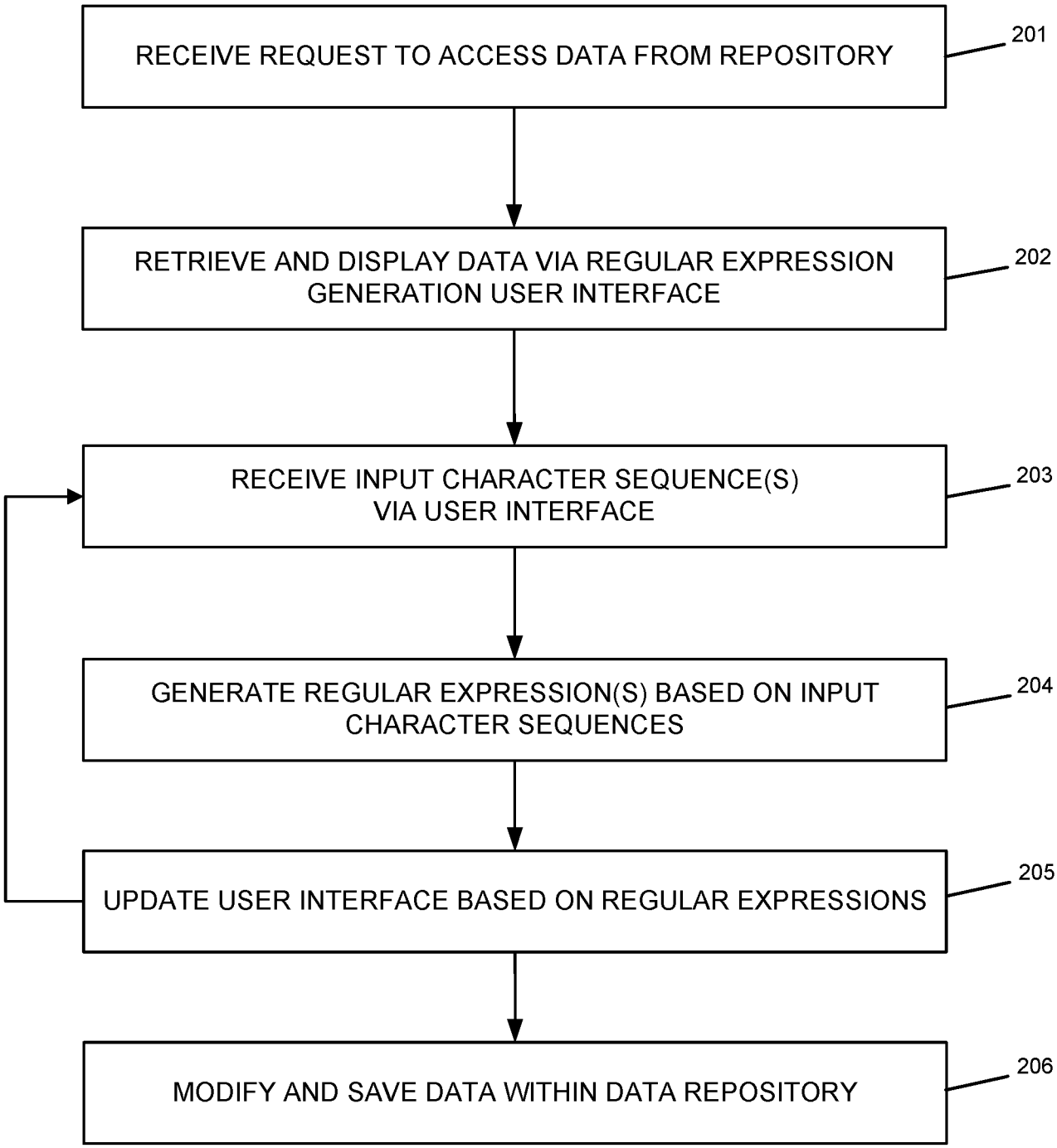

RECEIVE REQUEST TO ACCESS DATA FROM REPOSITORY
201
RETRIEVE AND DISPLAY DATA VIA REGULAR EXPRESSION GENERATION USER INTERFACE
202
RECEIVE INPUT CHARACTER SEQUENCE(S) VIA USER INTERFACE
203
GENERATE REGULAR EXPRESSION(S) BASED ON INPUT CHARACTER SEQUENCES
204
UPDATE USER INTERFACE BASED ON REGULAR EXPRESSIONS
205
MODIFY AND SAVE DATA WITHIN DATA REPOSITORY
206

FIG. 2

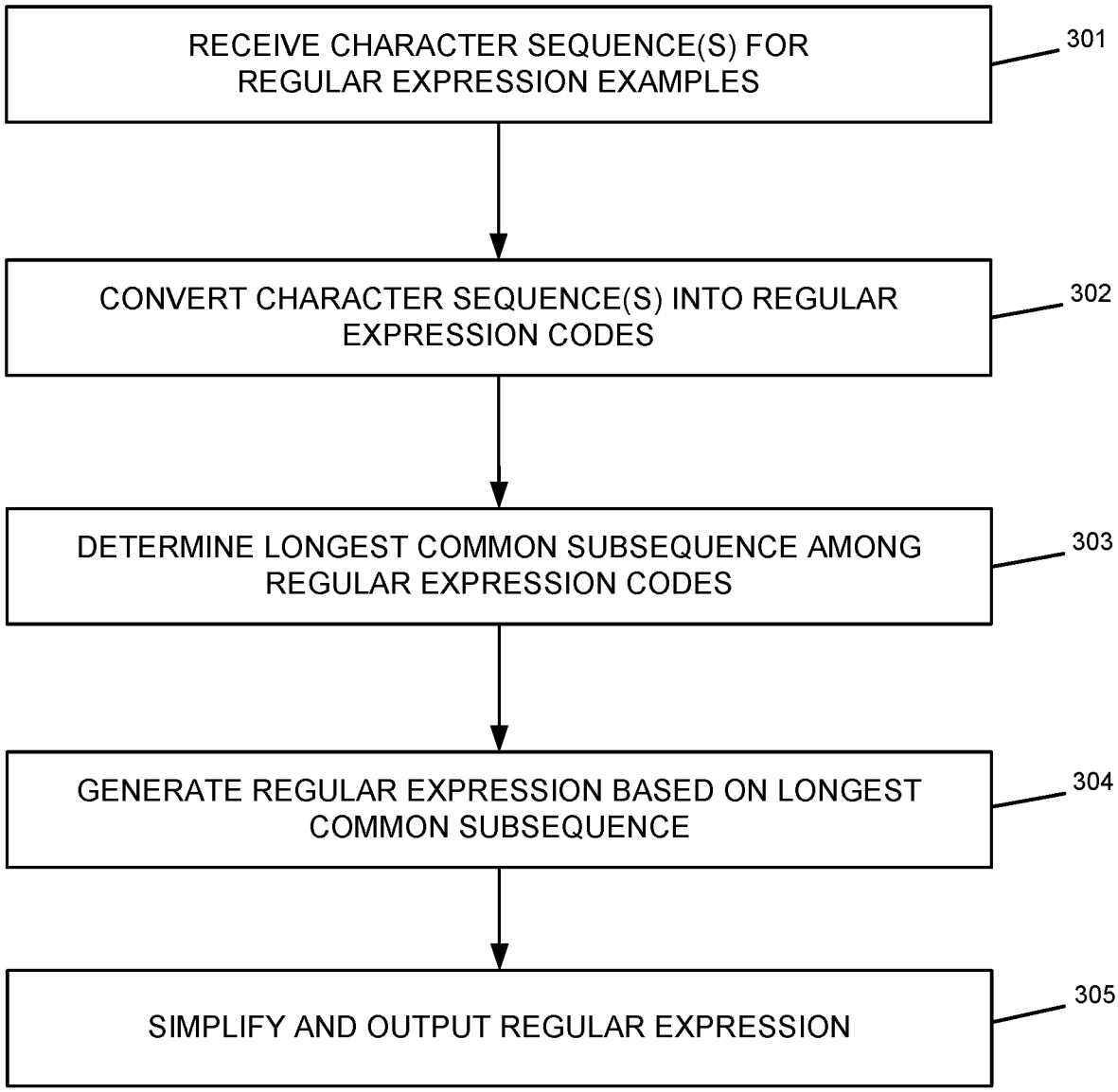

RECEIVE CHARACTER SEQUENCE(S) FOR REGULAR EXPRESSION EXAMPLES
301
CONVERT CHARACTER SEQUENCE(S) INTO REGULAR EXPRESSION CODES
302
DETERMINE LONGEST COMMON SUBSEQUENCE AMONG REGULAR EXPRESSION CODES
303
GENERATE REGULAR EXPRESSION BASED ON LONGEST COMMON SUBSEQUENCE
304
SIMPLIFY AND OUTPUT REGULAR EXPRESSION
305

FIG. 3

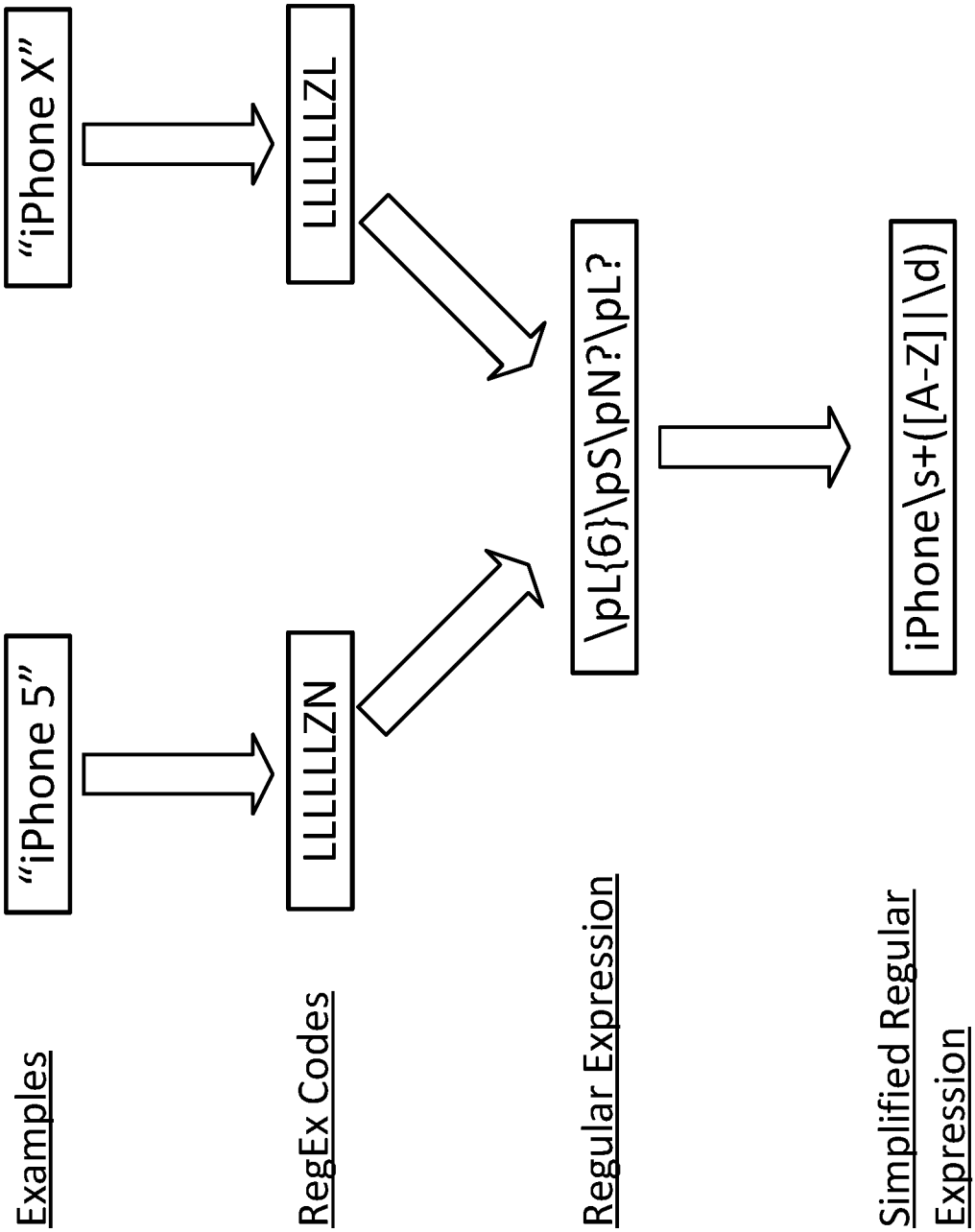

Examples
"iPhone 5"
"iPhone X"
RegEx Codes
LLLLLLZN
LLLLLLZL
Regular Expression
\pL{6}\pS\pN?\pL?
Simplified Regular Expression
iPhone\s+([A-Z]|\d)

FIG. 4

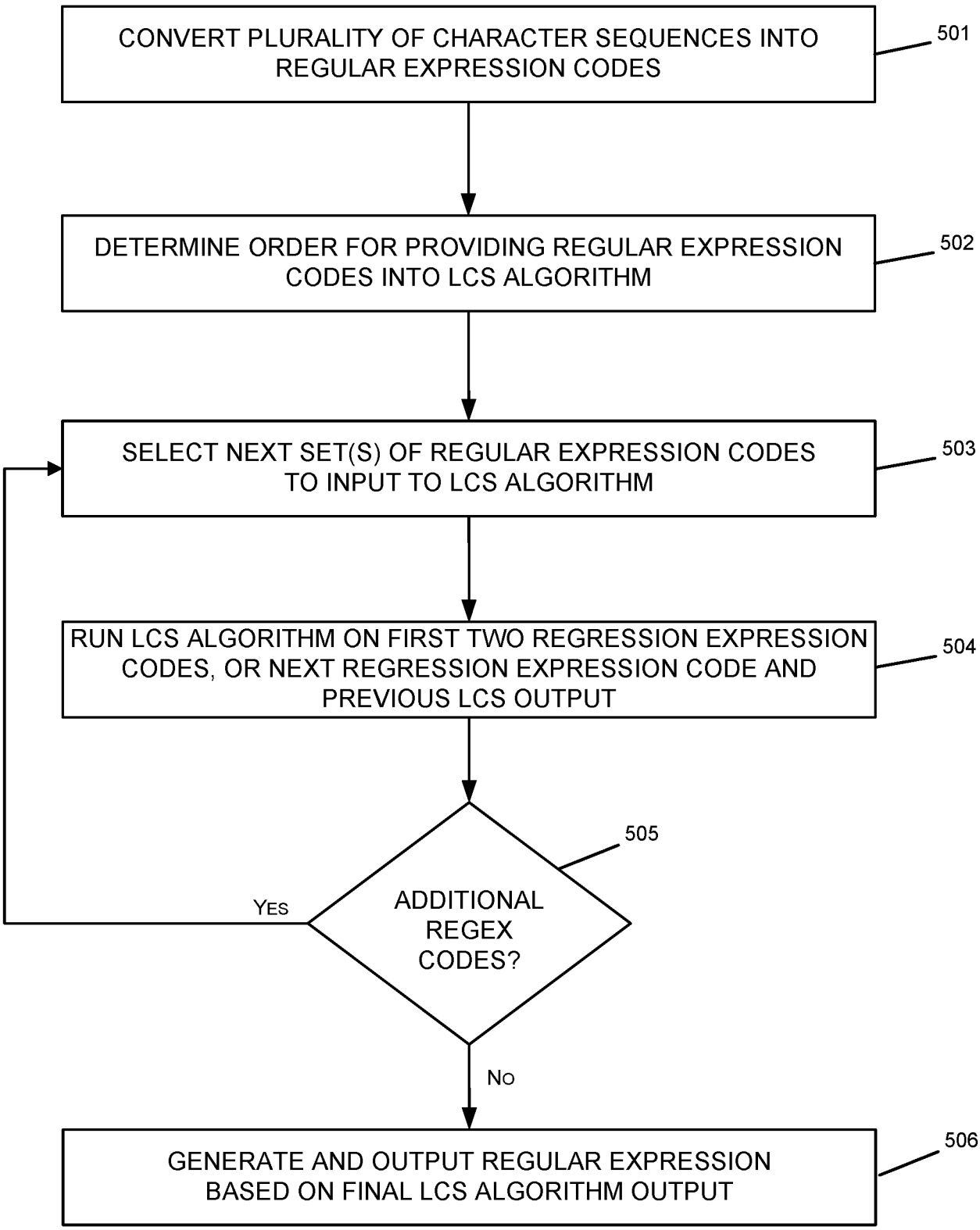

CONVERT PLURALITY OF CHARACTER SEQUENCES INTO REGULAR EXPRESSION CODES
501
DETERMINE ORDER FOR PROVIDING REGULAR EXPRESSION CODES INTO LCS ALGORITHM
502
SELECT NEXT SET(S) OF REGULAR EXPRESSION CODES TO INPUT TO LCS ALGORITHM
503
RUN LCS ALGORITHM ON FIRST TWO REGRESSION EXPRESSION CODES, OR NEXT REGRESSION EXPRESSION CODE AND PREVIOUS LCS OUTPUT
504
ADDITIONAL REGEX CODES?
505
YES
NO
GENERATE AND OUTPUT REGULAR EXPRESSION BASED ON FINAL LCS ALGORITHM OUTPUT
506

FIG. 5

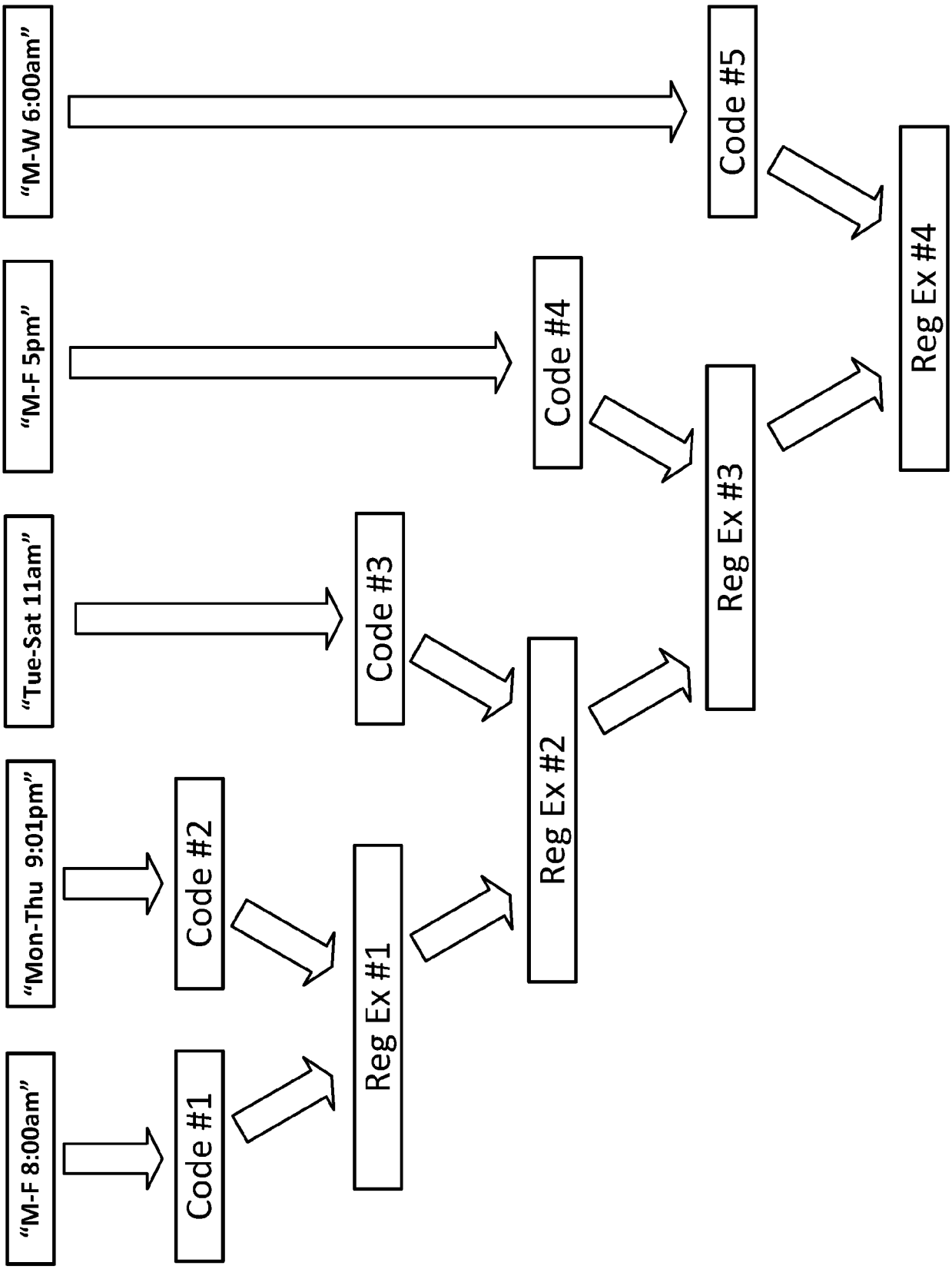

"M-F 8:00am"
"Mon-Thu 9:01pm"
"Tue-Sat 11am"
"M-F 5pm"
"M-W 6:00am"
Code #1
Code #2
Code #3
Code #4
Code #5
Reg Ex #1
Reg Ex #2
Reg Ex #3
Reg Ex #4

FIG. 6

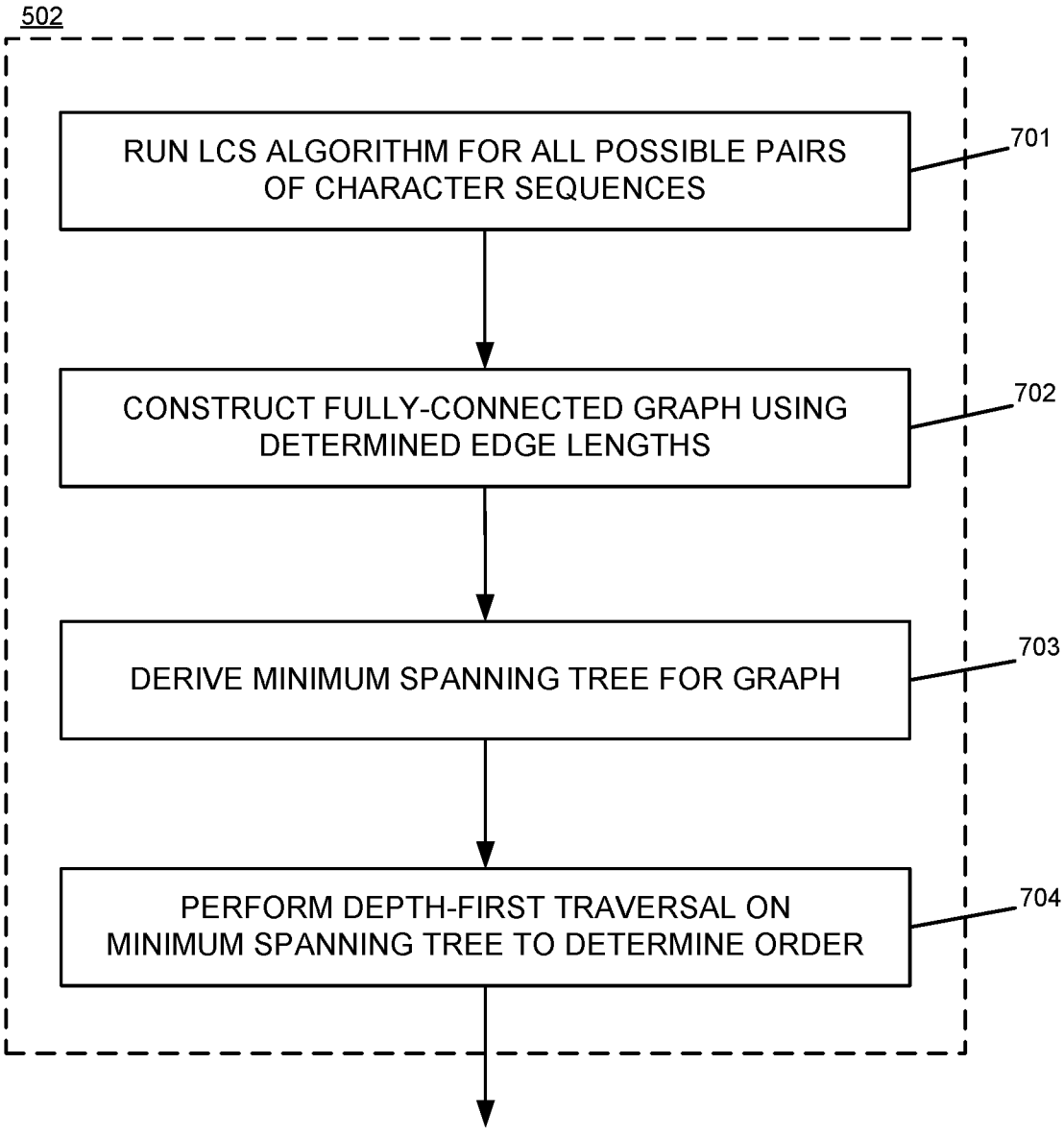

502
RUN LCS ALGORITHM FOR ALL POSSIBLE PAIRS OF CHARACTER SEQUENCES
701
CONSTRUCT FULLY-CONNECTED GRAPH USING DETERMINED EDGE LENGTHS
702
DERIVE MINIMUM SPANNING TREE FOR GRAPH
703
PERFORM DEPTH-FIRST TRAVERSAL ON MINIMUM SPANNING TREE TO DETERMINE ORDER
704

FIG. 7

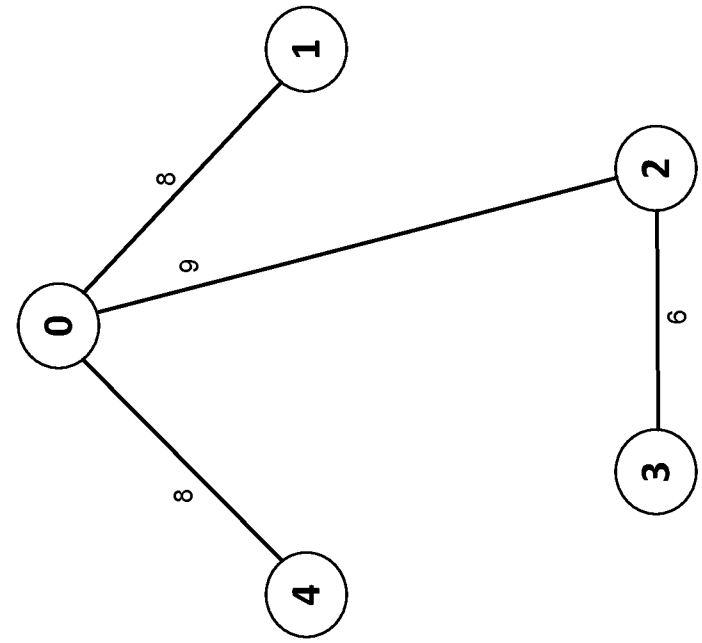


FIG. 8B

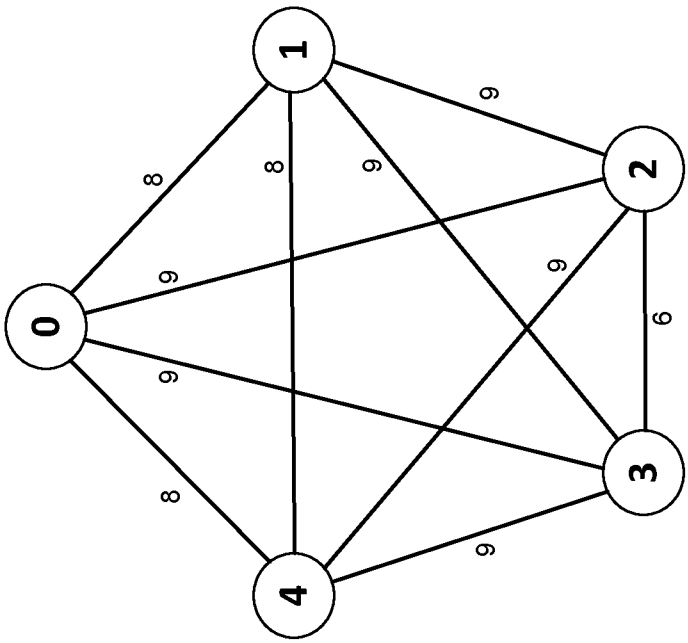


FIG. 8A

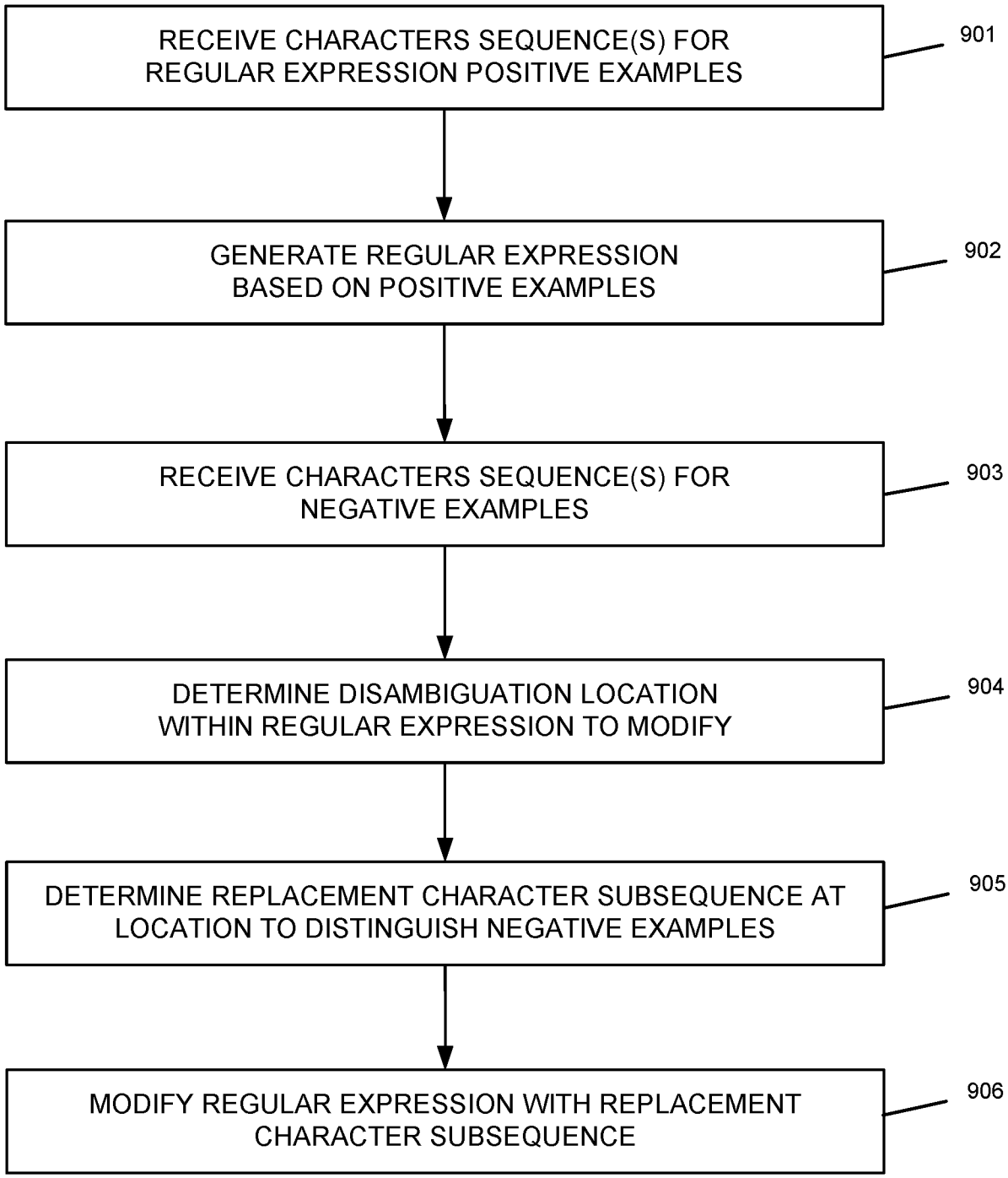

RECEIVE CHARACTERS SEQUENCE(S) FOR REGULAR EXPRESSION POSITIVE EXAMPLES
901
GENERATE REGULAR EXPRESSION BASED ON POSITIVE EXAMPLES
902
RECEIVE CHARACTERS SEQUENCE(S) FOR NEGATIVE EXAMPLES
903
DETERMINE DISAMBIGUATION LOCATION WITHIN REGULAR EXPRESSION TO MODIFY
904
DETERMINE REPLACEMENT CHARACTER SUBSEQUENCE AT LOCATION TO DISTINGUISH NEGATIVE EXAMPLES
905
MODIFY REGULAR EXPRESSION WITH REPLACEMENT CHARACTER SUBSEQUENCE
906

FIG. 9

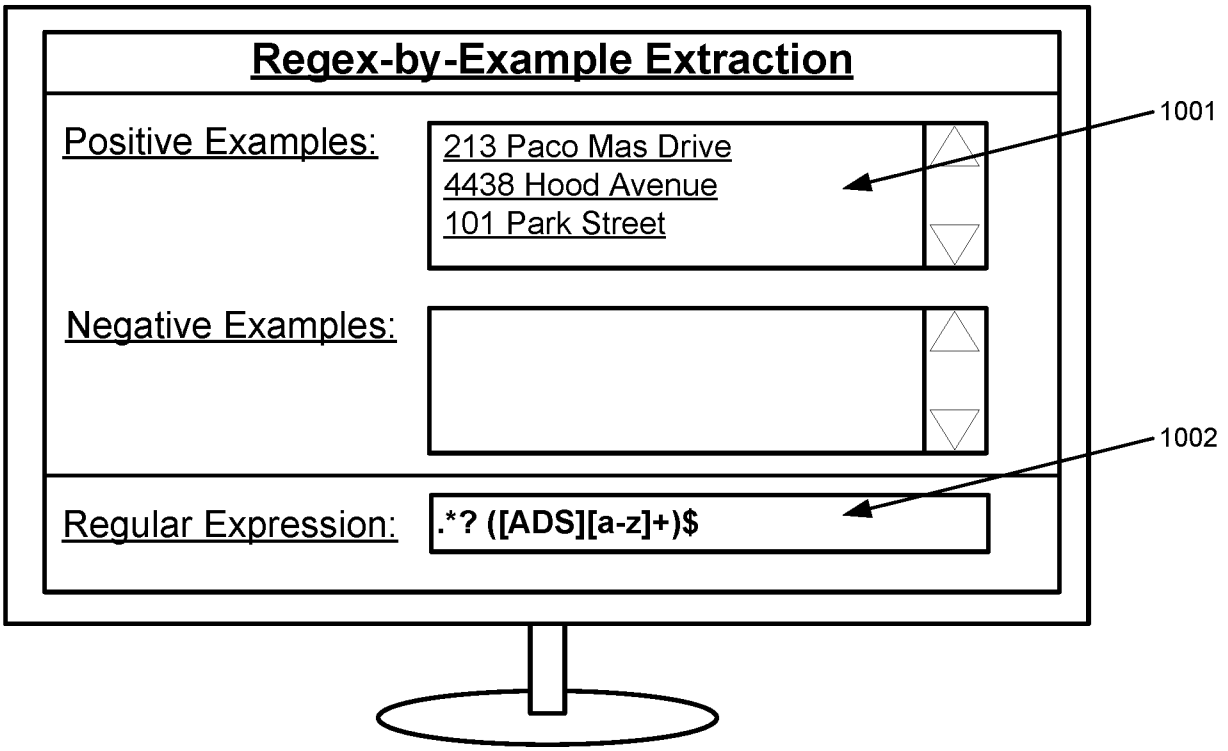


FIG. 10A

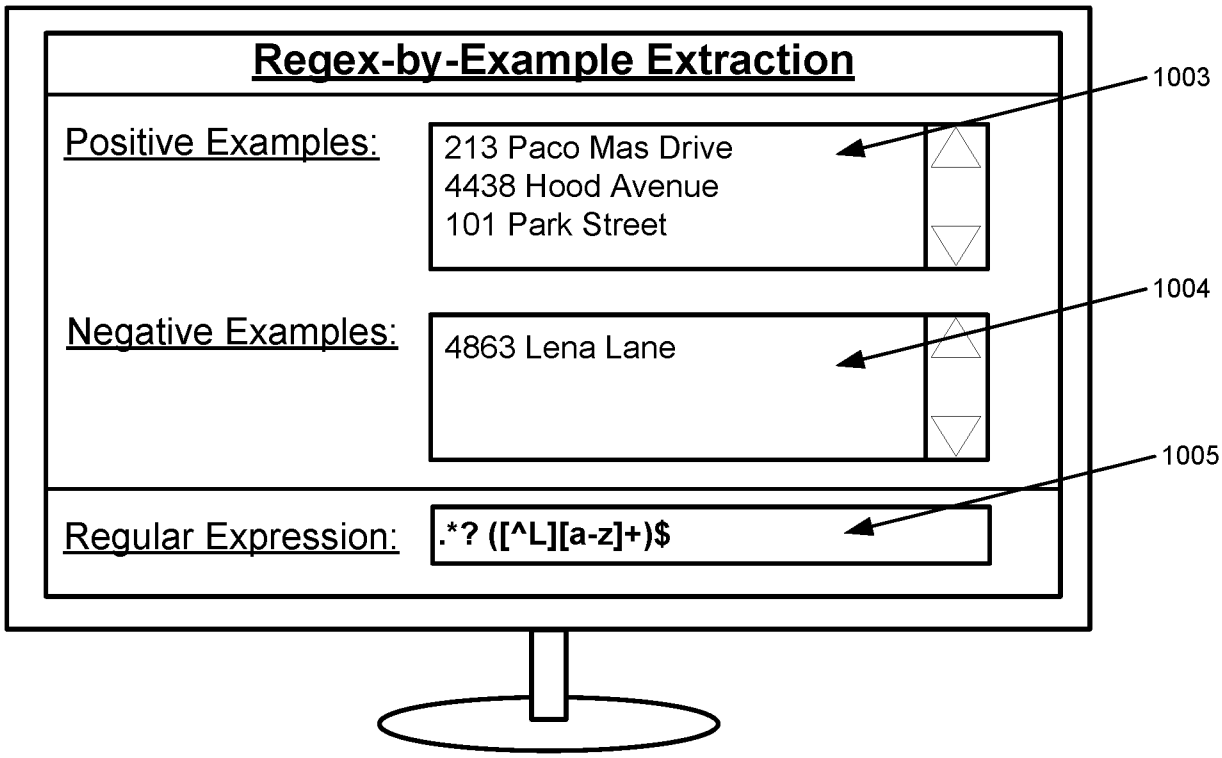


FIG. 10B

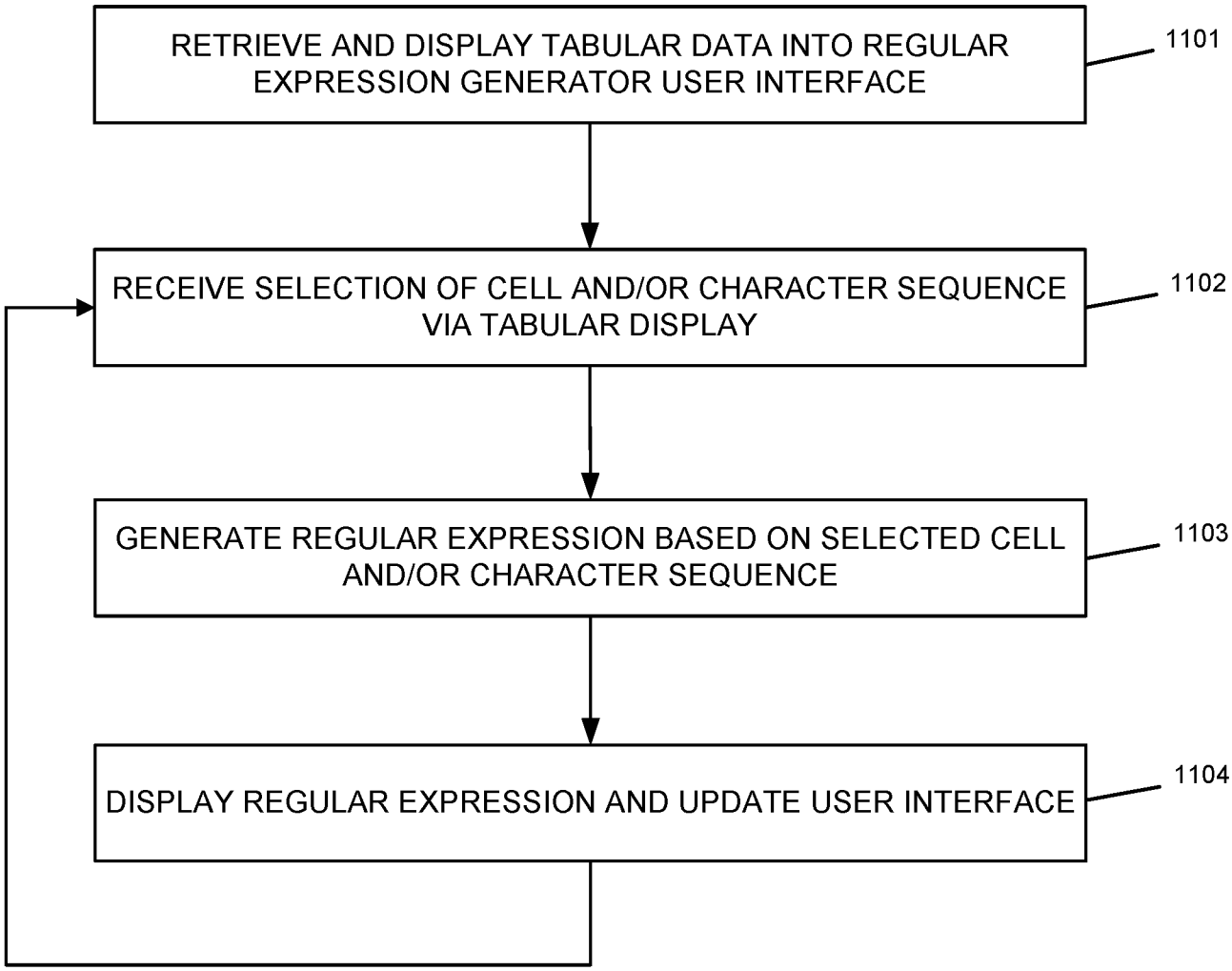

RETRIEVE AND DISPLAY TABULAR DATA INTO REGULAR EXPRESSION GENERATOR USER INTERFACE
1101
RECEIVE SELECTION OF CELL AND/OR CHARACTER SEQUENCE VIA TABULAR DISPLAY
1102
GENERATE REGULAR EXPRESSION BASED ON SELECTED CELL AND/OR CHARACTER SEQUENCE
1103
DISPLAY REGULAR EXPRESSION AND UPDATE USER INTERFACE
1104

FIG. 11

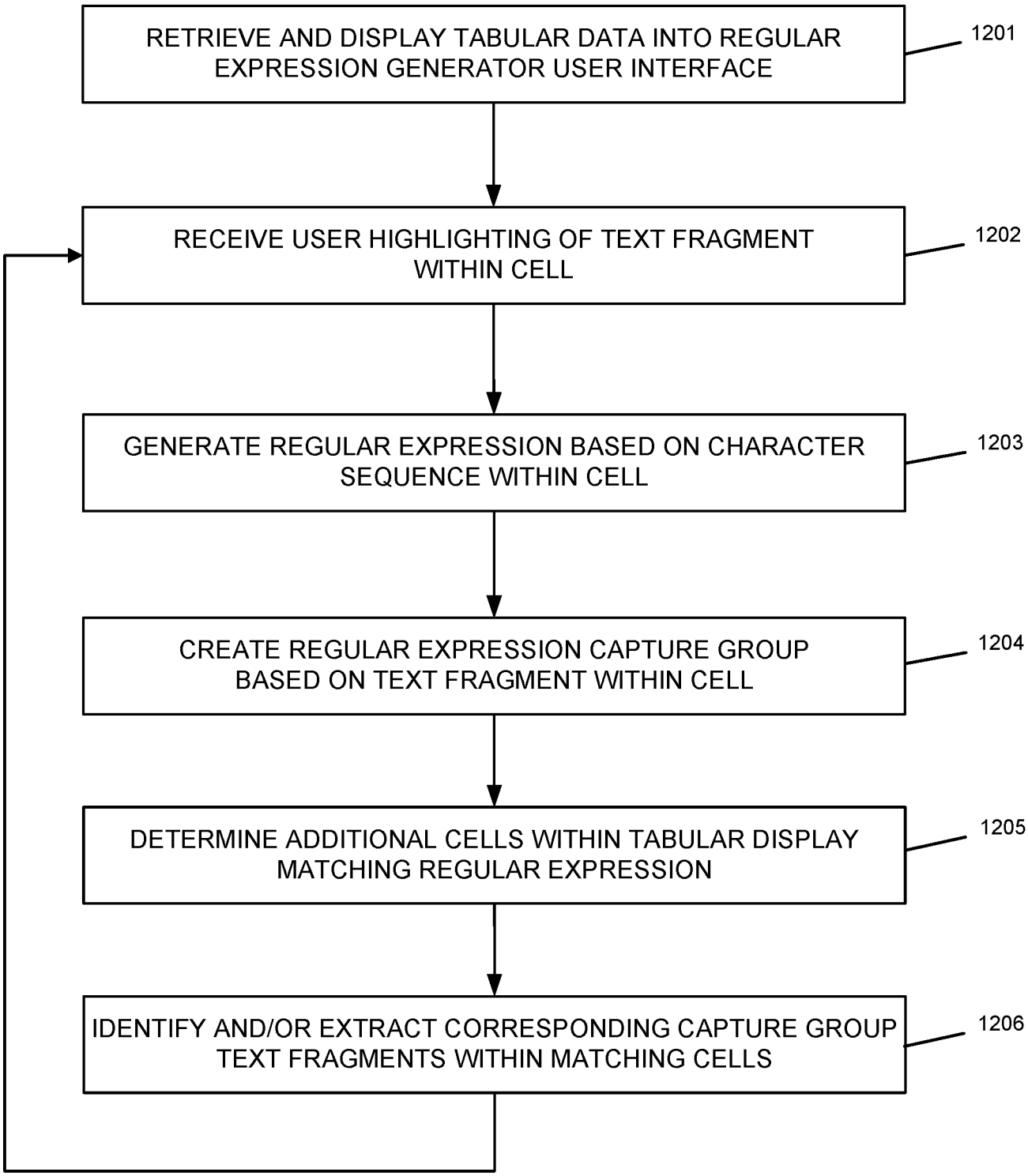

RETRIEVE AND DISPLAY TABULAR DATA INTO REGULAR EXPRESSION GENERATOR USER INTERFACE
1201
RECEIVE USER HIGHLIGHTING OF TEXT FRAGMENT WITHIN CELL
1202
GENERATE REGULAR EXPRESSION BASED ON CHARACTER SEQUENCE WITHIN CELL
1203
CREATE REGULAR EXPRESSION CAPTURE GROUP BASED ON TEXT FRAGMENT WITHIN CELL
1204
DETERMINE ADDITIONAL CELLS WITHIN TABULAR DISPLAY MATCHING REGULAR EXPRESSION
1205
IDENTIFY AND/OR EXTRACT CORRESPONDING CAPTURE GROUP TEXT FRAGMENTS WITHIN MATCHING CELLS
1206

FIG. 12

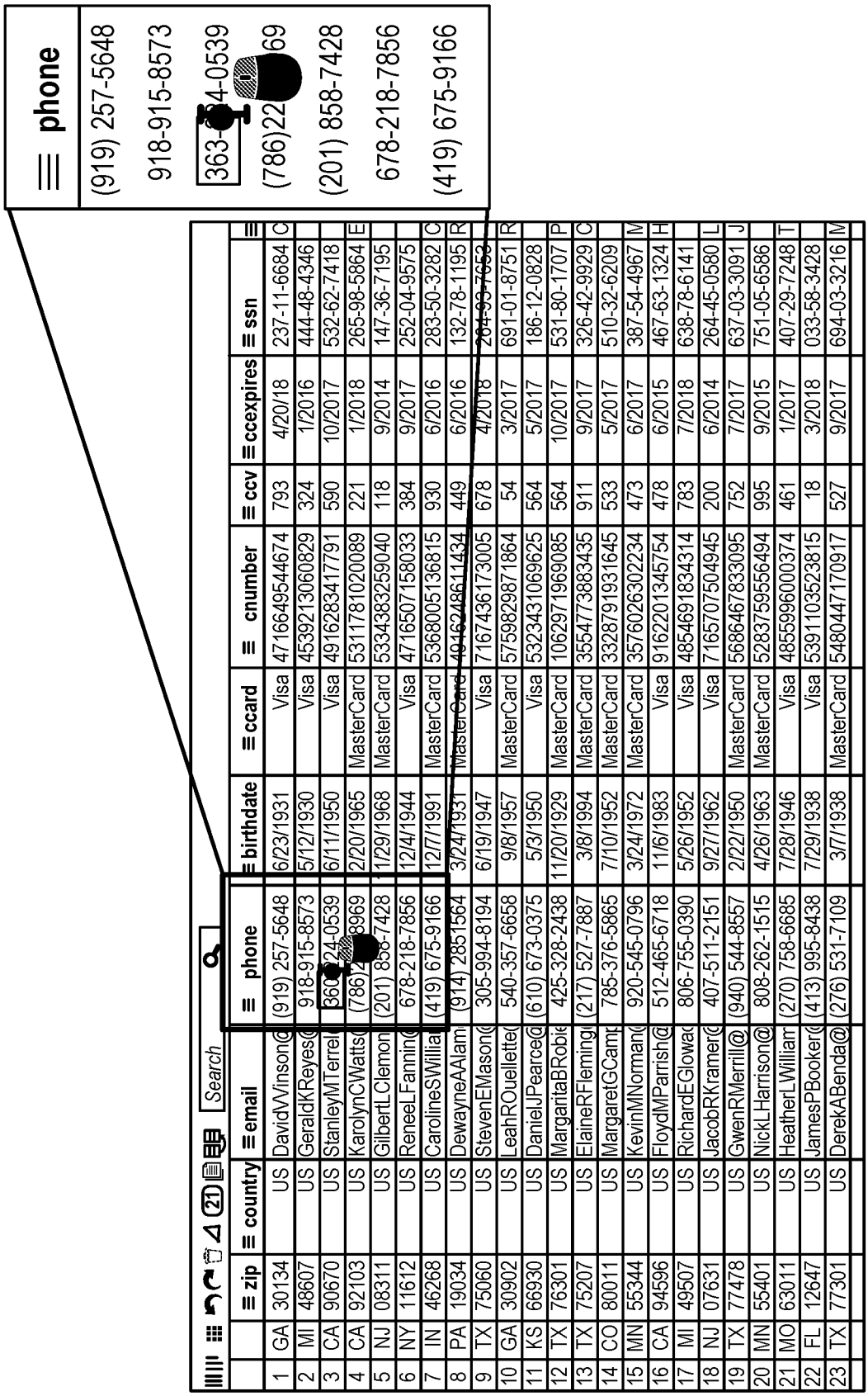


FIG. 13

SUBSTITUTE SHEET (RULE 26)

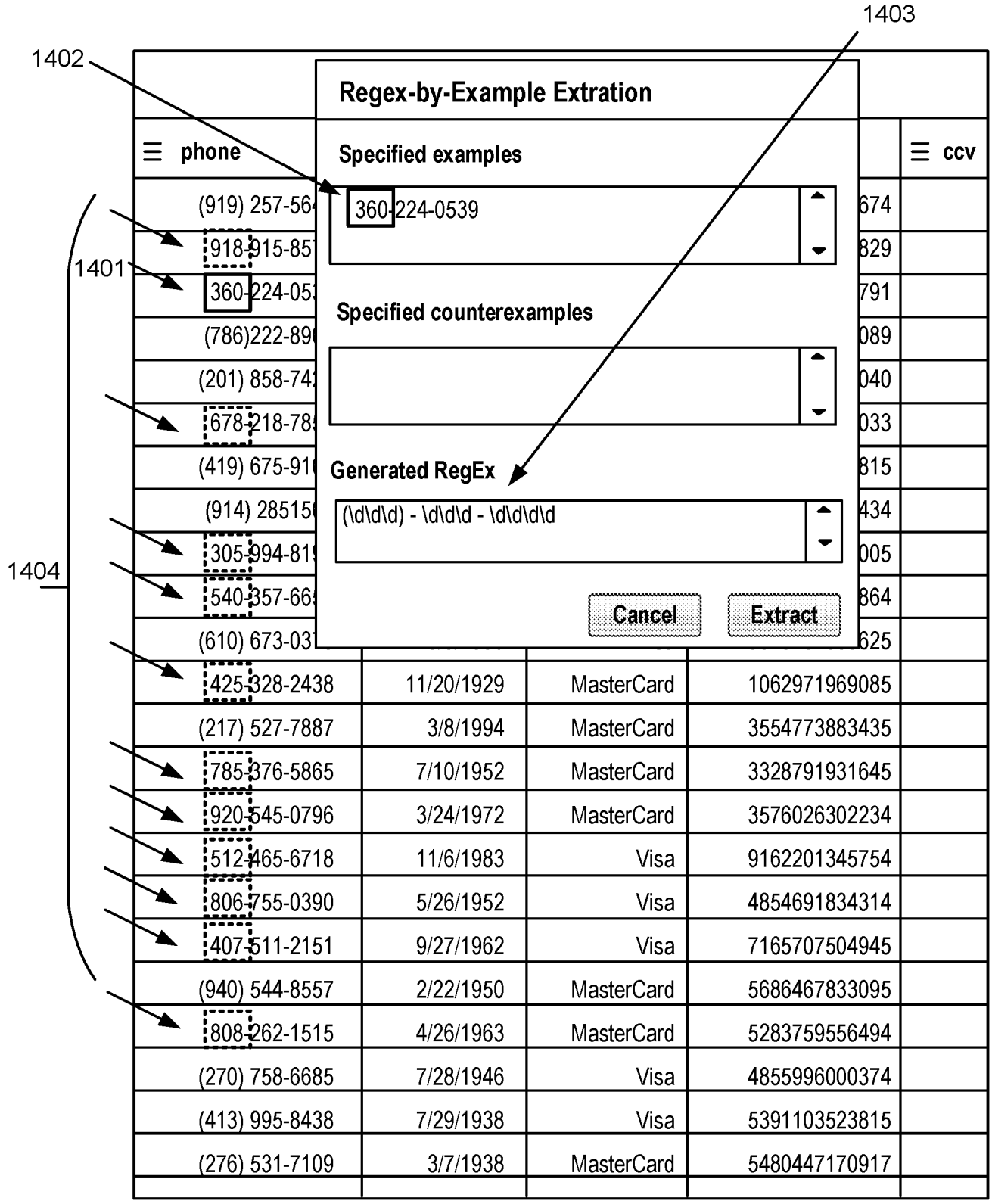
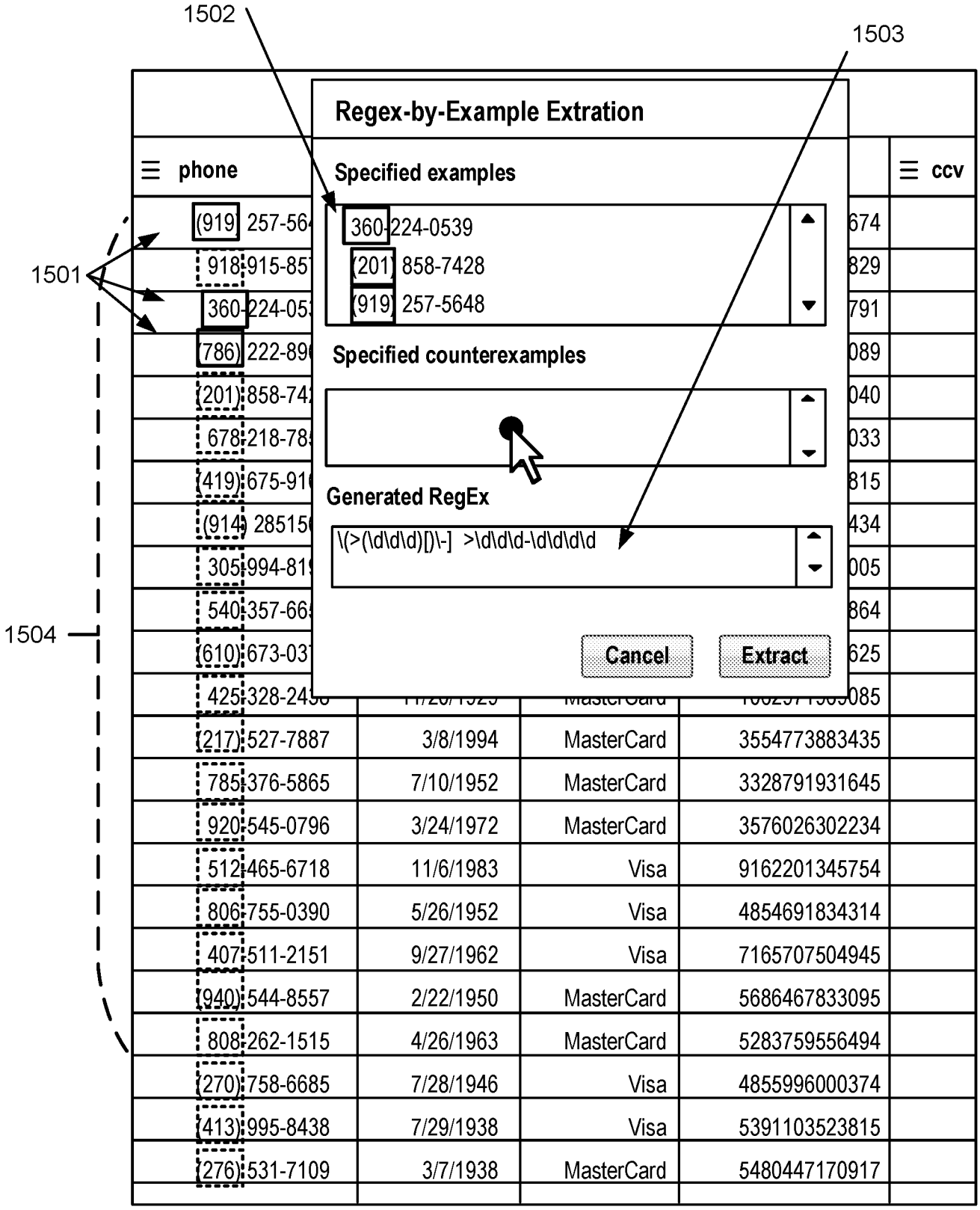

1403
1402
Regex-by-Example Extration
phone
Specified examples
360-224-0539
ccv
1401
Specified counterexamples
Generated RegEx
(\d\d\d) - \d\d\d - \d\d\d\d
Cancel
Extract
1404
425-328-2438 11/20/1929 MasterCard 1062971969085
(217) 527-7887 3/8/1994 MasterCard 3554773883435
785-376-5865 7/10/1952 MasterCard 3328791931645
920-545-0796 3/24/1972 MasterCard 3576026302234
512-465-6718 11/6/1983 Visa 9162201345754
806-755-0390 5/26/1952 Visa 4854691834314
407-511-2151 9/27/1962 Visa 7165707504945
(940) 544-8557 2/22/1950 MasterCard 5686467833095
808-262-1515 4/26/1963 MasterCard 5283759556494
(270) 758-6685 7/28/1946 Visa 4855996000374
(413) 995-8438 7/29/1938 Visa 5391103523815
(276) 531-7109 3/7/1938 MasterCard 5480447170917

FIG. 14



phone			ccv
(919) 257-56			
918 915-85			
360-224-05			
(786) 222-89			
(201) 858-74			
678 218-78			
(419) 675-91			
(914) 28515			
305 994-81			
540 357-66			
(610) 673-03			
425 328-245			
(217) 527-7887	3/8/1994	MasterCard	3554773883435
785 376-5865	7/10/1952	MasterCard	3328791931645
920 545-0796	3/24/1972	MasterCard	3576026302234
512 465-6718	11/6/1983	Visa	9162201345754
806 755-0390	5/26/1952	Visa	4854691834314
407 511-2151	9/27/1962	Visa	7165707504945
(940) 544-8557	2/22/1950	MasterCard	5686467833095
808 262-1515	4/26/1963	MasterCard	5283759556494
(270) 758-6685	7/28/1946	Visa	4855996000374
(413) 995-8438	7/29/1938	Visa	5391103523815
(276) 531-7109	3/7/1938	MasterCard	5480447170917

FIG. 15

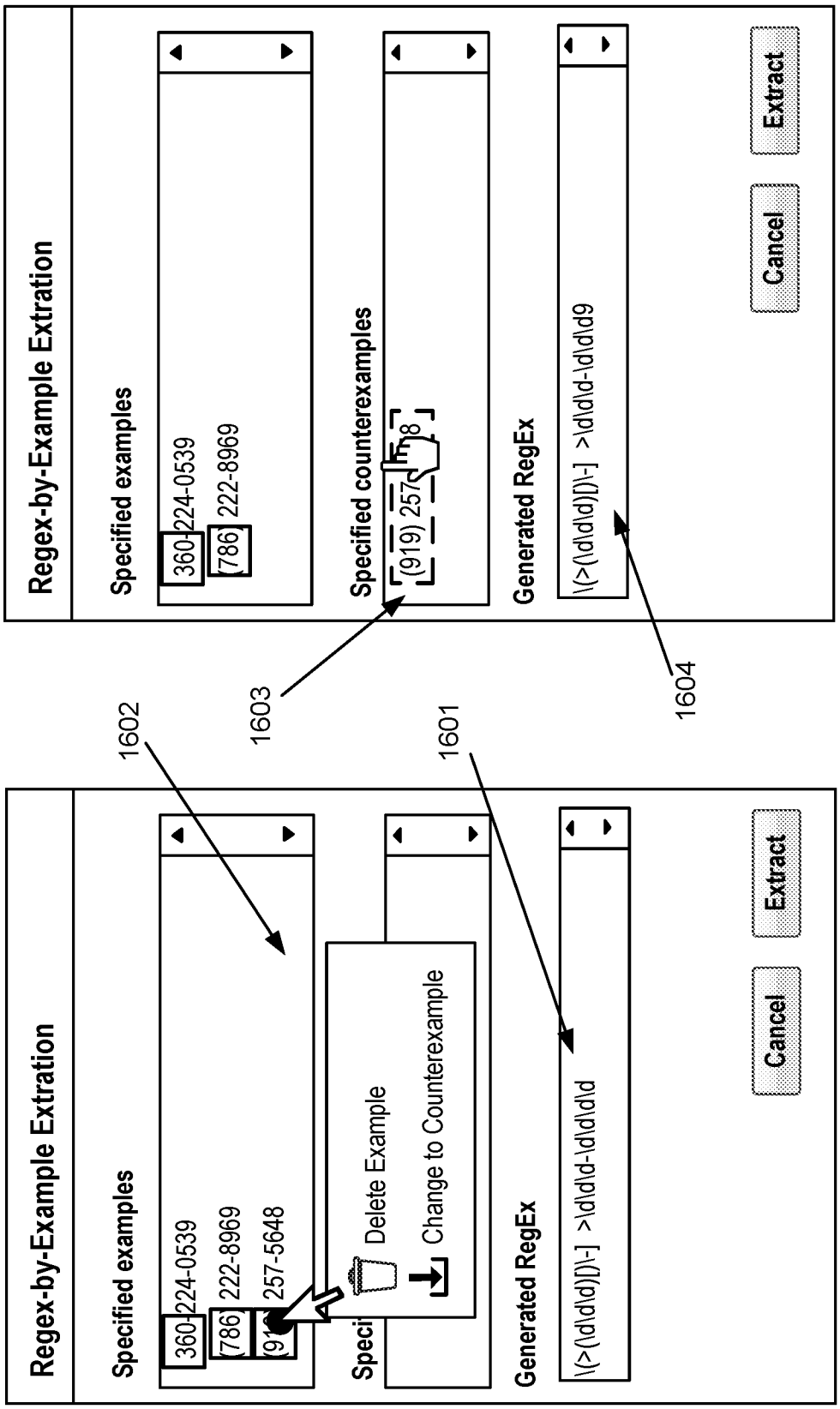


FIG. 16A

FIG. 16B

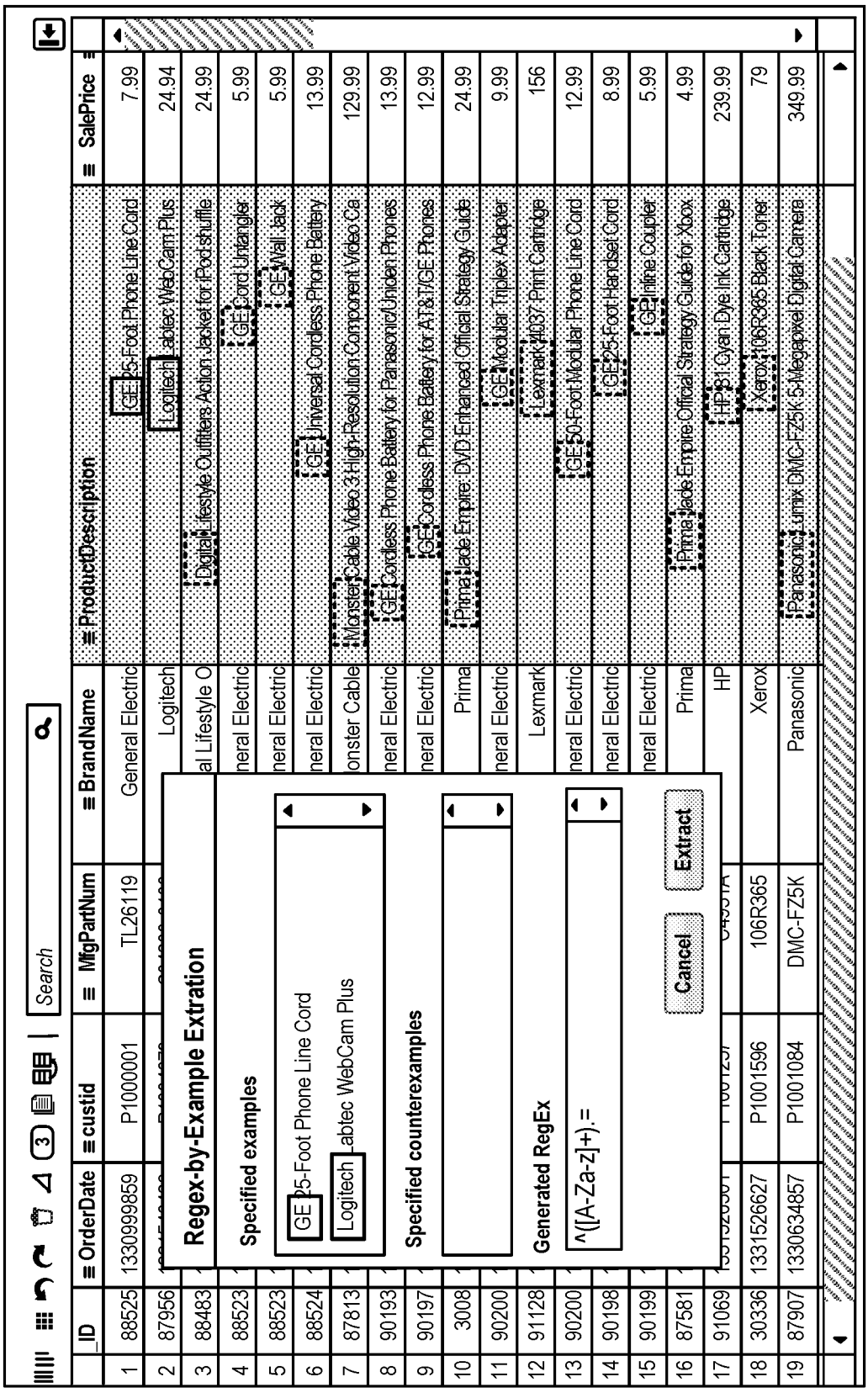


FIG. 17

SUBSTITUTE SHEET (RULE 26)

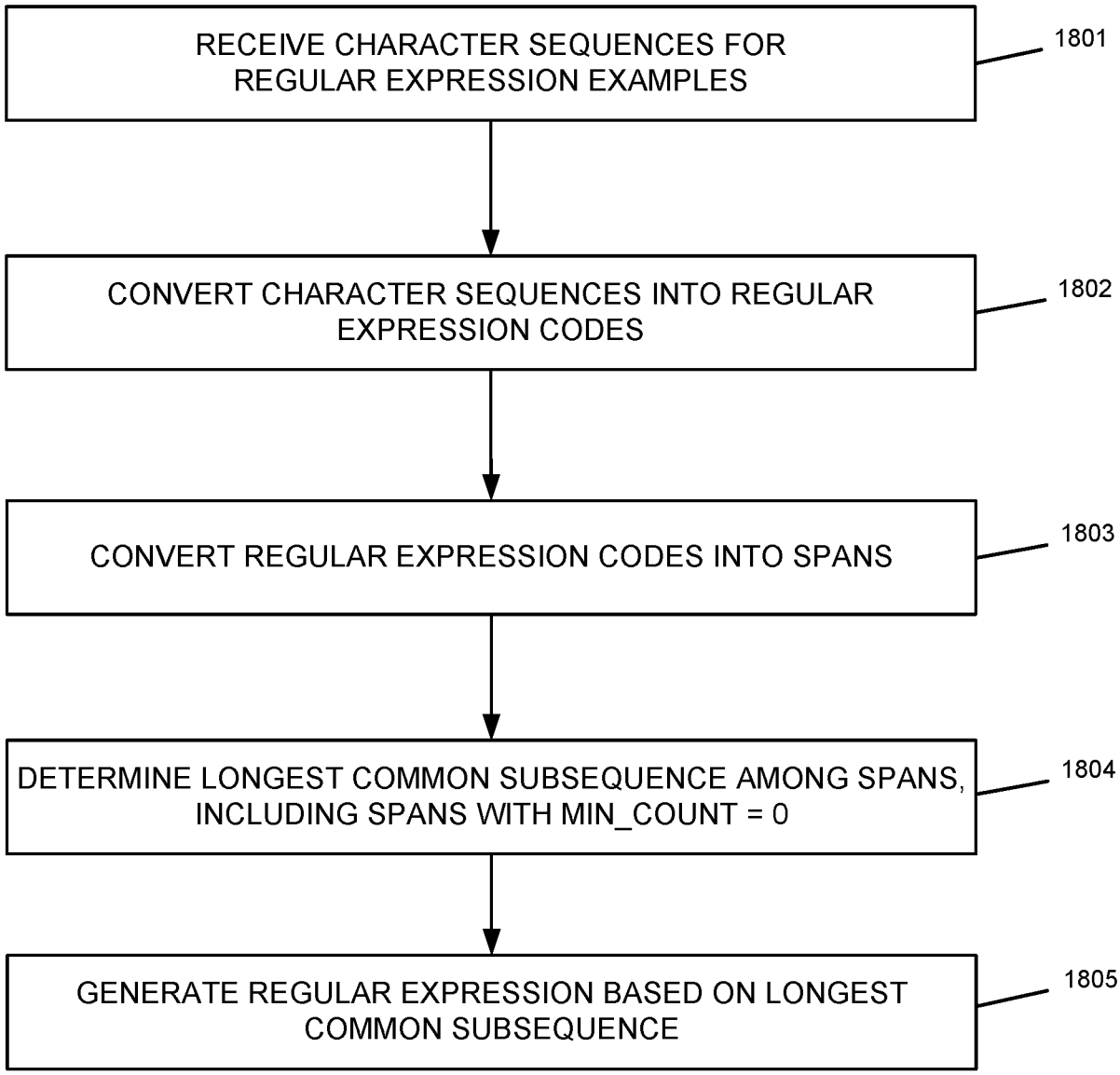

RECEIVE CHARACTER SEQUENCES FOR REGULAR EXPRESSION EXAMPLES
1801
CONVERT CHARACTER SEQUENCES INTO REGULAR EXPRESSION CODES
1802
CONVERT REGULAR EXPRESSION CODES INTO SPANS
1803
DETERMINE LONGEST COMMON SUBSEQUENCE AMONG SPANS, INCLUDING SPANS WITH MIN_COUNT = 0
1804
GENERATE REGULAR EXPRESSION BASED ON LONGEST COMMON SUBSEQUENCE
1805

FIG. 18

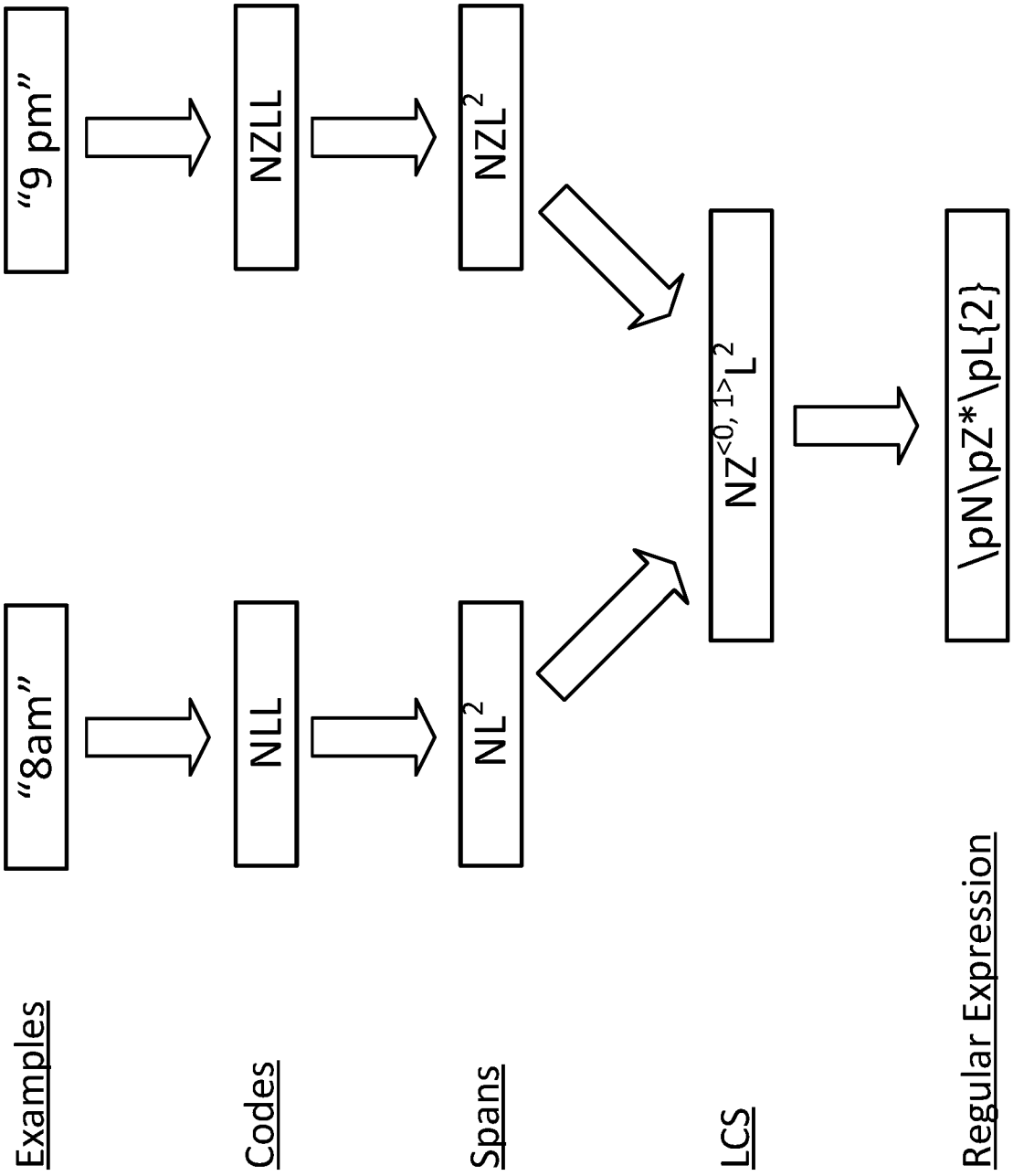

Examples
Codes
Spans
LCS
Regular Expression
"8am"
"9 pm"
NLL
NZLL
NL2
NZL2
NZ<0, 1>L2
\pN\pZ*\pL{2}

FIG. 19

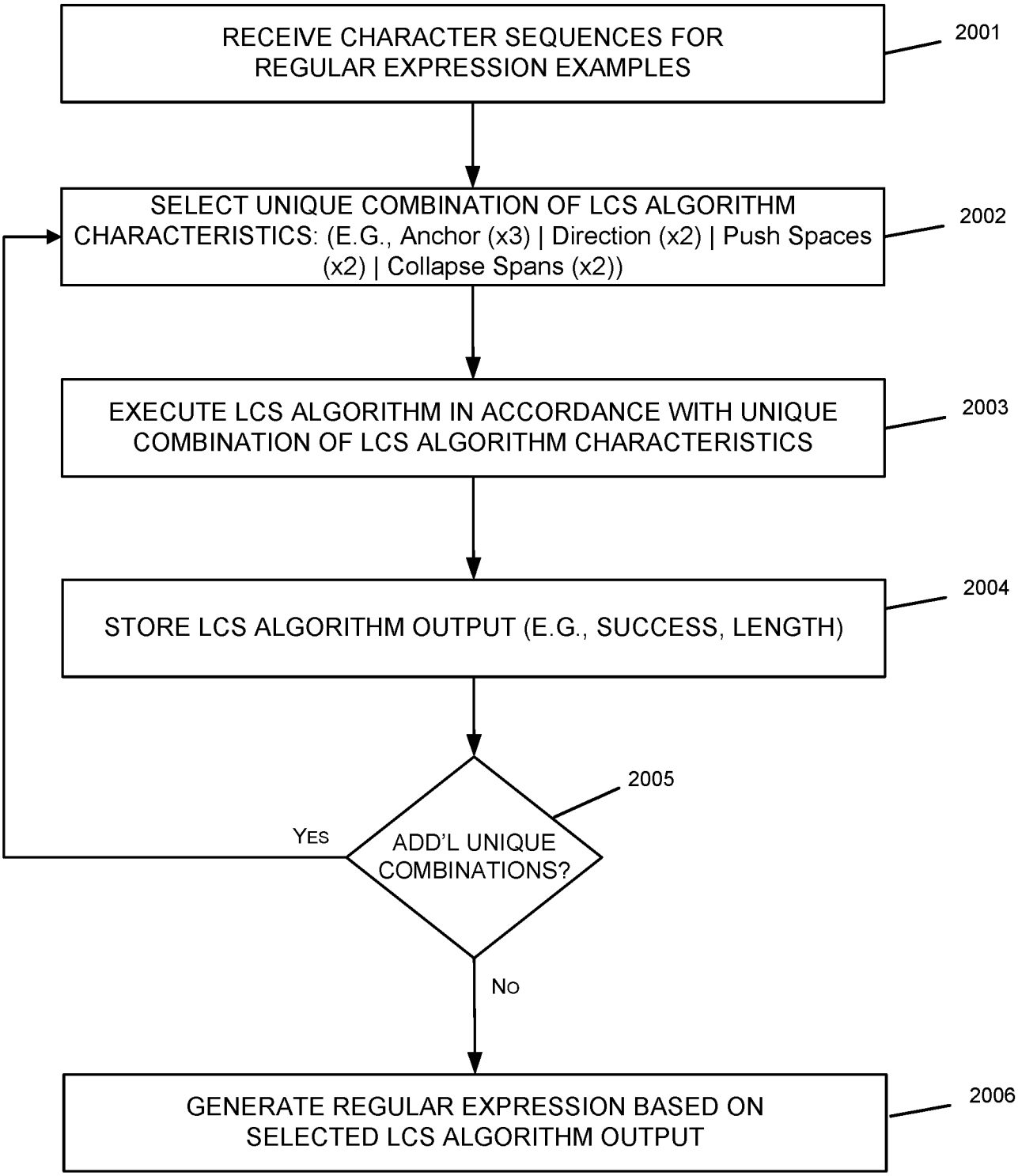

RECEIVE CHARACTER SEQUENCES FOR REGULAR EXPRESSION EXAMPLES
2001
SELECT UNIQUE COMBINATION OF LCS ALGORITHM CHARACTERISTICS: (E.G., Anchor (x3) | Direction (x2) | Push Spaces (x2) | Collapse Spans (x2))
2002
EXECUTE LCS ALGORITHM IN ACCORDANCE WITH UNIQUE COMBINATION OF LCS ALGORITHM CHARACTERISTICS
2003
STORE LCS ALGORITHM OUTPUT (E.G., SUCCESS, LENGTH)
2004
ADD'L UNIQUE COMBINATIONS?
2005
YES
NO
GENERATE REGULAR EXPRESSION BASED ON SELECTED LCS ALGORITHM OUTPUT
2006

FIG. 20

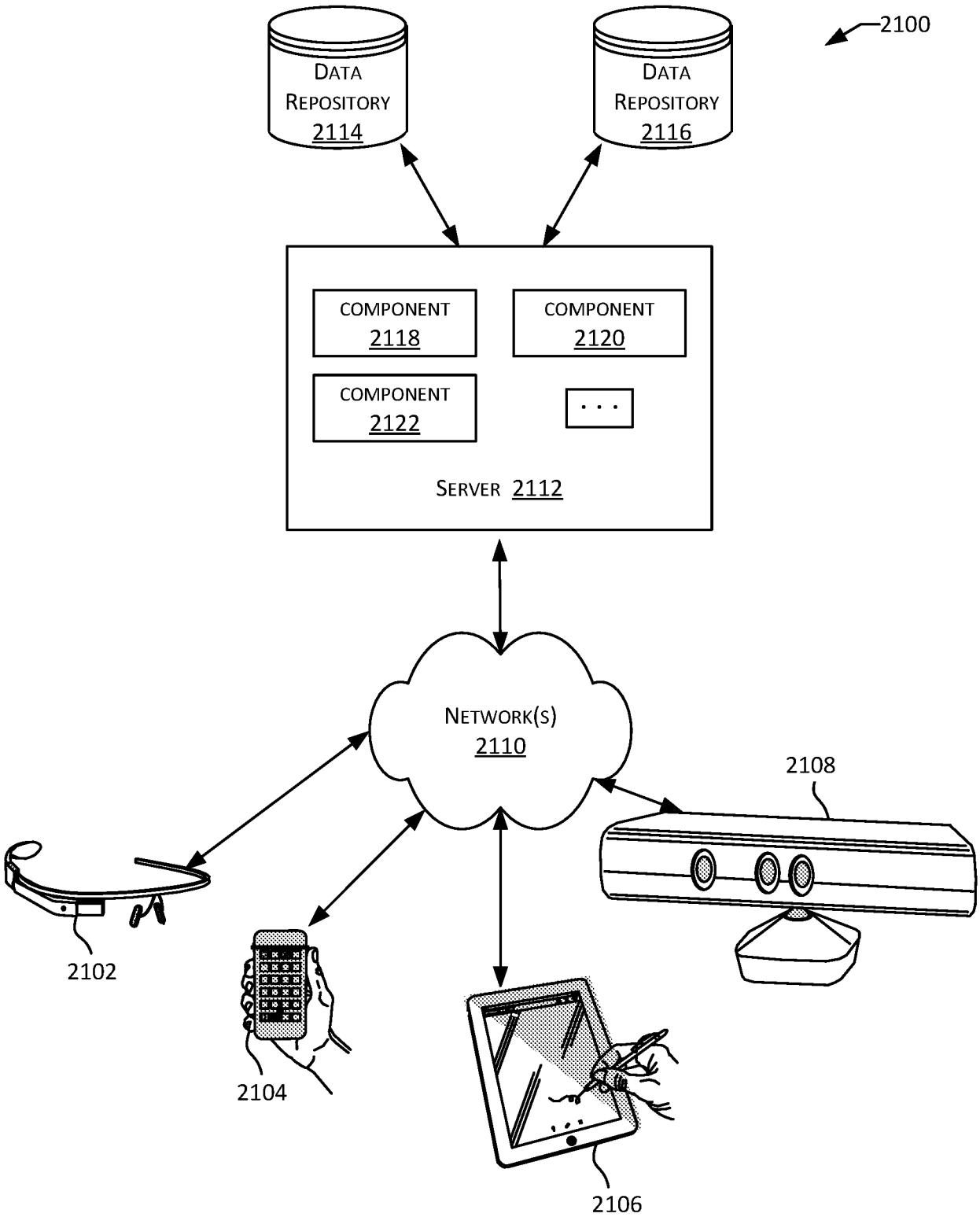

2100
DATA REPOSITORY 2114
DATA REPOSITORY 2116
COMPONENT 2118
COMPONENT 2120
COMPONENT 2122
. . .
SERVER 2112
NETWORK(S) 2110
2102
2104
2106
2108

FIG. 21

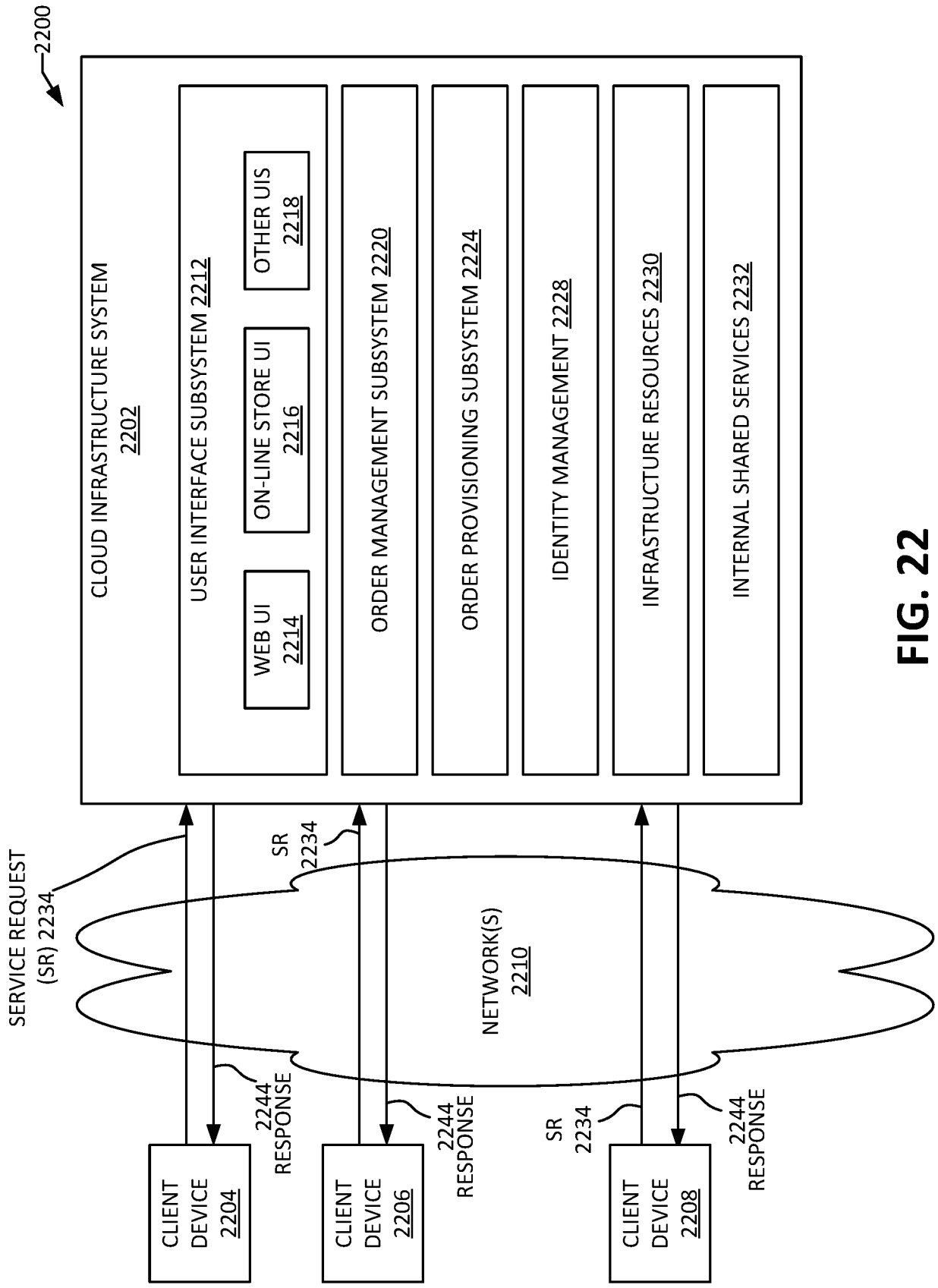

2200
CLOUD INFRASTRUCTURE SYSTEM
2202
USER INTERFACE SUBSYSTEM 2212
WEB UI 2214
ON-LINE STORE UI 2216
OTHER UIS 2218
ORDER MANAGEMENT SUBSYSTEM 2220
ORDER PROVISIONING SUBSYSTEM 2224
IDENTITY MANAGEMENT 2228
INFRASTRUCTURE RESOURCES 2230
INTERNAL SHARED SERVICES 2232
SERVICE REQUEST (SR) 2234
SR 2234
NETWORK(S) 2210
2244 RESPONSE
2244 RESPONSE
SR 2234
2244 RESPONSE
CLIENT DEVICE 2204
CLIENT DEVICE 2206
CLIENT DEVICE 2208

FIG. 22

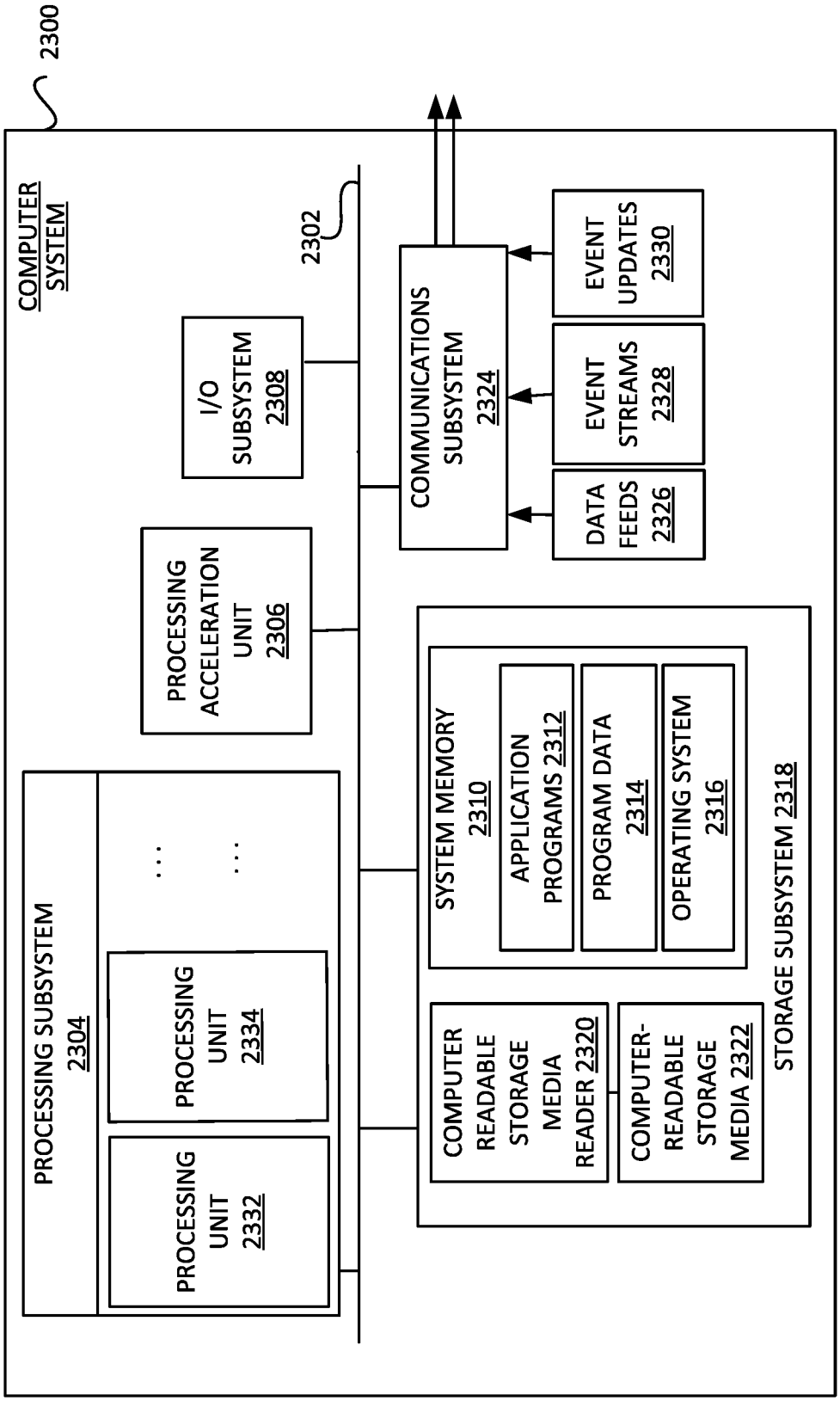

2300
COMPUTER SYSTEM
2302
I/O SUBSYSTEM 2308
PROCESSING ACCELERATION UNIT 2306
PROCESSING SUBSYSTEM 2304
PROCESSING UNIT 2332
PROCESSING UNIT 2334
COMMUNICATIONS SUBSYSTEM 2324
EVENT UPDATES 2330
EVENT STREAMS 2328
DATA FEEDS 2326
STORAGE SUBSYSTEM 2318
SYSTEM MEMORY 2310
APPLICATION PROGRAMS 2312
PROGRAM DATA 2314
OPERATING SYSTEM 2316
COMPUTER READABLE STORAGE MEDIA READER 2320
COMPUTER-READABLE STORAGE MEDIA 2322

FIG. 23

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2019/036829

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F17/22 G06F16/332 G06F16/33
ADD.

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2015/012812 A1 (RECOMMIND INC [US]) 29 January 2015 (2015-01-29) figures 6,7 paragraph [0039] - paragraph [0042] paragraph [0044] - paragraph [0045] paragraph [0058] paragraph [0059] - paragraph [0061] paragraph [0060] - paragraph [0065] paragraph [0072] - paragraph [0074] paragraph [0089] - paragraph [0091] ----- -/--	1-20

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

[X] See patent family annex.

* Special categories of cited documents :

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
14 August 2019	23/08/2019

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

1

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

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2019/036829

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2009/070327 A1 (LOESER ALEXANDER STEPHAN [DE] ET AL) 12 March 2009 (2009-03-12) the whole document paragraph [0028] paragraph [0035] paragraph [0036] paragraph [0037] - paragraph [0038] paragraph [0039] - paragraph [0041] paragraph [0036] - paragraph [0042] paragraph [0049] - paragraph [0056] -----	1-20
X	US 2017/083613 A1 (ATASU KUBILAY [CH]) 23 March 2017 (2017-03-23) the whole document -----	1-20
X	US 8 892 580 B2 (SYROWITZ CLINTON [US]; GANDHI MAUKTIK [US] ET AL.) 18 November 2014 (2014-11-18) the whole document -----	1-20
A	BERGROTH L ET AL: "A survey of longest common subsequence algorithms", STRING PROCESSING AND INFORMATION RETRIEVAL, 2000. SPIRE 2000. PROCEED INGS. SEVENTH INTERNATIONAL SYMPOSIUM ON SEPT. 27-29, 2000, PISCATAWAY, NJ, USA,IEEE, 27 September 2000 (2000-09-27), pages 39-48, XP010517586, ISBN: 978-0-7695-0746-0 the whole document -----	1-20

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2019/036829

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2015012812 A1	29-01-2015	US 8856642 B1	07-10-2014
		US 2015026559 A1	22-01-2015
		US 2018260370 A1	13-09-2018
		WO 2015012812 A1	29-01-2015
US 2009070327 A1	12-03-2009	US 2009070327 A1	12-03-2009
		US 2009070328 A1	12-03-2009
US 2017083613 A1	23-03-2017	US 2017083613 A1	23-03-2017
		US 2017083616 A1	23-03-2017
US 8892580 B2	18-11-2014	NONE	