



(51) International Patent Classification:

G06F 16/35 (2019.01)

(21) International Application Number:

PCT/US2020/031170

(22) International Filing Date:

01 May 2020 (01.05.2020)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

16/402,800

03 May 2019 (03.05.2019)

US

(71) Applicant: **SERVICENOW, INC.** [US/US]; 2225 Lawson

Lane, Santa Clara, California 95054 (US).

(72) Inventors: **JAYARAMAN, Baskar**; ServiceNow, Inc.,

2225 Lawson Lane, Santa Clara, California 95054 (US).

THAKUR, Aniruddha Madhusudhan; ServiceNow, Inc.,

2225 Lawson Lane, Santa Clara, California 95054 (US).

GOVINDARAJAN, Kannan; ServiceNow, Inc., 2225

Lawson Lane, Santa Clara, California 95054 (US). **WONG,**

Andrew Kai Chiu; ServiceNow, Inc., 2225 Lawson Lane,

Santa Clara, California 95054 (US). **PALAPUDI, Sriram**;

ServiceNow, Inc., 2225 Lawson Lane, Santa Clara, Califor-

nia 95054 (US).

(74) Agent: **FLETCHER, Michael G.** et al.; P.O. Box 692289,

Houston, Texas 77269 (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, ST, SV, SY, TH, TJ, TM, TN, TR,

TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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: CENTRALIZED MACHINE LEARNING PREDICTOR FOR A REMOTE NETWORK MANAGEMENT PLATFORM

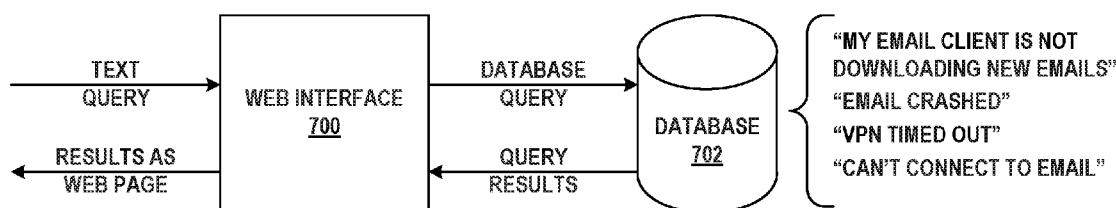


FIG. 7

(57) Abstract: A remote network management platform is provided that includes an end-user computational instance dedicated to a managed network, a training computational instance, and a prediction computational instance. The training instance is configured to receive a corpus of textual records from the end-user instance and to determine therefrom a machine learning (ML) model to determine the numerical similarity between input textual records and textual records in the corpus of textual records. The prediction instance is configured to receive the ML model and an additional textual record from the end-user instance, to use the ML model to determine respective numerical similarities between the additional textual record and the textual records in the corpus of textual records, and to transmit, based on the respective numerical similarities, representations of one or more of the textual records in the corpus of textual records to the end-user computational instance.

Declarations under Rule 4.17:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

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

CENTRALIZED MACHINE LEARNING PREDICTOR FOR A REMOTE NETWORK MANAGEMENT PLATFORM

BACKGROUND

[001] It is beneficial in a variety of applications to determine a similarity between samples of text or to otherwise compare samples of text. This can be done in order to identify, from a corpus of text samples, one or more samples that are similar in some respect to a target sample of text. For example, a target sample of text could contain a user's description of an information technology problem, and the target sample could be compared to articles in a database that describe a variety of information technology topics and/or solutions to information technology problems. By comparing the user's description of the problem to the articles in the database, one or more "most relevant" or otherwise similar articles can be provided to the user and/or to a technician in order to efficiently guide them to a solution.

SUMMARY

[001] Natural language processing (NLP) or other methods can be used to compare samples of text. This can be done in order to find patterns within the samples of text, to identify samples that are relevant to a query or otherwise similar to a target sample of text, or to provide some other benefit. For example, a query can include a sample of text describing a problem that a user is experiencing. Numerical similarity values can then be determined between the query text and the text of articles within a database, so that one or more articles that are most similar to the query may be provided to the user. This comparison may improve the quality of results provided to the user and/or reduce the amount of time spent by the user before determining and implementing a solution to the problem.

[002] However, machine learning (ML) models used to generate such degrees of numerical similarity between samples of text can be computationally expensive, requiring large amounts of processor cycles, local memory (e.g., processor cache, RAM), database storage, or other computational resources. The algorithms used to generate the ML models themselves can also be computationally expensive. Accordingly, training an ML model and/or applying the ML model to new textual records can result in a slowdown of other services being provided by a server or other computational resource. For example, a server in a remote network management system could be tasked with providing a web portal for users to submit incident reports and access knowledgebase articles, managing servers or other resources that are part of the end-user network being managed by the server, maintaining a database of received incident reports and knowledgebase articles, periodically training ML models based on the stored incident reports, and using the trained ML models to generate numerical similarity values between newly received incident reports and other incident reports and knowledgebase articles in the database. In such an example, the server training the ML model and/or executing the ML model to generate numerical similarities can result in slowing or failure of the other services, out-of-memory errors, or other unwanted effects on the operation of the server.

[003] To address this issue, the remote network management system can include separate computational instances, using separate sets of computational resources of the remote network management system, to provide different services. For example, an end-user instance can be configured to provide a web portal for users to submit incident reports and access knowledgebase articles, serve web pages, manage servers or other resources that are part of a network associated with the end-user, and maintain a database of received incident reports and knowledgebase articles. Such an end-user instance could request, from a training instance of the remote network management system, that an ML model be generated to predict numerical

similarities, clusters, or other relationships between incident reports, knowledgebase articles, or other textual records associated with the end-user instance. The end-user instance could, subsequently, request that a prediction instance of the remote network management system use the ML model to determine numerical similarities between textual records, for example between a newly-generated incident report and other incident reports associated with the end-user instance.

[004] By assigning different sets of computational resources to the end-user instance, the training instance, and the prediction instance, the impact of generating ML models and/or using such models to generate numerical similarities on the ongoing operations of the end-user instance can be reduced. Accordingly, the level of service provided by the end-user instance can be improved. For example, the latency of serving recorded incident reports and/or knowledgebase articles can be reduced. Additionally, the set of computational resources assigned to each computational instance can be optimized for their respective assigned functions. For example, training computational instances and prediction computational instances can be assigned more local memory (e.g., RAM, on-processor cache) to facilitate training and execution of ML models while the end-user instance can be assigned additional non-volatile storage to facilitate maintaining databases of incident reports, knowledgebase articles, server logs, or other information related to a managed network associated with the end-user instance. Further, the overall allocation of computational resources within a remote network management system that manages multiple different end-user networks can be improved by sharing a single training and/or prediction computational instance between multiple different end-user instances and/or by allocating a population of training and/or prediction computational instances to a population of end-user computational instances according to ongoing usage statistics.

[005] Note that the methods described herein to determine similarity between textual records, and to perform clustering based on such determined similarities, can be applied to other types of records. For example, records including both textual aspects (e.g., text fields of an incident report that describe a problem, a solution to the problem, or some other information) and non-textual aspects (e.g., dates, geographical locations, enumerated categories, an identifier related to a user or technician, or some other non-textual information of an incident report) could be clustered by an ML model based on both the textual and non-textual information. Indeed, the methods described herein could be applied to determine numerical similarities and/or to perform clustering based on such determined numerical similarities for completely non-textual records (e.g., combined genotype/phenotype information for individuals receiving a treatment, historical usage statistics for users of a media or network service, etc.).

[006] Accordingly, a first example embodiment may involve a remote network management platform including: (i) an end-user computational instance comprising a first set of computational resources of the remote network management platform and dedicated to a managed network; (ii) a training computational instance comprising a second set of computational resources of the remote network management platform; and (iii) a prediction computational instance comprising a third set of computational resources of the remote network management platform. The training computational instance is configured to perform operations including: receiving, from the end-user computational instance, a corpus of textual records; training, based on the corpus of textual records, an ML model to determine a degree of numerical similarity between input textual records and textual records in the corpus of textual records; and transmitting, to the end-user computational instance, the ML model. The prediction computational instance is configured to perform operations including: receiving, from the end-user computational instance, an additional textual record; receiving, from the end-user

computational instance, the ML model; determining, by the ML model, respective numerical similarities between the additional textual record and the textual records in the corpus of textual records; and, based on the respective numerical similarities, transmitting, to the end-user computational instance, representations of one or more of the textual records in the corpus of textual records.

[007] In a second example embodiment, a method may include: (i) operating a first set of computational resources of a remote network management platform to provide an end-user computational instance that is dedicated to a managed network; (ii) operating a second set of computational resources of the remote network management platform to provide a training computational instance; and (iii) operating a third set of computational resources of the remote network management platform to provide a prediction computational instance. The training computational instance is configured to perform operations including: receiving, from the end-user computational instance, a corpus of textual records; training, based on the corpus of textual records, an ML model to determine a degree of numerical similarity between input textual records and textual records in the corpus of textual records; and transmitting, to the end-user computational instance, the ML model. The prediction computational instance is configured to perform operations including: receiving, from the end-user computational instance, an additional textual record; receiving, from the end-user computational instance, the ML model; determining, by the ML model, respective numerical similarities between the additional textual record and the textual records in the corpus of textual records; and, based on the respective numerical similarities, transmitting, to the end-user computational instance, representations of one or more of the textual records in the corpus of textual records.

[008] In a third example embodiment, an article of manufacture may include a non-transitory computer-readable medium, having stored thereon program instructions that, upon

execution by a computing system, cause the computing system to perform operations in accordance with the second example embodiment.

[009] In a fourth example embodiment, a computing system may include at least one processor, as well as memory and program instructions. The program instructions may be stored in the memory, and upon execution by the at least one processor, cause the computing system to perform operations in accordance with the second example embodiment.

[010] In a fifth example embodiment, a system may include various means for carrying out each of the operations of the second example embodiment.

[011] These, as well as other embodiments, aspects, advantages, and alternatives, will become apparent to those of ordinary skill in the art by reading the following detailed description, with reference where appropriate to the accompanying drawings. Further, this summary and other descriptions and figures provided herein are intended to illustrate embodiments by way of example only and, as such, that numerous variations are possible. For instance, structural elements and process steps can be rearranged, combined, distributed, eliminated, or otherwise changed, while remaining within the scope of the embodiments as claimed.

BRIEF DESCRIPTION OF THE DRAWINGS

[012] Figure 1 illustrates a schematic drawing of a computing device, in accordance with example embodiments.

[013] Figure 2 illustrates a schematic drawing of a server device cluster, in accordance with example embodiments.

[014] Figure 3 depicts a remote network management architecture, in accordance with example embodiments.

[015] Figure 4 depicts a communication environment involving a remote network management architecture, in accordance with example embodiments.

[016] Figure 5A depicts another communication environment involving a remote network management architecture, in accordance with example embodiments.

[017] Figure 5B is a flow chart, in accordance with example embodiments.

[018] Figure 6 depicts an incident report, in accordance with example embodiments.

[019] Figure 7 depicts a database query architecture, in accordance with example embodiments.

[020] Figure 8A depicts an artificial neural network (ANN) configured for learning the contextual meanings of words, in accordance with example embodiments.

[021] Figure 8B depicts a set of training data for the ANN of figure 9A, in accordance with example embodiments.

[022] Figure 9A depicts an artificial neural network (ANN) configured for learning the contextual meanings of words, in accordance with example embodiments.

[023] Figure 9B depicts a set of training data for the ANN of figure 9A, in accordance with example embodiments.

[024] Figure 10A depicts training an ANN for paragraph vectors, in accordance with example embodiments.

[025] Figure 10B depicts training an ANN for paragraph vectors, in accordance with example embodiments.

[026] Figure 10C depicts training an ANN for paragraph vectors, in accordance with example embodiments.

[027] Figure 10D depicts using a trained ANN to determine the paragraph vector of a previously unseen paragraph, in accordance with example embodiments.

[028] Figure 11 depicts a remote network management architecture with centralized training and prediction, in accordance with example embodiments.

[029] Figure 12 is a flow chart, in accordance with example embodiments.

DETAILED DESCRIPTION

[030] Example methods, devices, and systems are described herein. It should be understood that the words “example” and “exemplary” are used herein to mean “serving as an example, instance, or illustration.” Any embodiment or feature described herein as being an “example” or “exemplary” is not necessarily to be construed as preferred or advantageous over other embodiments or features unless stated as such. Thus, other embodiments can be utilized and other changes can be made without departing from the scope of the subject matter presented herein.

[031] Accordingly, the example embodiments described herein are not meant to be limiting. It will be readily understood that the aspects of the present disclosure, as generally described herein, and illustrated in the figures, can be arranged, substituted, combined, separated, and designed in a wide variety of different configurations. For example, the separation of features into “client” and “server” components may occur in a number of ways.

[032] Further, unless context suggests otherwise, the features illustrated in each of the figures may be used in combination with one another. Thus, the figures should be generally viewed as component aspects of one or more overall embodiments, with the understanding that not all illustrated features are necessary for each embodiment.

[033] Additionally, any enumeration of elements, blocks, or steps in this specification or the claims is for purposes of clarity. Thus, such enumeration should not be interpreted to require

or imply that these elements, blocks, or steps adhere to a particular arrangement or are carried out in a particular order.

I. Introduction

[034] A large enterprise is a complex entity with many interrelated operations. Some of these are found across the enterprise, such as human resources (HR), supply chain, information technology (IT), and finance. However, each enterprise also has its own unique operations that provide essential capabilities and/or create competitive advantages.

[035] To support widely-implemented operations, enterprises typically use off-the-shelf software applications, such as customer relationship management (CRM) and human capital management (HCM) packages. However, they may also need custom software applications to meet their own unique requirements. A large enterprise often has dozens or hundreds of these custom software applications. Nonetheless, the advantages provided by the embodiments herein are not limited to large enterprises and may be applicable to an enterprise, or any other type of organization, of any size.

[036] Many such software applications are developed by individual departments within the enterprise. These range from simple spreadsheets to custom-built software tools and databases. But the proliferation of siloed custom software applications has numerous disadvantages. It negatively impacts an enterprise's ability to run and grow its operations, innovate, and meet regulatory requirements. The enterprise may find it difficult to integrate, streamline and enhance its operations due to lack of a single system that unifies its subsystems and data.

[037] To efficiently create custom applications, enterprises would benefit from a remotely-hosted application platform that eliminates unnecessary development complexity. The goal of such a platform would be to reduce time-consuming, repetitive application development

tasks so that software engineers and individuals in other roles can focus on developing unique, high-value features.

[038] In order to achieve this goal, the concept of Application Platform as a Service (aPaaS) is introduced, to intelligently automate workflows throughout the enterprise. An aPaaS system is hosted remotely from the enterprise, but may access data, applications, and services within the enterprise by way of secure connections. Such an aPaaS system may have a number of advantageous capabilities and characteristics. These advantages and characteristics may be able to improve the enterprise's operations and workflow for IT, HR, CRM, customer service, application development, and security.

[039] The aPaaS system may support development and execution of model-view-controller (MVC) applications. MVC applications divide their functionality into three interconnected parts (model, view, and controller) in order to isolate representations of information from the manner in which the information is presented to the user, thereby allowing for efficient code reuse and parallel development. These applications may be web-based, and offer create, read, update, delete (CRUD) capabilities. This allows new applications to be built on a common application infrastructure.

[040] The aPaaS system may support standardized application components, such as a standardized set of widgets for graphical user interface (GUI) development. In this way, applications built using the aPaaS system have a common look and feel. Other software components and modules may be standardized as well. In some cases, this look and feel can be branded or skinned with an enterprise's custom logos and/or color schemes.

[041] The aPaaS system may support the ability to configure the behavior of applications using metadata. This allows application behaviors to be rapidly adapted to meet specific needs. Such an approach reduces development time and increases flexibility. Further,

the aPaaS system may support GUI tools that facilitate metadata creation and management, thus reducing errors in the metadata.

[042] The aPaaS system may support clearly-defined interfaces between applications, so that software developers can avoid unwanted inter-application dependencies. Thus, the aPaaS system may implement a service layer in which persistent state information and other data are stored.

[043] The aPaaS system may support a rich set of integration features so that the applications thereon can interact with legacy applications and third-party applications. For instance, the aPaaS system may support a custom employee-onboarding system that integrates with legacy HR, IT, and accounting systems.

[044] The aPaaS system may support enterprise-grade security. Furthermore, since the aPaaS system may be remotely hosted, it should also utilize security procedures when it interacts with systems in the enterprise or third-party networks and services hosted outside of the enterprise. For example, the aPaaS system may be configured to share data amongst the enterprise and other parties to detect and identify common security threats.

[045] Other features, functionality, and advantages of an aPaaS system may exist. This description is for purpose of example and is not intended to be limiting.

[046] As an example of the aPaaS development process, a software developer may be tasked to create a new application using the aPaaS system. First, the developer may define the data model, which specifies the types of data that the application uses and the relationships therebetween. Then, via a GUI of the aPaaS system, the developer enters (e.g., uploads) the data model. The aPaaS system automatically creates all of the corresponding database tables, fields, and relationships, which can then be accessed via an object-oriented services layer.

[047] In addition, the aPaaS system can also build a fully-functional MVC application with client-side interfaces and server-side CRUD logic. This generated application may serve as the basis of further development for the user. Advantageously, the developer does not have to spend a large amount of time on basic application functionality. Further, since the application may be web-based, it can be accessed from any Internet-enabled client device. Alternatively or additionally, a local copy of the application may be able to be accessed, for instance, when Internet service is not available.

[048] The aPaaS system may also support a rich set of pre-defined functionality that can be added to applications. These features include support for searching, email, templating, workflow design, reporting, analytics, social media, scripting, mobile-friendly output, and customized GUIs.

[049] The following embodiments describe architectural and functional aspects of example aPaaS systems, as well as the features and advantages thereof.

II. Example Computing Devices and Cloud-Based Computing Environments

[050] Figure 1 is a simplified block diagram exemplifying a computing device 100, illustrating some of the components that could be included in a computing device arranged to operate in accordance with the embodiments herein. Computing device 100 could be a client device (e.g., a device actively operated by a user), a server device (e.g., a device that provides computational services to client devices), or some other type of computational platform. Some server devices may operate as client devices from time to time in order to perform particular operations, and some client devices may incorporate server features.

[051] In this example, computing device 100 includes processor 102, memory 104, network interface 106, and an input / output unit 108, all of which may be coupled by a system

bus 110 or a similar mechanism. In some embodiments, computing device 100 may include other components and/or peripheral devices (e.g., detachable storage, printers, and so on).

[052] Processor 102 may be one or more of any type of computer processing element, such as a central processing unit (CPU), a co-processor (e.g., a mathematics, graphics, or encryption co-processor), a digital signal processor (DSP), a network processor, and/or a form of integrated circuit or controller that performs processor operations. In some cases, processor 102 may be one or more single-core processors. In other cases, processor 102 may be one or more multi-core processors with multiple independent processing units. Processor 102 may also include register memory for temporarily storing instructions being executed and related data, as well as cache memory for temporarily storing recently-used instructions and data.

[053] Memory 104 may be any form of computer-usable memory, including but not limited to random access memory (RAM), read-only memory (ROM), and non-volatile memory (e.g., flash memory, hard disk drives, solid state drives, compact discs (CDs), digital video discs (DVDs), and/or tape storage). Thus, memory 104 represents both main memory units, as well as long-term storage. Other types of memory may include biological memory.

[054] Memory 104 may store program instructions and/or data on which program instructions may operate. By way of example, memory 104 may store these program instructions on a non-transitory, computer-readable medium, such that the instructions are executable by processor 102 to carry out any of the methods, processes, or operations disclosed in this specification or the accompanying drawings.

[055] As shown in Figure 1, memory 104 may include firmware 104A, kernel 104B, and/or applications 104C. Firmware 104A may be program code used to boot or otherwise initiate some or all of computing device 100. Kernel 104B may be an operating system, including modules for memory management, scheduling and management of processes, input /

output, and communication. Kernel 104B may also include device drivers that allow the operating system to communicate with the hardware modules (e.g., memory units, networking interfaces, ports, and busses), of computing device 100. Applications 104C may be one or more user-space software programs, such as web browsers or email clients, as well as any software libraries used by these programs. Memory 104 may also store data used by these and other programs and applications.

[056] Network interface 106 may take the form of one or more wireline interfaces, such as Ethernet (e.g., Fast Ethernet, Gigabit Ethernet, and so on). Network interface 106 may also support communication over one or more non-Ethernet media, such as coaxial cables or power lines, or over wide-area media, such as Synchronous Optical Networking (SONET) or digital subscriber line (DSL) technologies. Network interface 106 may additionally take the form of one or more wireless interfaces, such as IEEE 802.11 (Wifi), BLUETOOTH®, global positioning system (GPS), or a wide-area wireless interface. However, other forms of physical layer interfaces and other types of standard or proprietary communication protocols may be used over network interface 106. Furthermore, network interface 106 may comprise multiple physical interfaces. For instance, some embodiments of computing device 100 may include Ethernet, BLUETOOTH®, and Wifi interfaces.

[057] Input / output unit 108 may facilitate user and peripheral device interaction with computing device 100. Input / output unit 108 may include one or more types of input devices, such as a keyboard, a mouse, a touch screen, and so on. Similarly, input / output unit 108 may include one or more types of output devices, such as a screen, monitor, printer, and/or one or more light emitting diodes (LEDs). Additionally or alternatively, computing device 100 may communicate with other devices using a universal serial bus (USB) or high-definition multimedia interface (HDMI) port interface, for example.

[058] In some embodiments, one or more computing devices like computing device 100 may be deployed to support an aPaaS architecture. The exact physical location, connectivity, and configuration of these computing devices may be unknown and/or unimportant to client devices. Accordingly, the computing devices may be referred to as “cloud-based” devices that may be housed at various remote data center locations.

[059] Figure 2 depicts a cloud-based server cluster 200 in accordance with example embodiments. In Figure 2, operations of a computing device (e.g., computing device 100) may be distributed between server devices 202, data storage 204, and routers 206, all of which may be connected by local cluster network 208. The number of server devices 202, data storages 204, and routers 206 in server cluster 200 may depend on the computing task(s) and/or applications assigned to server cluster 200.

[060] For example, server devices 202 can be configured to perform various computing tasks of computing device 100. Thus, computing tasks can be distributed among one or more of server devices 202. To the extent that these computing tasks can be performed in parallel, such a distribution of tasks may reduce the total time to complete these tasks and return a result. For purpose of simplicity, both server cluster 200 and individual server devices 202 may be referred to as a “server device.” This nomenclature should be understood to imply that one or more distinct server devices, data storage devices, and cluster routers may be involved in server device operations.

[061] Data storage 204 may be data storage arrays that include drive array controllers configured to manage read and write access to groups of hard disk drives and/or solid state drives. The drive array controllers, alone or in conjunction with server devices 202, may also be configured to manage backup or redundant copies of the data stored in data storage 204 to protect against drive failures or other types of failures that prevent one or more of server devices

202 from accessing units of data storage 204. Other types of memory aside from drives may be used.

[062] Routers 206 may include networking equipment configured to provide internal and external communications for server cluster 200. For example, routers 206 may include one or more packet-switching and/or routing devices (including switches and/or gateways) configured to provide (i) network communications between server devices 202 and data storage 204 via local cluster network 208, and/or (ii) network communications between the server cluster 200 and other devices via communication link 210 to network 212.

[063] Additionally, the configuration of routers 206 can be based at least in part on the data communication requirements of server devices 202 and data storage 204, the latency and throughput of the local cluster network 208, the latency, throughput, and cost of communication link 210, and/or other factors that may contribute to the cost, speed, fault-tolerance, resiliency, efficiency and/or other design goals of the system architecture.

[064] As a possible example, data storage 204 may include any form of database, such as a structured query language (SQL) database. Various types of data structures may store the information in such a database, including but not limited to tables, arrays, lists, trees, and tuples. Furthermore, any databases in data storage 204 may be monolithic or distributed across multiple physical devices.

[065] Server devices 202 may be configured to transmit data to and receive data from data storage 204. This transmission and retrieval may take the form of SQL queries or other types of database queries, and the output of such queries, respectively. Additional text, images, video, and/or audio may be included as well. Furthermore, server devices 202 may organize the received data into web page representations. Such a representation may take the form of a markup language, such as the hypertext markup language (HTML), the extensible markup

language (XML), or some other standardized or proprietary format. Moreover, server devices 202 may have the capability of executing various types of computerized scripting languages, such as but not limited to Perl, Python, PHP Hypertext Preprocessor (PHP), Active Server Pages (ASP), JAVASCRIPT®, and so on. Computer program code written in these languages may facilitate the providing of web pages to client devices, as well as client device interaction with the web pages.

III. Example Remote Network Management Architecture

[066] Figure 3 depicts a remote network management architecture, in accordance with example embodiments. This architecture includes three main components, managed network 300, remote network management platform 320, and third-party networks 340, all connected by way of Internet 350.

[067] Managed network 300 may be, for example, an enterprise network used by an entity for computing and communications tasks, as well as storage of data. Thus, managed network 300 may include client devices 302, server devices 304, routers 306, virtual machines 308, firewall 310, and/or proxy servers 312. Client devices 302 may be embodied by computing device 100, server devices 304 may be embodied by computing device 100 or server cluster 200, and routers 306 may be any type of router, switch, or gateway.

[068] Virtual machines 308 may be embodied by one or more of computing device 100 or server cluster 200. In general, a virtual machine is an emulation of a computing system, and mimics the functionality (e.g., processor, memory, and communication resources) of a physical computer. One physical computing system, such as server cluster 200, may support up to thousands of individual virtual machines. In some embodiments, virtual machines 308 may be managed by a centralized server device or application that facilitates allocation of physical computing resources to individual virtual machines, as well as performance and error reporting.

Enterprises often employ virtual machines in order to allocate computing resources in an efficient, as needed fashion. Providers of virtualized computing systems include VMWARE® and MICROSOFT®.

[069] Firewall 310 may be one or more specialized routers or server devices that protect managed network 300 from unauthorized attempts to access the devices, applications, and services therein, while allowing authorized communication that is initiated from managed network 300. Firewall 310 may also provide intrusion detection, web filtering, virus scanning, application-layer gateways, and other applications or services. In some embodiments not shown in Figure 3, managed network 300 may include one or more virtual private network (VPN) gateways with which it communicates with remote network management platform 320 (see below).

[070] Managed network 300 may also include one or more proxy servers 312. An embodiment of proxy servers 312 may be a server device that facilitates communication and movement of data between managed network 300, remote network management platform 320, and third-party networks 340. In particular, proxy servers 312 may be able to establish and maintain secure communication sessions with one or more computational instances of remote network management platform 320. By way of such a session, remote network management platform 320 may be able to discover and manage aspects of the architecture and configuration of managed network 300 and its components. Possibly with the assistance of proxy servers 312, remote network management platform 320 may also be able to discover and manage aspects of third-party networks 340 that are used by managed network 300.

[071] Firewalls, such as firewall 310, typically deny all communication sessions that are incoming by way of Internet 350, unless such a session was ultimately initiated from behind the firewall (i.e., from a device on managed network 300) or the firewall has been explicitly

configured to support the session. By placing proxy servers 312 behind firewall 310 (e.g., within managed network 300 and protected by firewall 310), proxy servers 312 may be able to initiate these communication sessions through firewall 310. Thus, firewall 310 might not have to be specifically configured to support incoming sessions from remote network management platform 320, thereby avoiding potential security risks to managed network 300.

[072] In some cases, managed network 300 may consist of a few devices and a small number of networks. In other deployments, managed network 300 may span multiple physical locations and include hundreds of networks and hundreds of thousands of devices. Thus, the architecture depicted in Figure 3 is capable of scaling up or down by orders of magnitude.

[073] Furthermore, depending on the size, architecture, and connectivity of managed network 300, a varying number of proxy servers 312 may be deployed therein. For example, each one of proxy servers 312 may be responsible for communicating with remote network management platform 320 regarding a portion of managed network 300. Alternatively or additionally, sets of two or more proxy servers may be assigned to such a portion of managed network 300 for purposes of load balancing, redundancy, and/or high availability.

[074] Remote network management platform 320 is a hosted environment that provides aPaaS services to users, particularly to the operators of managed network 300. These services may take the form of web-based portals, for instance. Thus, a user can securely access remote network management platform 320 from, for instance, client devices 302, or potentially from a client device outside of managed network 300. By way of the web-based portals, users may design, test, and deploy applications, generate reports, view analytics, and perform other tasks.

[075] As shown in Figure 3, remote network management platform 320 includes four computational instances 322, 324, 326, and 328. Each of these instances may represent one or more server devices and/or one or more databases that provide a set of web portals, services, and

applications (e.g., a wholly-functioning aPaaS system) available to a particular customer. In some cases, a single customer may use multiple computational instances. For example, managed network 300 may be an enterprise customer of remote network management platform 320, and may use computational instances 322, 324, and 326. The reason for providing multiple instances to one customer is that the customer may wish to independently develop, test, and deploy its applications and services. Thus, computational instance 322 may be dedicated to application development related to managed network 300, computational instance 324 may be dedicated to testing these applications, and computational instance 326 may be dedicated to the live operation of tested applications and services. A computational instance may also be referred to as a hosted instance, a remote instance, a customer instance, or by some other designation. Any application deployed onto a computational instance may be a scoped application, in that its access to databases within the computational instance can be restricted to certain elements therein (e.g., one or more particular database tables or particular rows with one or more database tables).

[076] For purpose of clarity, the disclosure herein refers to the physical hardware, software, and arrangement thereof as a “computational instance.” Note that users may colloquially refer to the graphical user interfaces provided thereby as “instances.” But unless it is defined otherwise herein, a “computational instance” is a computing system disposed within remote network management platform 320.

[077] The multi-instance architecture of remote network management platform 320 is in contrast to conventional multi-tenant architectures, over which multi-instance architectures exhibit several advantages. In multi-tenant architectures, data from different customers (e.g., enterprises) are comingled in a single database. While these customers’ data are separate from one another, the separation is enforced by the software that operates the single database. As a consequence, a security breach in this system may impact all customers’ data, creating additional

risk, especially for entities subject to governmental, healthcare, and/or financial regulation. Furthermore, any database operations that impact one customer will likely impact all customers sharing that database. Thus, if there is an outage due to hardware or software errors, this outage affects all such customers. Likewise, if the database is to be upgraded to meet the needs of one customer, it will be unavailable to all customers during the upgrade process. Often, such maintenance windows will be long, due to the size of the shared database.

[078] In contrast, the multi-instance architecture provides each customer with its own database in a dedicated computing instance. This prevents comingling of customer data, and allows each instance to be independently managed. For example, when one customer's instance experiences an outage due to errors or an upgrade, other computational instances are not impacted. Maintenance down time is limited because the database only contains one customer's data. Further, the simpler design of the multi-instance architecture allows redundant copies of each customer database and instance to be deployed in a geographically diverse fashion. This facilitates high availability, where the live version of the customer's instance can be moved when faults are detected or maintenance is being performed.

[079] In some embodiments, remote network management platform 320 may include one or more central instances, controlled by the entity that operates this platform. Like a computational instance, a central instance may include some number of physical or virtual servers and database devices. Such a central instance may serve as a repository for data that can be shared amongst at least some of the computational instances. For instance, definitions of common security threats that could occur on the computational instances, software packages that are commonly discovered on the computational instances, and/or an application store for applications that can be deployed to the computational instances may reside in a central instance.

Computational instances may communicate with central instances by way of well-defined interfaces in order to obtain this data.

[080] In order to support multiple computational instances in an efficient fashion, remote network management platform 320 may implement a plurality of these instances on a single hardware platform. For example, when the aPaaS system is implemented on a server cluster such as server cluster 200, it may operate a virtual machine that dedicates varying amounts of computational, storage, and communication resources to instances. But full virtualization of server cluster 200 might not be necessary, and other mechanisms may be used to separate instances. In some examples, each instance may have a dedicated account and one or more dedicated databases on server cluster 200. Alternatively, computational instance 322 may span multiple physical devices.

[081] In some cases, a single server cluster of remote network management platform 320 may support multiple independent enterprises. Furthermore, as described below, remote network management platform 320 may include multiple server clusters deployed in geographically diverse data centers in order to facilitate load balancing, redundancy, and/or high availability.

[082] Third-party networks 340 may be remote server devices (e.g., a plurality of server clusters such as server cluster 200) that can be used for outsourced computational, data storage, communication, and service hosting operations. These servers may be virtualized (i.e., the servers may be virtual machines). Examples of third-party networks 340 may include AMAZON WEB SERVICES® and MICROSOFT® AZURE®. Like remote network management platform 320, multiple server clusters supporting third-party networks 340 may be deployed at geographically diverse locations for purposes of load balancing, redundancy, and/or high availability.

[083] Managed network 300 may use one or more of third-party networks 340 to deploy applications and services to its clients and customers. For instance, if managed network 300 provides online music streaming services, third-party networks 340 may store the music files and provide web interface and streaming capabilities. In this way, the enterprise of managed network 300 does not have to build and maintain its own servers for these operations.

[084] Remote network management platform 320 may include modules that integrate with third-party networks 340 to expose virtual machines and managed services therein to managed network 300. The modules may allow users to request virtual resources and provide flexible reporting for third-party networks 340. In order to establish this functionality, a user from managed network 300 might first establish an account with third-party networks 340, and request a set of associated resources. Then, the user may enter the account information into the appropriate modules of remote network management platform 320. These modules may then automatically discover the manageable resources in the account, and also provide reports related to usage, performance, and billing.

[085] Internet 350 may represent a portion of the global Internet. However, Internet 350 may alternatively represent a different type of network, such as a private wide-area or local-area packet-switched network.

[086] Figure 4 further illustrates the communication environment between managed network 300 and computational instance 322, and introduces additional features and alternative embodiments. In Figure 4, computational instance 322 is replicated across data centers 400A and 400B. These data centers may be geographically distant from one another, perhaps in different cities or different countries. Each data center includes support equipment that facilitates communication with managed network 300, as well as remote users.

[087] In data center 400A, network traffic to and from external devices flows either through VPN gateway 402A or firewall 404A. VPN gateway 402A may be peered with VPN gateway 412 of managed network 300 by way of a security protocol such as Internet Protocol Security (IPSEC) or Transport Layer Security (TLS). Firewall 404A may be configured to allow access from authorized users, such as user 414 and remote user 416, and to deny access to unauthorized users. By way of firewall 404A, these users may access computational instance 322, and possibly other computational instances. Load balancer 406A may be used to distribute traffic amongst one or more physical or virtual server devices that host computational instance 322. Load balancer 406A may simplify user access by hiding the internal configuration of data center 400A, (e.g., computational instance 322) from client devices. For instance, if computational instance 322 includes multiple physical or virtual computing devices that share access to multiple databases, load balancer 406A may distribute network traffic and processing tasks across these computing devices and databases so that no one computing device or database is significantly busier than the others. In some embodiments, computational instance 322 may include VPN gateway 402A, firewall 404A, and load balancer 406A.

[088] Data center 400B may include its own versions of the components in data center 400A. Thus, VPN gateway 402B, firewall 404B, and load balancer 406B may perform the same or similar operations as VPN gateway 402A, firewall 404A, and load balancer 406A, respectively. Further, by way of real-time or near-real-time database replication and/or other operations, computational instance 322 may exist simultaneously in data centers 400A and 400B.

[089] Data centers 400A and 400B as shown in Figure 4 may facilitate redundancy and high availability. In the configuration of Figure 4, data center 400A is active and data center 400B is passive. Thus, data center 400A is serving all traffic to and from managed network 300, while the version of computational instance 322 in data center 400B is being updated in near-

real-time. Other configurations, such as one in which both data centers are active, may be supported.

[090] Should data center 400A fail in some fashion or otherwise become unavailable to users, data center 400B can take over as the active data center. For example, domain name system (DNS) servers that associate a domain name of computational instance 322 with one or more Internet Protocol (IP) addresses of data center 400A may re-associate the domain name with one or more IP addresses of data center 400B. After this re-association completes (which may take less than one second or several seconds), users may access computational instance 322 by way of data center 400B.

[091] Figure 4 also illustrates a possible configuration of managed network 300. As noted above, proxy servers 312 and user 414 may access computational instance 322 through firewall 310. Proxy servers 312 may also access configuration items 410. In Figure 4, configuration items 410 may refer to any or all of client devices 302, server devices 304, routers 306, and virtual machines 308, any applications or services executing thereon, as well as relationships between devices, applications, and services. Thus, the term “configuration items” may be shorthand for any physical or virtual device, or any application or service remotely discoverable or managed by computational instance 322, or relationships between discovered devices, applications, and services. Configuration items may be represented in a configuration management database (CMDB) of computational instance 322.

[092] As noted above, VPN gateway 412 may provide a dedicated VPN to VPN gateway 402A. Such a VPN may be helpful when there is a significant amount of traffic between managed network 300 and computational instance 322, or security policies otherwise suggest or require use of a VPN between these sites. In some embodiments, any device in managed network 300 and/or computational instance 322 that directly communicates via the

VPN is assigned a public IP address. Other devices in managed network 300 and/or computational instance 322 may be assigned private IP addresses (e.g., IP addresses selected from the 10.0.0.0 – 10.255.255.255 or 192.168.0.0 – 192.168.255.255 ranges, represented in shorthand as subnets 10.0.0.0/8 and 192.168.0.0/16, respectively).

IV. Example Device, Application, and Service Discovery

[093] In order for remote network management platform 320 to administer the devices, applications, and services of managed network 300, remote network management platform 320 may first determine what devices are present in managed network 300, the configurations and operational statuses of these devices, and the applications and services provided by the devices, and well as the relationships between discovered devices, applications, and services. As noted above, each device, application, service, and relationship may be referred to as a configuration item. The process of defining configuration items within managed network 300 is referred to as discovery, and may be facilitated at least in part by proxy servers 312.

[094] For purpose of the embodiments herein, an “application” may refer to one or more processes, threads, programs, client modules, server modules, or any other software that executes on a device or group of devices. A “service” may refer to a high-level capability provided by multiple applications executing on one or more devices working in conjunction with one another. For example, a high-level web service may involve multiple web application server threads executing on one device and accessing information from a database application that executes on another device.

[095] Figure 5A provides a logical depiction of how configuration items can be discovered, as well as how information related to discovered configuration items can be stored. For sake of simplicity, remote network management platform 320, third-party networks 340, and Internet 350 are not shown.

[096] In Figure 5A, CMDB 500 and task list 502 are stored within computational instance 322. Computational instance 322 may transmit discovery commands to proxy servers 312. In response, proxy servers 312 may transmit probes to various devices, applications, and services in managed network 300. These devices, applications, and services may transmit responses to proxy servers 312, and proxy servers 312 may then provide information regarding discovered configuration items to CMDB 500 for storage therein. Configuration items stored in CMDB 500 represent the environment of managed network 300.

[097] Task list 502 represents a list of activities that proxy servers 312 are to perform on behalf of computational instance 322. As discovery takes place, task list 502 is populated. Proxy servers 312 repeatedly query task list 502, obtain the next task therein, and perform this task until task list 502 is empty or another stopping condition has been reached.

[098] To facilitate discovery, proxy servers 312 may be configured with information regarding one or more subnets in managed network 300 that are reachable by way of proxy servers 312. For instance, proxy servers 312 may be given the IP address range 192.168.0/24 as a subnet. Then, computational instance 322 may store this information in CMDB 500 and place tasks in task list 502 for discovery of devices at each of these addresses.

[099] Figure 5A also depicts devices, applications, and services in managed network 300 as configuration items 504, 506, 508, 510, and 512. As noted above, these configuration items represent a set of physical and/or virtual devices (e.g., client devices, server devices, routers, or virtual machines), applications executing thereon (e.g., web servers, email servers, databases, or storage arrays), relationships therebetween, as well as services that involve multiple individual configuration items.

[100] Placing the tasks in task list 502 may trigger or otherwise cause proxy servers 312 to begin discovery. Alternatively or additionally, discovery may be manually triggered or

automatically triggered based on triggering events (e.g., discovery may automatically begin once per day at a particular time).

[101] In general, discovery may proceed in four logical phases: scanning, classification, identification, and exploration. Each phase of discovery involves various types of probe messages being transmitted by proxy servers 312 to one or more devices in managed network 300. The responses to these probes may be received and processed by proxy servers 312, and representations thereof may be transmitted to CMDB 500. Thus, each phase can result in more configuration items being discovered and stored in CMDB 500.

[102] In the scanning phase, proxy servers 312 may probe each IP address in the specified range of IP addresses for open Transmission Control Protocol (TCP) and/or User Datagram Protocol (UDP) ports to determine the general type of device. The presence of such open ports at an IP address may indicate that a particular application is operating on the device that is assigned the IP address, which in turn may identify the operating system used by the device. For example, if TCP port 135 is open, then the device is likely executing a WINDOWS® operating system. Similarly, if TCP port 22 is open, then the device is likely executing a UNIX® operating system, such as LINUX®. If UDP port 161 is open, then the device may be able to be further identified through the Simple Network Management Protocol (SNMP). Other possibilities exist. Once the presence of a device at a particular IP address and its open ports have been discovered, these configuration items are saved in CMDB 500.

[103] In the classification phase, proxy servers 312 may further probe each discovered device to determine the version of its operating system. The probes used for a particular device are based on information gathered about the devices during the scanning phase. For example, if a device is found with TCP port 22 open, a set of UNIX®-specific probes may be used. Likewise, if a device is found with TCP port 135 open, a set of WINDOWS®-specific probes

may be used. For either case, an appropriate set of tasks may be placed in task list 502 for proxy servers 312 to carry out. These tasks may result in proxy servers 312 logging on, or otherwise accessing information from the particular device. For instance, if TCP port 22 is open, proxy servers 312 may be instructed to initiate a Secure Shell (SSH) connection to the particular device and obtain information about the operating system thereon from particular locations in the file system. Based on this information, the operating system may be determined. As an example, a UNIX® device with TCP port 22 open may be classified as AIX®, HPUX, LINUX®, MACOS®, or SOLARIS®. This classification information may be stored as one or more configuration items in CMDB 500.

[104] In the identification phase, proxy servers 312 may determine specific details about a classified device. The probes used during this phase may be based on information gathered about the particular devices during the classification phase. For example, if a device was classified as LINUX®, a set of LINUX®-specific probes may be used. Likewise, if a device was classified as WINDOWS® 2012, as a set of WINDOWS®-2012-specific probes may be used. As was the case for the classification phase, an appropriate set of tasks may be placed in task list 502 for proxy servers 312 to carry out. These tasks may result in proxy servers 312 reading information from the particular device, such as basic input / output system (BIOS) information, serial numbers, network interface information, media access control address(es) assigned to these network interface(s), IP address(es) used by the particular device and so on. This identification information may be stored as one or more configuration items in CMDB 500.

[105] In the exploration phase, proxy servers 312 may determine further details about the operational state of a classified device. The probes used during this phase may be based on information gathered about the particular devices during the classification phase and/or the identification phase. Again, an appropriate set of tasks may be placed in task list 502 for proxy

servers 312 to carry out. These tasks may result in proxy servers 312 reading additional information from the particular device, such as processor information, memory information, lists of running processes (applications), and so on. Once more, the discovered information may be stored as one or more configuration items in CMDB 500.

[106] Running discovery on a network device, such as a router, may utilize SNMP. Instead of or in addition to determining a list of running processes or other application-related information, discovery may determine additional subnets known to the router and the operational state of the router's network interfaces (e.g., active, inactive, queue length, number of packets dropped, etc.). The IP addresses of the additional subnets may be candidates for further discovery procedures. Thus, discovery may progress iteratively or recursively.

[107] Once discovery completes, a snapshot representation of each discovered device, application, and service is available in CMDB 500. For example, after discovery, operating system version, hardware configuration and network configuration details for client devices, server devices, and routers in managed network 300, as well as applications executing thereon, may be stored. This collected information may be presented to a user in various ways to allow the user to view the hardware composition and operational status of devices, as well as the characteristics of services that span multiple devices and applications.

[108] Furthermore, CMDB 500 may include entries regarding dependencies and relationships between configuration items. More specifically, an application that is executing on a particular server device, as well as the services that rely on this application, may be represented as such in CMDB 500. For instance, suppose that a database application is executing on a server device, and that this database application is used by a new employee onboarding service as well as a payroll service. Thus, if the server device is taken out of operation for maintenance, it is clear that the employee onboarding service and payroll service will be impacted. Likewise, the

dependencies and relationships between configuration items may be able to represent the services impacted when a particular router fails.

[109] In general, dependencies and relationships between configuration items may be displayed on a web-based interface and represented in a hierarchical fashion. Thus, adding, changing, or removing such dependencies and relationships may be accomplished by way of this interface.

[110] Furthermore, users from managed network 300 may develop workflows that allow certain coordinated activities to take place across multiple discovered devices. For instance, an IT workflow might allow the user to change the common administrator password to all discovered LINUX® devices in a single operation.

[111] In order for discovery to take place in the manner described above, proxy servers 312, CMDB 500, and/or one or more credential stores may be configured with credentials for one or more of the devices to be discovered. Credentials may include any type of information needed in order to access the devices. These may include userid / password pairs, certificates, and so on. In some embodiments, these credentials may be stored in encrypted fields of CMDB 500. Proxy servers 312 may contain the decryption key for the credentials so that proxy servers 312 can use these credentials to log on to or otherwise access devices being discovered.

[112] The discovery process is depicted as a flow chart in Figure 5B. At block 520, the task list in the computational instance is populated, for instance, with a range of IP addresses. At block 522, the scanning phase takes place. Thus, the proxy servers probe the IP addresses for devices using these IP addresses, and attempt to determine the operating systems that are executing on these devices. At block 524, the classification phase takes place. The proxy servers attempt to determine the operating system version of the discovered devices. At block 526, the identification phase takes place. The proxy servers attempt to determine the hardware and/or

software configuration of the discovered devices. At block 528, the exploration phase takes place. The proxy servers attempt to determine the operational state and applications executing on the discovered devices. At block 530, further editing of the configuration items representing the discovered devices and applications may take place. This editing may be automated and/or manual in nature.

[113] The blocks represented in Figure 5B are for purpose of example. Discovery may be a highly configurable procedure that can have more or fewer phases, and the operations of each phase may vary. In some cases, one or more phases may be customized, or may otherwise deviate from the exemplary descriptions above.

V. Natural Language Processing of Text Queries

[114] Natural language processing is a discipline that involves, among other activities, using computers to understand the structure and meaning of human language. This determined structure and meaning may be applicable to the processing of IT incidents, as described below.

[115] Each incident may be represented as an incident report. While incident reports may exist in various formats and contain various types of information, an example incident report 600 is shown in Figure 6. Incident report 600 consists of a number of fields in the left column, at least some of which are associated with values in the right column.

[116] Field 602 identifies the originator of the incident, in this case Bob Smith. Field 604 identifies the time at which the incident was created, in this case 9:56AM on February 7, 2018. Field 605 is a text string that provides a short description of the problem. Field 606 identifies the description of the problem, as provided by the originator. Thus, field 606 may be a free-form text string containing anywhere from a few words to several sentences or more. Field 608 is a categorization of the incident, in this case email. This categorization may be provided

by the originator, the IT personnel to whom the incident is assigned, or automatically based on the context of the problem description field.

[117] Field 610 identifies the IT personnel to whom the incident is assigned (if applicable), in this case Alice Jones. Field 612 identifies the status of the incident. The status may be one of “open,” “assigned,” “working,” or “resolved” for instance. Field 614 identifies how the incident was resolved (if applicable). This field may be filled out by the IT personnel to whom the incident is assigned or another individual. Field 616 identifies the time at which the incident was resolved, in this case 10:10AM on February 7, 2018. Field 618 specifies the closure code of the incident (if applicable) and can take on values such as “closed (permanently)”, “closed (work around)”, “closed (cannot reproduce)”, etc. Field 620 identifies any additional notes added to the record, such as by the IT personnel to whom the incident is assigned. Field 622 identifies a link to an online article that may help users avoid having to address a similar issue in the future.

[118] Incident report 600 is presented for purpose of example. Other types of incident reports may be used, and these reports may contain more, fewer, and/or different fields.

[119] Incident reports, such as incident report 600, may be created in various ways. For instance, by way of a web form, an email sent to a designated address, a voicemail box using speech-to-text conversion, and so on. These incident reports may be stored in an incident report database that can be queried. As an example, a query in the form of a text string could return one or more incident reports that contain the words in the text string. Additionally or alternatively, one or more elements of an incident report (e.g., a “short description” field) may be used to query a database of knowledgebase articles, other incident reports, or some other corpus of text. This may be done in order to identify other incident reports, resolved past incident reports, reports on the resolution of past problems, knowledgebase articles, or other information that may

be relevant to the incident report in order to facilitate resolution of a problem represented in the incident report.

[120] This process is illustrated in Figure 7. A text query may be entered into web interface 700. This web interface may be supplied by way of a computational instance of remote network management platform 320. Web interface 700 converts the text query into a database query (e.g., an SQL query), and provides the SQL query to database 702. This database may be CMDB 500 or some other database. Database 702 contains a number of incident reports with problem description fields as shown in Figure 6. Regardless, database 702 conducts the query and returns matching results to web interface 700. One or more such results may be returned. Web interface 700 provides these results as a web page.

[121] For example, if the text query is “email”, web interface 700 may convert this query into an SQL query of database 702. For example, the query may look at the problem description field of a table containing incident reports. Any such incident report that matches the query – i.e., includes the term “email” – may be provided in the query results. Thus, the incident reports with the problem descriptions of “My email client is not downloading new emails”, “Email crashed”, and “Can’t connect to email” may be provided, while the incident report with the problem description “VPN timed out” is not returned.

[122] This matching technique is simplistic and has a number of drawbacks. It only considers the presence of the text of the query in the incidents. Thus, it does not consider contextual information, such as words appearing before and after the query text. Also, synonyms of the query text (e.g., “mail” or “message”) and misspellings of the query text (e.g., “emial”) would not return any results in this example.

[123] Furthermore, deploying such a solution would involve use of an inefficient sparse matrix, with entries in one dimension for each word in the English language and entries in the

other dimension for the problem description of each incident. While the exact number of English words is a matter of debate, there are at least 150,000 – 200,000, with less than about 20,000 in common use. Given that a busy IT department can have a database of tens of thousands of incidents, this matrix would be quite large and wasteful to store even if just the 20,000 most commonly used words are included.

[124] Thus, the above methods of comparison may be replaced by and/or augmented with a variety of methods that compare the semantic content and/or context of text samples. These methods can improve a variety of machine learning techniques to facilitate natural language processing. Such techniques can include determining word and/or paragraph vectors from samples of text, applying artificial neural networks or other deep learning algorithms, sentiment analysis, or other techniques in order to determine a similarity between samples of text. For example, these or other natural language processing techniques can be applied to determine the similarity between one or more text fields of an incident report and other incident reports, resolved incident reports, knowledgebase articles, or other potentially relevant samples of text.

VI. Natural Language Processing of Text Queries based on Semantic Content

[125] The degree of similarity between two samples of text can be determined in a variety of ways. The two samples of text could be a text field of an incident report and a text field of another incident report, a text field of a resolved incident report, a knowledgebase article, or some other sample of text that may be relevant to the resolution, classification, or other aspects of an incident report. Additionally or alternatively, one or both of the samples could be segments of text within a larger sample of text. As noted above, a degree of overlap between the identities of words present in the two samples of text and/or a word matrix method could be used to determine the degree of similarity. Additionally or alternatively, one or more techniques of

natural language processing could be applied to compare the samples of text such that the context or other semantic content of the texts affects the determined similarity value between the samples of text.

[126] Such techniques may be applied to improve text query matching related to incident reports. These techniques may include a variety of machine learning algorithms that can be trained based on samples of text. The samples of text used for training can include past examples of incident reports, knowledgebase articles, or other text samples of the same nature as the text samples to which the trained model will be applied. This has the benefit of providing a model that has been uniquely adapted to the vocabulary, topics, and idiomatic word use common in its intended application.

[127] Such techniques can include determining word and/or paragraph vectors from samples of text, applying ANNs or other deep learning algorithms, performing sentiment analysis, or other techniques in order to determine a similarity between samples of text, to group multiple samples of text together according to topic or content, to partition a sample of text into discrete internally-related segments, to determine statistical associations between words, or to perform some other language processing task. Below, a particular method for determining similarity values between samples of text using an ANN model that provides compact semantic representations of words and text strings is provided as a non-limiting example of such techniques. However, other techniques may be applied to generate similarity values between samples of text as applied elsewhere herein. In the discussion below, there are two approaches for training an ANN model to represent the semantic meanings of words: word vectors and paragraph vectors. These techniques may be combined with one another or with other techniques.

A. Word Vectors

[128] A “word vector” may be determined for each word present in a corpus of text records such that words having similar meanings (or “semantic content”) are associated with word vectors that are near each other within a semantically encoded vector space. Such vectors may have dozens, hundreds, or more elements. These word vectors allow the underlying meaning of words to be compared or otherwise operated on by a computing device. Accordingly, the use of word vectors may allow for a significant improvement over simpler word list or word matrix methods.

[129] Word vectors can be used to quickly and efficiently compare the overall semantic content of samples of text, allowing a similarity value between the samples of text to be determined. This can include determining a distance, a cosine similarity, or some other measure of similarity between the word vectors of the words in each of the text samples. For example, a mean of the word vectors in each of the text samples could be determined and a cosine similarity between the means then used as a measure of similarity between the text samples. Additionally or alternatively, the word vectors may be provided as input to an ANN, a support vector machine, a decision tree, or some other machine learning algorithm in order to perform sentiment analysis, to classify or cluster samples of text, to determine a level of similarity between samples of text, or to perform some other language processing task.

[130] Word vectors may be determined for a set of words in a variety of ways. In an example, a matrix of the word vectors can be an input layer of an ANN. The ANN (including the matrix of word vectors) can then be trained with a large number of text strings from a database to determine the contextual relationships between words appearing in these text strings. Such an ANN 800 is shown in Figure 8A. ANN 800 includes input layer 802, which feeds into hidden layer 804, which in turn feeds into output layer 806. The number of nodes in input layer 802 and

output layer 806 may be equivalent to the number of words in a pre-defined vocabulary or dictionary (e.g., 20,000, 50,000, or 100,000). The number of nodes in hidden layer 804 may be much smaller (e.g., 64 as shown in Figure 8A, or other values such as 16, 32, 128, 512, 1024, etc.).

[131] For each text string in the database, ANN 800 is trained with one or more arrangements of words. For instance, in Figure 8B, ANN 800 is shown being trained with input word “email” and output (context) words “can’t”, “connect” and “to”. The output words serve as the ground truth output values to which the results produced by output layer 806 are compared. This arrangement reflects that “email” appears proximate to “can’t”, “connect” and “to” in a text string in database 702.

[132] In an implementation, this could be represented as node I_2 receiving an input of 1, and all other nodes in input layer 802 receiving an input of 0. Similarly, node O_1 is associated with a ground truth value of “can’t”, node O_2 is associated with a ground truth value of “connect”, and node O_3 is associated with a ground truth value of “to”. In the implementation, this could be represented as nodes O_1 , O_2 , and O_3 being associated with ground truth values of 1 and all other nodes in output layer 806 being associated with ground truth values of 0. The loss function may be a sum of squared errors, for example, between the outputs generated by output layer 806 in response to the input described above and a vector containing the ground truth values associated with the output layer nodes.

[133] Other arrangements of this text string from database 702 may be used to train an ANN. For instance, as shown in Figure A the input word to input layer 902 of an ANN 900 may be “can’t” and the output words of output layer 906 may be “connect”, “to”, and “email.” In another example, as shown in Figure 9B, the input word may be “connect” and the output words may be “can’t”, “to”, and “email.”

[134] In general, these arrangements may be selected so that the output words are within w words of the input word (e.g., where w could be 1, 2, 3, 5, etc.), the output words are in the same sentence as the input word, the output words are in the same paragraph as the input word, and so on. Furthermore, various word arrangements of each text string in database 702 may be used to train ANN (e.g., ANN 800 or ANN 900). These text strings may be selected from short description field 605, problem description field 606, category field 608, resolution field 614, notes field 620, and/or any other field or combination of fields in an incident report.

[135] After an ANN (e.g., 800, 900) is trained with these arrangements of text strings, the hidden layer (e.g., 804, 904) of the ANN becomes a compact vector representation of the context and meaning of an input word. That is, the weightings from a particular node (e.g., I_3) in the input layer to the hidden layer represent the elements of the word vector of the word corresponding to the particular node (e.g., “can’t”). For example, assuming that ANN is fully-trained with a corpus of 10,000 or so text strings (though more or fewer text strings may be used), an input word of “email” may have a similar vector representation of an input word of “mail”. Intuitively, since hidden layer is all that ANN has to determine the context of an input word, if two words have similar contexts, then they are highly likely to have similar vector representations.

[136] In some embodiments, an ANN can be trained with input words associated with the output nodes $O_1 \dots O_n$ and the output (context) words associated with input nodes $I_1 \dots I_n$. This arrangement may produce an identical or similar vector for hidden layer.

[137] Furthermore, vectors generated in this fashion are additive. Thus, subtracting the vector representation of “mail” from the vector representation of “email” is expected to produce a vector with values close to 0. However, subtracting the vector representation of “VPN” from

the vector representation of “email” is expected to produce a vector with higher values. In this manner, the model indicates that “email” and “mail” have closer meanings than “email” and “VPN”.

[138] Once vector representations have been determined for all words of interest, linear and/or multiplicative aggregations of these vectors may be used to represent text strings. For instance, a vector for the text string “can’t connect to email” can be found by adding together the individual vectors for the words “can’t”, “connect”, “to”, and “email”. In some cases, an average or some other operation may be applied to the vectors for the words. This can be expressed below as the vector sum of m vectors $\mathbf{v}_i$ with each entry therein divided by m , where $i = \{1 \dots m\}$. But other possibilities, such as weighted averages, exist.

$$\mathbf{v}_{avg} = \frac{1}{m} \sum_{i=1}^m \mathbf{v}_i \quad (1)$$

[139] Regardless of how the aggregations are determined, this general technique allows vector representations for each text string in database 702 to be found. These vector representations may be stored in database 702 as well, either along with their associated text strings or separately. These vector representations can then be used to compare the text strings, cluster or group the text strings, train some other machine learning classifier, or to perform some other task. For example, a matching text string for a particularly query text may be determined by determining a cosine similarity or other similarity value between the vector representation of the query text and the stored vector representations of samples of text in the database 702.

[140] The comparison may identify one or more text string vectors from database 702 that “match” in this fashion. In some cases this may be the k text string vectors with the highest similarity, or any text string vector with a similarity that is greater than a pre-determined value. The identified text string vectors could correspond to a subset of incident reports, within a

greater corpus of incident reports that is recorded in the database 702, that are relevant to an additional incident report that corresponds to the query text string vector. For each of the identified text string vectors, the associated text string may be looked up in database 702 and provided as an output text string. In some cases, the associated incident reports may be provided as well.

[141] In some cases, only incident reports that are not older than a pre-determined age are provided. For instance, the system may be configured to identify text string vectors only from incident reports that were resolved within the last 3 months, 6 months, or 12 months. Alternatively, the system may be configured to identify text string vectors only from incident reports that were opened within the last 3 months, 6 months, or 12 months.

[142] In this fashion, incident reports with similar problem descriptions as that of the input text string can be rapidly identified. Notably, this system provides contextual results that are more likely to be relevant and meaningful to the input text string. Consequently, an individual can review these incident reports to determine how similar problems as that in the problem description have been reported and addressed in the past. This may result in the amount of time it takes to resolve incidents being dramatically reduced.

[143] Additionally or alternatively, these embodiments can be applied to detect and identify clusters of semantically and/or contextually similar incident reports within a corpus of incident reports. For example, clusters of incident reports related to a similar issue that is likely to affect users of an IT system, an ongoing misconfiguration of one or more aspects of an IT system, a progressive hardware failure in a component of an IT system, or some other recurring issue within an IT system. Identifying such clusters of related incident reports can allow the IT system to be repaired or upgraded (e.g., by replacing and/or reconfiguring failing or

inconsistently performing hardware or software), users to be trained to avoid common mistakes, rarely-occurring hardware or software issues to be detected and rectified, or other benefits.

[144] Such clusters of relevant incident reports can be detected and/or identified by identifying, within the semantically encoded vector space, aggregated word (and/or paragraph) vectors corresponding to the incident reports. A variety of methods could be employed to detect such clusters within the semantically encoded vector space, e.g., k-means clustering, support vector machines, ANNs (e.g., unsupervised ANNs configured and/or trained to identify relevant subsets of training examples within a corpus of available training examples), or some other classifier or other method for identifying clusters of related vectors within a vector space.

B. Paragraph Vectors

[145] As discussed previously, an ANN model (e.g., 800, 900) uses the surrounding context to provide compact, semantically relevant vector representations of words. After training, words with similar meanings can map to a similar position in the vector space. For example, the vectors for “powerful” and “strong” may appear close to each other, whereas the vectors for “powerful” and “Paris” may be farther apart. Additions and subtractions between word vectors also carry meaning. Using vector algebra on the determined word vectors, we can answer analogy questions such as “King” - “man” + “woman” = “Queen.”

[146] However, the complete semantic meaning of a sentence or other passage (e.g., a phrase, several sentences, a paragraph, a text segment within a larger sample of text, or a document) cannot always be captured from the individual word vectors of a sentence (e.g., by applying vector algebra). Word vectors can represent the semantic content of individual words and may be trained using short context windows. Thus, the semantic content of word order and any information outside the short context window is lost when operating based only on word vectors.

[147] Take for example the sentence “I want a big green cell right now.” In this case, simple vector algebra of the individual words may fail to provide the correct semantic meaning of the word “cell,” as the word “cell” has multiple possible meanings and thus can be ambiguous. Depending on the context, “cell” could be a biological cell, a prison cell, or a cell of a cellular communications network. Accordingly, the paragraph, sentence, or phrase from which a given word is sampled can provide crucial contextual information.

[148] In another example, given the sentence “Where art thou __,” it is easy to predict the missing word as “Romeo” if sentence was said to derive from a paragraph about Shakespeare. Thus, learning a semantic vector representation of an entire paragraph can help contribute to predicting the context of words sampled from that paragraph.

[149] Similar to the methods above for learning word vectors, an ANN or other machine learning structures may be trained using a large number of paragraphs in a corpus to determine the contextual meaning of entire paragraphs, sentences, phrases, or other multi-word text samples as well as to determine the meaning of the individual words that make up the paragraphs in the corpus. Such an ANN 1000 is shown in Figure 10A. ANN 1000 includes input layer 1002, which feeds into hidden layer 1004, which in turn feeds into output layer 1006. Note that input layer 1002 consists of two types of input substructures, the top substructure 1008 (consisting of input nodes $I_1 \dots I_n$) representing words and the bottom substructure 1010 (consisting of input nodes $D_1 \dots D_m$) representing paragraphs (documents). The number of nodes in output layer 1006 and the top input layer substructure 1008 may be equal to the number of unique words in the entire corpus. The number of nodes in the bottom input layer substructure 1010 may be equivalent to the number of unique paragraphs in the entire corpus. Note that “paragraph,” as used herein, may be a sentence, a paragraph, one or more fields of an incident report, a segment of a larger string of text, or some other multi-word string of text.

[150] For each paragraph in the corpus, ANN 1000 is trained with fixed-length contexts generated from moving a sliding window over the paragraph. Thus, a given paragraph vector is shared across all training contexts created from its source paragraph, but not across training contexts created from other paragraphs. Word vectors are shared across training contexts created from all paragraphs, e.g., the vector for “cannot” is the same for all paragraphs. Paragraphs are not limited in size; they can be as large as entire documents or as small as a sentence or phrase. In Figure 10A, ANN 1000 is shown in a single training iteration, being trained with input word context “can’t,” “connect” and “to,” input paragraph context DOC 1, and output word “email.” The output word serves as the ground truth output value to which the result produced by output layer 1006 is compared. This arrangement reflects that “email” appears proximate to “can’t,” “connect,” and “to,” and is within DOC 1.

[151] In an implementation, this could be represented as output node O_4 receiving a ground truth value of 1 and all other nodes in output layer 1006 having ground truth values of 0. Similarly, node I_1 has a ground truth value of “can’t,” node I_2 has a ground truth value of “connect,” node I_3 has a ground truth value of “to,” and node D_1 has ground truth value of DOC 1. In the implementation, this could be represented as nodes I_1 , I_2 , I_3 , and D_1 being associated with values of 1 and all other nodes in input layer 1002 having values of 0. The loss function may be a sum of squared errors, for example, between the output of output layer 1006 and a vector containing the ground truth values. The weight values of the corresponding word vectors and paragraph vectors, as well all the output layer parameters (e.g., softmax weights) are updated based on the loss function (e.g., via backpropagation).

[152] Figure 10B shows ANN 1000 being trained with a subsequent context window. This context window derives from the same document, but shifts ahead a word in the document and uses input word context “connect,” “to” and “email,” input paragraph context DOC 1, and

output word “server.” In an implementation, these inputs and outputs can be encoded with ground truth values as similarly described above.

[153] Figure 10C shows an instance of ANN 1000 trained with another document within the corpus. The context window derives from this document and uses input word context “can’t”, “load”, and “my”, input paragraph context DOC 2, and output word “database.” In an implementation, these inputs and outputs can be encoded with ground truth values as similarly described above.

[154] After ANN 1000 is trained, the weights associated with hidden layer 1004 become a compact vector representation of the context and meaning of input words and paragraphs. For example, assuming that ANN 1000 is fully-trained with a corpus of 1,000 paragraphs, with the entire corpus containing 10,000 unique words, each paragraph and each word can be represented by a unique vector with a length equal to the number of hidden nodes in hidden layer 1004. These unique vectors encode the contextual meaning of words within the paragraphs or the paragraphs themselves.

[155] Figure 10D shows ANN 1000 at prediction time performing an inference step to compute the paragraph vector for a new, previously unseen paragraph. This inference step begins by adding an additional input node 1012 to input layer substructure 1010 that represents the unseen paragraph (DOC M+1). During this inference process, the coefficients of the word vectors substructure 1008 and the learned weights between hidden layer 1004 and output layer 1006 are held fixed. Thus, the model generates an additional paragraph vector 1012, corresponding to the unseen paragraph in the input paragraph vector substructure 1010, to obtain the new semantic vector representation of the unseen paragraph. Any additional unseen paragraphs can be trained through a similar process by adding input nodes to input layer substructure 1010.

[156] Alternatively, paragraph vectors can be trained by ignoring word context in the input layer, only using the paragraph vector as the input, and forcing the model to predict different word contexts randomly sampled from the paragraph in the output layer. The input layer of such an ANN only consists of paragraph vectors, while the output layer represents a single context window that is randomly generated from a given paragraph. Training such an ANN may result in a vector representation for the semantic content of paragraphs in the corpus, but will not necessarily provide any semantic vector representations for the words therein.

[157] Once vector representations have been determined for paragraphs in the corpus, linear and/or multiplicative aggregation of these vectors may be used to represent topics of interest. Furthermore, if the dimensions of paragraph vectors are the same as the dimensions of word vectors, as shown in ANN 1000, then linear and multiplicative aggregation between word vectors and paragraphs vectors can be obtained. For example, finding the Chinese equivalent of “Julius Caesar” using an encyclopedia as a corpus can be achieved by vector operations $PV(\text{“Julius Caesar”}) - WV(\text{“Roman”}) + WV(\text{“Chinese”})$, where PV is a paragraph vector (representing an entire Wikipedia article) and WV are word vectors. Thus, paragraph vectors can achieve the same kind of analogies to word vectors with more context-based results.

[158] In practice, such learned paragraph vectors can be used as inputs into other supervised learning models, such as sentiment prediction models. In such models, which can include but are not limited to ANNs, Support Vector Machines (SVMs), or Naïve Bayes Classifiers, paragraph vectors are used as input with a corresponding sentiment label as output. Other metrics such as cosine similarity and nearest neighbors clustering algorithms can be applied to paragraph vectors to find or group paragraphs on similar topics within the corpus of paragraphs.

[159] In the present embodiments, a combination of learned word vectors and paragraph vectors can help determine the structure and meaning of incidents reports, for example incident report 600 as shown in Figure 6. Incident report 600 consists of a number of fields in the left column, at least some of which are associated with values in the right column. For longer text fields, such as short description field 605, problem description field 606, resolution field 614, and notes field 620, it may be preferable to represent the associated right column text as a paragraph vector, or as multiple paragraph vectors corresponding to respective text segments within the right column text, to gain more contextual meaning rather than aggregating the individual word vectors that form the text. Incident report 600 is presented for purpose of example. Various fields of an incident report can be arranged to be represented as paragraph vectors, word vectors, or weighted combinations of the two. Other types of incident reports, problem reports, case files, or knowledgebase articles may also be used, and these reports may contain more, fewer, and/or different fields.

[160] After representing different fields as paragraph vectors, word vectors, or weighted combinations of the two, a single vector to represent the entire incident can be generated by concatenating, generating a vector sum, or otherwise aggregating the word and/or paragraph vector representations of the individual incident fields. With a single aggregate incident vector representation, a system can be configured to identify similar aggregate vectors (and therefore similar incident reports) based on cosine similarity or other metrics as discussed above. Alternatively, a search for similar incident reports may use just the paragraph text of one or more individual fields. In this fashion, text from one or more individual fields in an incident report could be combined into a single paragraph of text. A paragraph vector could then be generated from this single, large paragraph of concatenated text and used to search for similar incidents.

[161] This process can be illustrated in terms of the previously described ANN structures. Initially, text strings are obtained from database 702 of Figure 7. As noted above, these text strings may be from parts of incident reports. Then, words are extracted from the text strings. The words extracted may be all of the words in the text strings or some of these words. These extracted words are provided as input to ANN 900 of Figures 9A-9B. The substring contexts of these words are extracted from the text strings. The substring contexts may be one or more substrings containing words before, after, or surrounding the associated words that were extracted. These vector representations may then be used to compare (e.g., using cosine similarity) their respective text samples.

[162] The comparison may identify one or more incident reports from database 702 that “match” in this fashion. In some cases this may be the k incident reports with the highest similarity, or any incident report with a similarity that is greater than a pre-determined value. The user may be provided with these identified incident reports or references thereto.

[163] In some cases, only incident reports that are not older than a pre-determined age are provided. For instance, the system may be configured to only identify incident reports that were resolved within the last 3 months, 6 months, or 12 months. Alternatively, the system may be configured to only identify incident reports that were opened within the last 3 months, 6 months, or 12 months.

[164] In this fashion, incident reports with similar content as that of the input incident report can be rapidly identified. Consequently, an individual can review these incident reports to determine how similar problems as that in the incident have been reported and addressed in the past. This may result in the amount of time it takes to resolve incidents being dramatically reduced.

[165] While this section describes some possible embodiments of word vectors and paragraph vectors, other embodiments may exist. For example, different ANN structures and different training procedures can be used.

VII. Example Distribution of Functionality Across a Remote Network Management Platform

[166] A remote network management platform as described herein may be configured to provide a variety of services to one or more managed networks associated with one or more users. These services can include serving web pages over the Internet, managing servers or other systems of the managed network(s), providing an interface for users or technicians to input incident reports or queries related to the operation of the managed network(s), maintaining databases of such incidents and/or queries, knowledgebase articles, system logs, or other information about the operation of the managed network(s), providing responses to received queries, managing resolved and outstanding incident reports, allocating technicians or other assets to outstanding incident reports, detecting patterns within sets of received incident reports, or other services related to the ongoing operation of end-user (managed) networks.

[167] Some of the services provided by the remote network management platform may include, or may be enhanced by, generating and applying one or more ML models to determine a numerical similarity between textual records. For example, the remote network management platform could determine respective degrees of numerical similarity between one or more fields of a newly-generated incident report and fields of stored incident reports and/or knowledgebase articles in order to identify, based on the numerical similarities, stored incident reports and/or knowledgebase articles that may be relevant to a user or technician attempting to resolve the newly-generated incident report. In another example, the remote network management platform could determine respective degrees of numerical similarity between one or more fields of stored

incident reports in order to identify related clusters of incident reports. These identified clusters could then be used to identify current or historical issues within a managed network, to develop a diagnostic solution for a common technological problem associated with an identified cluster or incident reports, to assign newly-generated incident reports to an existing cluster to facilitate resolution of the incident report, or to provide some other benefit. Such ML models could include natural language processing or other techniques described herein.

[168] The generation and application of such ML models can be computationally expensive. For example, the training of an ML model, based on a corpus of incident reports or other textual records, could require significant RAM or other memory, numerous processor cycles, or other computational resources. If a server used to serve web pages, manage the generation and storage of incident reports, manage servers of a managed network, or to perform other services related to a managed network is also used to train and/or apply such ML models, the server's performance of the former tasks may be significantly degraded while training and/or apply the ML models. For example, a latency of processing web page requests, receiving new incident reports, determining and transmitting commands to managed servers, or other operations may increase and/or the remote network management platform may fail to perform some of these tasks. In another example, the server may experience an out-of-memory error or other system errors.

[169] To reduce or avoid such issues, the remote network management platform could allocate separate sets of computational resources (e.g., separate servers, sets of one or more servers, processor cores or other portions of servers) to perform different operations. For example, a first set of computational resources of the remote network management platform could be allocated to an end-user computational instance that is configured to serve web pages, manage the generation and storage of incident reports, maintain and serve a set of

knowledgebase articles, manage servers of a managed network, or to perform some other operations related to the maintenance and operation of a managed network. Additional sets of computational resources could be allocated to generate and/or apply ML models based on incident reports, knowledgebase articles, or other textual records related to the end-user computational instance and/or to the managed network.

[170] For example, a second set of computational resources of the remote network management platform could be allocated to a training computational instance that is configured to generate, for the end-user computational instance, ML models based on a corpus of textual records (e.g., incident reports, knowledgebase articles) related to the end-user computational instance and/or to the managed network. The end-user instance could transmit a request for the ML model, along with the corpus of textual records, to the training instance periodically in order to update the ML model over time as new incident reports or other textual records are received / generated. Additionally or alternatively, the training instance could acquire the corpus of textual records from a database or other source outside of the end-user instance.

[171] In such an example, a third set of computational resources of the remote network management platform could be allocated to a prediction computational instance that is configured to use the ML model(s) generated by the training computational instance to determine, for the end-user computational instance, numerical similarities between incident reports or other textual records specified by the end-user instance and textual records within the corpus of textual records. Such numerical similarities could be used to identify other incident reports, knowledgebase articles, or other resources that may be relevant to the textual record specified by the end-user instance. The prediction instance can then, based on the determined numerical similarities, transmit to the end-user instance a representation of one or more textual records within the corpus of textual records. For example, the prediction instance could transmit

a representation of the textual record within the corpus of textual records that is most similar to the textual record specified by the end-user instance. In another example, the prediction instance could transmit a representation of a cluster of textual records within the corpus that are, as a group, most similar to the textual record specified by the end-user instance. The end-user instance could provide the ML model to the prediction instance. Additionally or alternatively, the prediction instance could receive the ML model directly from the training instance, from a database that contains the ML model, from another predication instance, or from some other source.

[172] Partitioning the functionality of a remote network management system across different computational instances in this manner can provide a variety of benefits. A particular end-user instance can gain access to ML model predictions (by sending a request to a training instance and/or prediction instance) without significantly degrading its performance of ongoing network management tasks or other ongoing processes. The balance of computational resources provided to each computational instance can be tailored to the intended use of the computational instance, decreasing cost and/or increasing resource utilization. For example, computational resources allocated to an end-user instance could include less memory, more non-volatile storage, and more bandwidth for communication across the internet. Conversely, computational resources allocated to a training instance and/or a prediction instance could include more memory. Additionally, the overall computational resources of the remote network management system can be allocated amongst multiple end-user instances, training instances, and prediction instances according to ongoing usage statistics. For example, if end-user instances are generating fewer requests for new ML models and/or fewer requests for predictions generated using such ML models, computational resources used for training and prediction instances could be re-allocated to end-user instances.

[173] Figure 11 depicts a remote network management architecture, in accordance with example embodiments. This architecture includes two main components, managed networks 1100 and 1105 associated with respective different groups of end-users and remote network management platform 1120, all connected by way of Internet 1150.

[174] Managed network 1100 and/or 1105 may each be, for example, an enterprise network used by an entity for computing and communications tasks, as well as storage of data. Thus, managed network 1100 and/or 1105 may each include client devices, server devices, routers, virtual machines, firewall(s), and/or proxy servers. Client devices may be embodied by computing device 100, server devices may be embodied by computing device 100 or server cluster 200, and routers may be any type of router, switch, or gateway.

[175] In some cases, managed network 1100 and/or 1105 may consist of a few devices and a small number of networks. In other deployments, managed network 1100 and/or 1105 may span multiple physical locations and include hundreds of networks and hundreds of thousands of devices. Additionally, remote network management platform 1120 may be configured to provide services to additional managed networks, for example, hundreds or thousands of managed networks. Thus, the architecture depicted in Figure 11 is capable of scaling up or down by orders of magnitude.

[176] Remote network management platform 1120 is a hosted environment that provides aPaaS services to users, particularly to the operators of managed networks 1100 and 1105. These services may take the form of web-based portals, for instance. Thus, a user can securely access remote network management platform 1120 from, for instance, client devices within the managed network(s) 1100, 1105, or potentially from a client device outside of managed networks. By way of the web-based portals, users may design, test, and deploy applications, generate reports, view analytics, and perform other tasks. Remote network

management platform 1120 may be an embodiment of remote network management platform 320, for example.

[177] As shown in Figure 11, remote network management platform 1120 includes five computational instances 1122, 1124, 1126, 1128, and 1130. Each of these instances may represent one or more server devices and/or one or more databases that provide a set of web portals, services, and applications (e.g., a wholly-functioning aPaaS system) available to a particular customer. Additionally or alternatively, each of these instances may represent one or more server devices that provide training and/or application of ML models based on sets of textual records or other information (e.g., ML models that use NLP to generate word and/or paragraph vectors to represent the semantic content of fields of incident reports and/or other textual records) to a particular customer or to multiple customers. Each computational instance has a respective set of computational resources of the remote network management platform 1120 allocated to it. Such resources may include processors, processor cores, local memory (e.g., RAM), non-volatile memory (e.g., individual hard disks, segments of a logical array of disks, a portion of a shared database), servers, modems, or other computational resources. For example, a first end-user instance 1122 includes a database 1123 that may contain incident reports, knowledgebase articles, and/or other textual records of use to the first end-user instance 1122. In another example, a first prediction instance 1128 includes a memory 1129 (e.g., a RAM, a non-volatile memory) that may contain an ML model that the prediction instance 1128 may apply.

[178] In some cases, a single customer may use multiple end-user computational instances. For example, managed network 1100 may be an enterprise customer of remote network management platform 1120, and may use computational instances 1122 and 1124. The reason for providing multiple instances to one customer is that the customer may wish to

independently develop, test, and deploy its applications and services. Thus, computational instance 1122 may be dedicated to application development related to managed network 1100, computational instance 1124 may be dedicated to testing these applications, and additional computational instance(s) (not shown) may be dedicated to the live operation of tested applications and services. A computational instance may also be referred to as a hosted instance, a remote instance, a customer instance, or by some other designation. Any application deployed onto a computational instance may be a scoped application, in that its access to databases within the computational instance can be restricted to certain elements therein (e.g., one or more particular database tables or particular rows with one or more database tables).

[179] In some cases, a single computational instance may be used by multiple customers. For example, the training instance 1126 and/or prediction instance(s) 1128, 1130 may be used by managed networks 1100, 1105 that correspond to respective different customers. In such an example, first 1122 and second 1124 end-user instances may correspond to first and second customers, respectively, and may operate to manage first 1100 and second 1105 managed networks, respectively. In such an example, both of the end-user instances 1122, 1124 may send requests to the training instance 1126 to generate an ML model and/or to the prediction instance(s) 1128, 1130 to determine numerical similarities using such an ML model.

[180] Internet 1150 may represent a portion of the global Internet. However, Internet 1150 may alternatively represent a different type of network, such as a private wide-area or local-area packet-switched network.

[181] A prediction computational instance as described herein is configured to receive, from an end-user computational instance or from some other requestor instance/system, a request to use an ML model to determine respective degrees of numerical similarity between an additional textual record provided by the end-user instance (e.g., as part of the request) and

textual records in a corpus of textual records used to generate the ML model. The prediction instance then uses the ML model to determine the numerical similarities and transmits, to the end-user instance and based on the determined numerical similarities, a representation of one or more textual records in the corpus of textual records. Transmitting a representation of one or more textual records in the corpus of textual records could include transmitting a representation of the textual record in the corpus that is, with respect to the determined numerical similarities, most similar to the additional textual record. Alternatively, transmitting a representation of one or more textual records in the corpus of textual records could include transmitting a representation of a specified number n of textual records in the corpus that are, with respect to the determined numerical similarities, the most similar to the additional textual record. In yet another example, transmitting a representation of one or more textual records in the corpus of textual records could include transmitting a representation of a cluster or other group of textual records within the corpus that are, as a group and with respect to the determined numerical similarities, the most similar to the additional textual record. Such a cluster could be selected from a set of clusters detected within the corpus of textual records, e.g., using a k-means clustering algorithm or some other clustering method.

[182] Determining a numerical similarity between the additional textual record and a particular textual record in the corpus of textual records could include determining respective vector representations of the textual records. Such a vector representation could be, as described above, a word vector, a paragraph vector, or some other vector representation of the textual record, in a semantically-encoded vector space, that represents the meaning of words and/or of sets of words in the textual records. Determining the numerical similarity between textual records in such embodiments could then include determining a distance (e.g., a Euclidean distance, a Manhattan distance) between the vector representations of the textual records, and

angle (e.g., a cosine angle) between the vector representations of the textual records, or determining some other metric of similarity between the vector representations.

[183] Determining a numerical similarity between the additional textual record and a cluster of textual records within the corpus of textual records could include determining an aggregate vector representation of the textual records of the cluster and then determining a distance, angle, or other metric of similarity between a vector representation of the additional textual record and the aggregate vector representation of the cluster. Such an aggregate vector representation could be an arithmetic, geometric, or other type of average of the vector representations of textual records that are members of the cluster. In another example, the aggregate vector representation could be a centroid or other vector representation used as part of a clustering algorithm used to identify the cluster within the corpus of textual records (e.g., a centroid of a cluster identified using k-means clustering). Additionally or alternatively, similarity between the additional textual record and a particular cluster could be determined by determining whether the vector representation of the additional textual record is located within a representative volume of the particular cluster. For example, such a representative volume could be a hyper-sphere defined by being less than a threshold distance, within the vector space, from a centroid or other representative vector for the particular cluster.

[184] Transmitting a representation, to an end-user computational instance, of one or more textual records in a corpus of textual records can include transmitting a variety of different information. In some examples, transmitting a representation of a textual record includes transmitting an identification number (e.g., a global unique identifier (GUID)), time and date stamp, a location within a database, or some other identifying information that can enable the end-user instance to identify the represented textual record in a database or to otherwise access the represented textual record. Additionally or alternatively, transmitting a representation of a

textual record can include transmitting a copy of the textual record itself or a portion thereof. For example, transmitting a representation of a textual record can include transmitting a copy of a ‘problem resolution’ field of an incident report.

[185] Transmitting a representation of one or more textual records in a corpus of textual records can include transmitting a representation of one or more clusters of textual records within a corpus of textual records. This can include transmitting an identification number (e.g., a GUID), a location within a database, or some other identifying information that can enable the end-user instance to identify the represented cluster of textual records. Additionally or alternatively, a representation of the contents of the cluster of textual records could be transmitted. This could include transmitting identification information for the textual records within the cluster (e.g., GUIDs, locations within a database), all or part of the contents of the textual records within the cluster (e.g., text from ‘problem resolution’ fields of incident reports within the cluster), the contents of one or more representative textual records associated with the cluster (e.g., a knowledgebase article associated with the cluster, a ‘problem resolution’ field or other contents of an incident report that is near a centroid of the cluster), or some other content associated with the cluster.

[186] A training computational instance (e.g., training instance 1126) is configured to generate, for end-user instances, ML models that can then be used by prediction instances as described herein. The prediction instances may access the ML models in a variety of ways. In some examples, they may receive the ML models directly from the generating training instances, or from a database that contains the ML model. Additionally or alternatively, the prediction instances may receive the ML models from the end-use instances (e.g., as part of a request for a prediction). A prediction instance may store the ML model in a memory of the prediction instance (e.g., in memory 1129). Such a prediction instance may maintain the ML model in the

memory for a specified period of time in order to provide multiple predictions to an end-user instance. Such a prediction instance may operate periodically to delete ML models from the memory in order to free up space in the memory, in order to provide increased data security for the end-user instance, or to provide some other benefit. In such embodiments, the prediction instance may, in response to receiving a request for a prediction from an end-user instance, search the memory to determine whether an appropriate ML model is present. If the prediction instance cannot find an ML model suitable to satisfy the end-user instance's request, the prediction instance can send, to the end-user instance, a request for the ML model. The end-user instance can then send the ML model to the prediction instance. Additionally or alternatively, the prediction instance may send a request for the ML model to a training instance, a database, or some other location where the ML model may be stored and/or generated.

VIII. Example Operations

[187] Figure 12 is a flow chart illustrating an example embodiment. The process illustrated by Figure 12 may be carried out by a computing device, such as computing device 100, and/or a cluster of computing devices, such as server cluster 200 or remote network management platforms 320 or 1120. However, the process can be carried out by other types of devices or device subsystems. For example, the process could be carried out by a portable computer, such as a laptop or a tablet device.

[188] The embodiments of Figure 12 may be simplified by the removal of any one or more of the features shown therein. Further, these embodiments may be combined with features, aspects, and/or implementations of any of the previous figures or otherwise described herein.

[189] The example embodiment of Figure 12 includes operating a first set of computational resources of a remote network management platform to provide an end-user computational instance that is dedicated to a managed network (1200).

[190] The example embodiment of Figure 12 additionally includes operating a second set of computational resources of the remote network management platform to provide a training computational instance (1202). The training computational instance is configured to perform operations including: (i) receiving, from the end-user computational instance or from a database associated therewith, a corpus of textual records, (ii) training, based on the corpus of textual records, an ML model to determine a degree of numerical similarity between input textual records and textual records in the corpus of textual records, and (iii) transmitting, to the end-user computational instance, the ML model.

[191] The example embodiment of Figure 12 additionally includes operating a third set of computational resources of the remote network management platform to provide a prediction computational instance (1204). The prediction computational instance is configured to perform operations including: (i) receiving, from the end-user computational instance, an additional textual record, (ii) receiving, from the end-user computational instance, from the training computational instance, or from some other source, the ML model, (iii) determining, by the ML model, respective numerical similarities between the additional textual record and the textual records in the corpus of textual records, and (iv) based on the respective numerical similarities, transmitting, to the end-user computational instance, representations of one or more of the textual records in the corpus of textual records.

[192] In some embodiments, the third set of computational resources includes more memory than the first set of computational resources.

[193] In some embodiments, the third set of computational resources includes a memory, and the operations performed by the prediction computational instance additionally include: in response to receiving the additional textual record, determining that the ML model is

not stored in the memory; and requesting, from the end-user computational instance, the ML model.

[194] In some embodiments, determining, by the ML model, respective numerical similarities between the additional textual record and one or more of the textual records in the corpus of textual records includes using the ML model to determine at least one of (i) word vectors that describe, in a first semantically-encoded vector space, a meaning of respective words of the additional textual record, or (ii) a paragraph vector that describes, in a second semantically-encoded vector space, a meaning of multiple words of the additional textual record.

[195] In some embodiments, the ML model represents a set of clusters of the textual records in the corpus of textual records, and determining, by the ML model, respective numerical similarities between the additional textual record and the textual records in the corpus of textual records further comprises selecting, for the additional textual record, a cluster from the set of clusters based on at least one of the word vectors or paragraph vector.

[196] In some embodiments, training the ML model to determine degree of numerical similarity between input textual records and textual records in the corpus of textual records includes training the ML model to: determine, for an input textual record and for a given textual record in the corpus of textual records, respective vector representations in a vector space; and determine a degree of numerical similarity between the input textual record and the given textual record by determining a distance, within the vector space, between the vector representations of the input textual record and the given textual record.

[197] In some embodiments, training the ML model to determine the degree of numerical similarity between input textual records and textual records in the corpus of textual records includes training the ML model to: determine, for an input textual record and for a given textual record in the corpus of textual records, respective vector representations in a vector

space; and determine a degree of numerical similarity between the input textual record and the given textual record by determining an angle, within the vector space, between the vector representations of the input textual record and the given textual record.

[198] In some embodiments, training the ML model to determine the degree of numerical similarity between input textual records and textual records in the corpus of textual records includes training the ML model to: determine, for a cluster of textual records in the corpus of textual records, an aggregate vector representation in a vector space; determine, for an input textual record, a vector representation in the vector space; and determine a degree of numerical similarity between the input textual record and the cluster of textual records in the corpus of textual records by determining an angle, within the vector space, between the vector representation of the input textual record and the aggregate vector representation.

[199] In some embodiments, training the ML model to determine the degree of numerical similarity between input textual records and textual records in the corpus of textual records includes training the ML model to: determine, for a cluster of textual records in the corpus of textual records, an aggregate vector representation in a vector space; determine, for an input textual record, a vector representation in the vector space; and determine a degree of numerical similarity between the input textual record and the cluster of textual records in the corpus of textual records by determining a distance, within the vector space, between the vector representation of the input textual record and the aggregate vector representation.

[200] In some embodiments, training the ML model to determine the degree of numerical similarity between input textual records and textual records in the corpus of textual records includes training the ML model to: determine, for a cluster of textual records in the corpus of textual records, a representative volume within a vector space; determine, for an input textual record, a vector representation in the vector space; and determine a degree of numerical

similarity between the input textual record and the cluster of textual records in the corpus of textual records by determining whether the vector representation of the input textual record is within the representative volume.

[201] In some embodiments, the operations performed by the training computational instance additionally include: receiving, from an additional end-user computational instance that is dedicated to an additional managed network, an additional corpus of textual records; training, based on the additional corpus of textual records, an additional ML model to determine a degree of similarity between input textual records and textual records in the additional corpus of textual records; and transmitting, to the additional end-user computational instance, the additional ML model. In such embodiments, the operations performed by the prediction computational instance may additionally include: receiving, from the additional end-user computational instance, a further textual record; receiving, from the additional end-user computational instance, the additional ML model; determining, by the additional ML model, additional respective numerical similarities between the further textual record and the textual records in the additional corpus of textual records; and, based on the additional respective numerical similarities, transmitting, to the additional end-user computational instance, representations of one or more of the textual records in the additional corpus of textual records.

IX. Conclusion

[202] The present disclosure is not to be limited in terms of the particular embodiments described in this application, which are intended as illustrations of various aspects. Many modifications and variations can be made without departing from its scope, as will be apparent to those skilled in the art. Functionally equivalent methods and apparatuses within the scope of the disclosure, in addition to those described herein, will be apparent to those skilled in the art from

the foregoing descriptions. Such modifications and variations are intended to fall within the scope of the appended claims.

[203] The above detailed description describes various features and operations of the disclosed systems, devices, and methods with reference to the accompanying figures. The example embodiments described herein and in the figures are not meant to be limiting. Other embodiments can be utilized, and other changes can be made, without departing from the scope of the subject matter presented herein. It will be readily understood that the aspects of the present disclosure, as generally described herein, and illustrated in the figures, can be arranged, substituted, combined, separated, and designed in a wide variety of different configurations.

[204] With respect to any or all of the message flow diagrams, scenarios, and flow charts in the figures and as discussed herein, each step, block, and/or communication can represent a processing of information and/or a transmission of information in accordance with example embodiments. Alternative embodiments are included within the scope of these example embodiments. In these alternative embodiments, for example, operations described as steps, blocks, transmissions, communications, requests, responses, and/or messages can be executed out of order from that shown or discussed, including substantially concurrently or in reverse order, depending on the functionality involved. Further, more or fewer blocks and/or operations can be used with any of the message flow diagrams, scenarios, and flow charts discussed herein, and these message flow diagrams, scenarios, and flow charts can be combined with one another, in part or in whole.

[205] A step or block that represents a processing of information can correspond to circuitry that can be configured to perform the specific logical functions of a herein-described method or technique. Alternatively or additionally, a step or block that represents a processing of information can correspond to a module, a segment, or a portion of program code (including

related data). The program code can include one or more instructions executable by a processor for implementing specific logical operations or actions in the method or technique. The program code and/or related data can be stored on any type of computer readable medium such as a storage device including RAM, a disk drive, a solid state drive, or another storage medium.

[206] The computer readable medium can also include non-transitory computer readable media such as computer readable media that store data for short periods of time like register memory and processor cache. The computer readable media can further include non-transitory computer readable media that store program code and/or data for longer periods of time. Thus, the computer readable media may include secondary or persistent long term storage, like ROM, optical or magnetic disks, solid state drives, compact-disc read only memory (CD-ROM), for example. The computer readable media can also be any other volatile or non-volatile storage systems. A computer readable medium can be considered a computer readable storage medium, for example, or a tangible storage device.

[207] Moreover, a step or block that represents one or more information transmissions can correspond to information transmissions between software and/or hardware modules in the same physical device. However, other information transmissions can be between software modules and/or hardware modules in different physical devices.

[208] The particular arrangements shown in the figures should not be viewed as limiting. It should be understood that other embodiments can include more or less of each element shown in a given figure. Further, some of the illustrated elements can be combined or omitted. Yet further, an example embodiment can include elements that are not illustrated in the figures.

[209] While various aspects and embodiments have been disclosed herein, other aspects and embodiments will be apparent to those skilled in the art. The various aspects and

embodiments disclosed herein are for purpose of illustration and are not intended to be limiting, with the true scope being indicated by the following claims.

CLAIMS

What is claimed is:

1. A remote network management platform comprising:

an end-user computational instance comprising a first set of computational resources of the remote network management platform and dedicated to a managed network;

a training computational instance comprising a second set of computational resources of the remote network management platform, and configured to perform operations including:

receiving, from the end-user computational instance, a corpus of textual records,

training, based on the corpus of textual records, a machine learning (ML) model to determine a degree of numerical similarity between input textual records and textual records in the corpus of textual records, and

transmitting, to the end-user computational instance, the ML model; and

a prediction computational instance comprising a third set of computational resources of the remote network management platform, and configured to perform operations including:

receiving, from the end-user computational instance, an additional textual record,

receiving, from the end-user computational instance, the ML model,

determining, by the ML model, respective numerical similarities between the additional textual record and the textual records in the corpus of textual records, and

based on the respective numerical similarities, transmitting, to the end-user computational instance, representations of one or more of the textual records in the corpus of textual records.
2. The remote network management platform of claim 1, wherein the third set of

computational resources includes more memory than the first set of computational resources.

3. The remote network management platform of claim 1, wherein the third set of computational resources includes a memory, and wherein the operations performed by the prediction computational instance additionally include:

in response to receiving the additional textual record, determining that the ML model is not stored in the memory; and

requesting, from the end-user computational instance, the ML model.

4. The remote network management platform of claim 1, wherein determining, by the ML model, respective numerical similarities between the additional textual record and one or more of the textual records in the corpus of textual records comprises using the ML model to determine at least one of (i) word vectors that describe, in a first semantically-encoded vector space, a meaning of respective words of the additional textual record, or (ii) a paragraph vector that describes, in a second semantically-encoded vector space, a meaning of multiple words of the additional textual record.

5. The remote network management platform of claim 4, wherein the ML model represents a set of clusters of the textual records in the corpus of textual records, and wherein determining, by the ML model, respective numerical similarities between the additional textual record and the textual records in the corpus of textual records further comprises selecting, for the additional textual record, a cluster from the set of clusters based on at least one of the word vectors or the paragraph vector.

6. The remote network management platform of claim 1, wherein the ML model represents a set of clusters of textual records in the corpus of textual records, and wherein determining, by the ML model, respective numerical similarities between the additional textual record and the textual records in the corpus of textual records comprises selecting, for the additional textual record, a cluster from the set of clusters.

7. The remote network management platform of claim 1, wherein training the ML model to determine the degree of numerical similarity between input textual records and textual records in the corpus of textual records comprises training the ML model to:

determine, for an input textual record and for a given textual record in the corpus of textual records, respective vector representations in a vector space; and

determine a degree of numerical similarity between the input textual record and the given textual record by determining a distance, within the vector space, between the vector representations of the input textual record and the given textual record.

8. The remote network management platform of claim 1, wherein training the ML model to determine the degree of numerical similarity between input textual records and textual records in the corpus of textual records comprises training the ML model to:

determine, for an input textual record and for a given textual record in the corpus of textual records, respective vector representations in a vector space; and

determine a degree of numerical similarity between the input textual record and the given textual record by determining an angle, within the vector space, between the vector representations of the input textual record and the given textual record.

9. The remote network management platform of claim 1, wherein training the ML model to determine the degree of numerical similarity between input textual records and textual records in the corpus of textual records comprises training the ML model to:

determine, for a cluster of textual records in the corpus of textual records, an aggregate vector representation in a vector space;

determine, for an input textual record, a vector representation in the vector space; and

determine a degree of numerical similarity between the input textual record and the cluster of textual records in the corpus of textual records by determining an angle, within the vector space, between the vector representation of the input textual record and the aggregate vector representation.

10. The remote network management platform of claim 1, wherein training the ML model to determine the degree of numerical similarity between input textual records and textual records in the corpus of textual records comprises training the ML model to:

determine, for a cluster of textual records in the corpus of textual records, an aggregate vector representation in a vector space;

determine, for an input textual record, a vector representation in the vector space; and

determine a degree of numerical similarity between the input textual record and the cluster of textual records in the corpus of textual records by determining a distance, within the vector space, between the vector representation of the input textual record and the aggregate vector representation.

11. The remote network management platform of claim 1, wherein training the ML model to determine the degree of numerical similarity between input textual records and textual

records in the corpus of textual records comprises training the ML model to:

determine, for a cluster of textual records in the corpus of textual records, a representative volume within a vector space;

determine, for an input textual record, a vector representation in the vector space; and

determine a degree of numerical similarity between the input textual record and the cluster of textual records in the corpus of textual records by determining whether the vector representation of the input textual record is within the representative volume.

12. The remote network management platform of claim 1, further comprising:

an additional end-user computational instance comprising a fourth set of computational resources of the remote network management platform and dedicated to an additional managed network;

wherein the operations performed by the training computational instance additionally include:

receiving, from the additional end-user computational instance, an additional corpus of textual records,

training, based on the additional corpus of textual records, an additional ML model to determine a degree of similarity between input textual records and textual records in the additional corpus of textual records, and

transmitting, to the additional end-user computational instance, the additional ML model; and

wherein the operations performed by the prediction computational instance additionally include:

receiving, from the additional end-user computational instance, a further textual record,

receiving, from the additional end-user computational instance, the additional ML model,

determining, by the additional ML model, additional respective numerical similarities between the further textual record and the textual records in the additional corpus of textual records, and

based on the additional respective numerical similarities, transmitting, to the additional end-user computational instance, representations of one or more of the textual records in the additional corpus of textual records.

13. An article of manufacture including a non-transitory computer-readable medium, having stored thereon program instructions that, upon execution by a remote network management platform, cause the remote network management platform to perform operations comprising:

operating a first set of computational resources of the remote network management platform to provide an end-user computational instance that is dedicated to a managed network;

operating a second set of computational resources of the remote network management platform to provide a training computational instance that is configured to perform operations including: (i) receiving, from the end-user computational instance, a corpus of textual records, (ii) training, based on the corpus of textual records, a machine learning (ML) model to determine a degree of numerical similarity between input textual records and textual records in the corpus of textual records, and (iii) transmitting, to the end-user computational instance, the ML model; and

operating a third set of computational resources of the remote network management platform to provide a prediction computational instance that is configured to perform operations including: (i) receiving, from the end-user computational instance, an additional textual record, (ii) receiving, from the end-user computational instance, the ML model, (iii) determining, by the ML model, respective numerical similarities between the additional textual record and the textual records in the corpus of textual records, and (iv) based on the respective numerical similarities, transmitting, to the end-user computational instance, representations of one or more of the textual records in the corpus of textual records.

14. The article of manufacture of claim 13, wherein the operations performed by the prediction computational instance additionally include:

in response to receiving the additional textual record, determining that the ML model is not stored in a memory of the third set of computational resources; and

requesting, from the end-user computational instance, the ML model.

15. The article of manufacture of claim 13, wherein determining, by the ML model, respective numerical similarities between the additional textual record and one or more of the textual records in the corpus of textual records comprises using the ML model to determine at least one of (i) word vectors that describe, in a first semantically-encoded vector space, a meaning of respective words of the additional textual record, or (ii) a paragraph vector that describes, in a second semantically-encoded vector space, a meaning of multiple words of the additional textual record.

16. The article of manufacture of claim 15, wherein the ML model represents a set of

clusters of the textual records in the corpus of textual records, and wherein determining, by the ML model, respective numerical similarities between the additional textual record and the textual records in the corpus of textual records further comprises selecting, for the additional textual record, a cluster from the set of clusters based on at least one of the word vectors or the paragraph vector.

17. The article of manufacture of claim 13, wherein the ML model represents a set of clusters of textual records in the corpus of textual records, and wherein determining, by the ML model, respective numerical similarities between the additional textual record and the textual records in the corpus of textual records comprises selecting, for the additional textual record, a cluster from the set of clusters.

18. The article of manufacture of claim 13, wherein training the ML model to determine the degree of numerical similarity between input textual records and textual records in the corpus of textual records comprises training the ML model to:

determine, for an input textual record and for a given textual record in the corpus of textual records, respective vector representations in a vector space; and

determine a degree of numerical similarity between the input textual record and the given textual record by determining a distance, within the vector space, between the vector representations of the input textual record and the given textual record.

19. The article of manufacture of claim 13, wherein training the ML model to determine the degree of numerical similarity between input textual records and textual records in the corpus of textual records comprises training the ML model to:

determine, for an input textual record and for a given textual record in the corpus of textual records, respective vector representations in a vector space; and

determine a degree of numerical similarity between the input textual record and the given textual record by determining an angle, within the vector space, between the vector representations of the input textual record and the given textual record.

20. A method comprising:

operating a first set of computational resources of a remote network management platform to provide an end-user computational instance that is dedicated to a managed network;

operating a second set of computational resources of the remote network management platform to provide a training computational instance that is configured to perform operations including: (i) receiving, from the end-user computational instance, a corpus of textual records, (ii) training, based on the corpus of textual records, a machine learning (ML) model to determine a degree of numerical similarity between input textual records and textual records in the corpus of textual records, and (iii) transmitting, to the end-user computational instance, the ML model; and

operating a third set of computational resources of the remote network management platform to provide a prediction computational instance that is configured to perform operations including: (i) receiving, from the end-user computational instance, an additional textual record, (ii) receiving, from the end-user computational instance, the ML model, (iii) determining, by the ML model, respective numerical similarities between the additional textual record and the textual records in the corpus of textual records, and (iv) based on the respective numerical similarities, transmitting, to the end-user computational instance, representations of one or more of the textual records in the corpus of textual records.

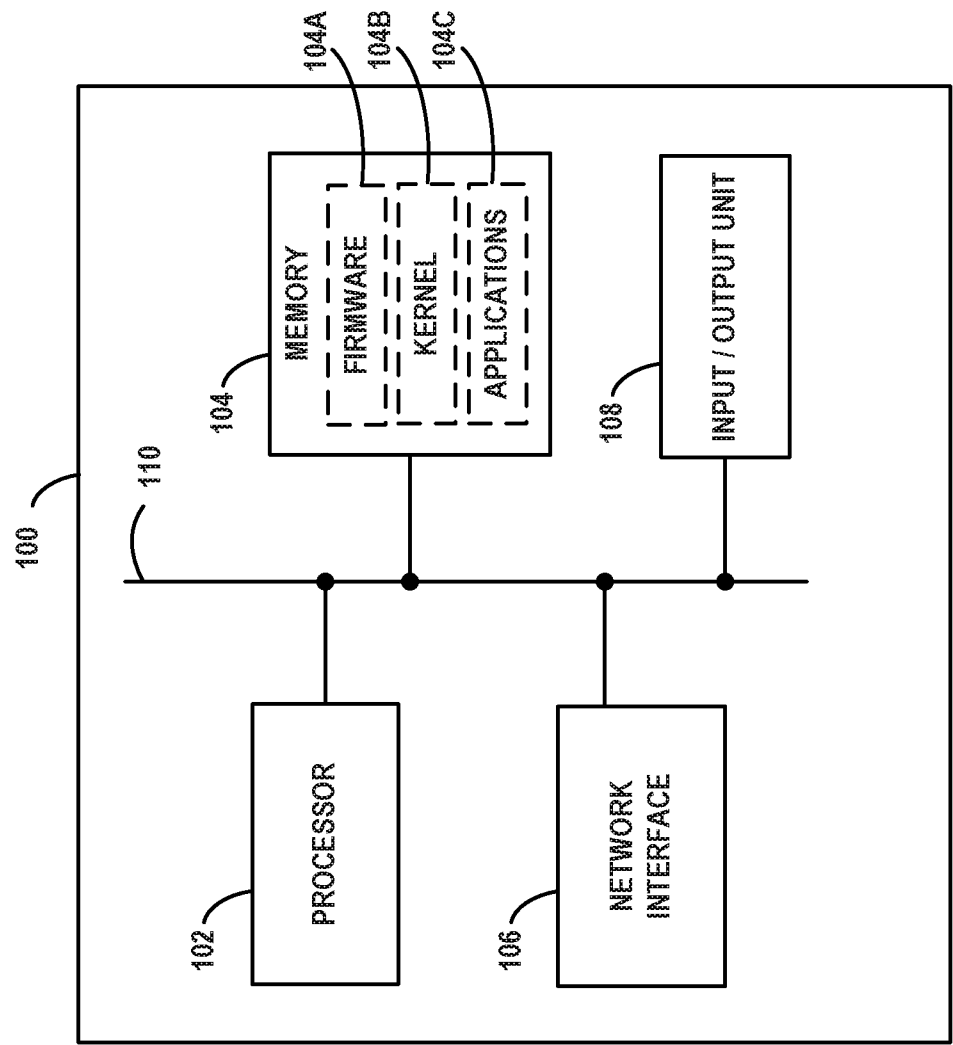


FIG. 1

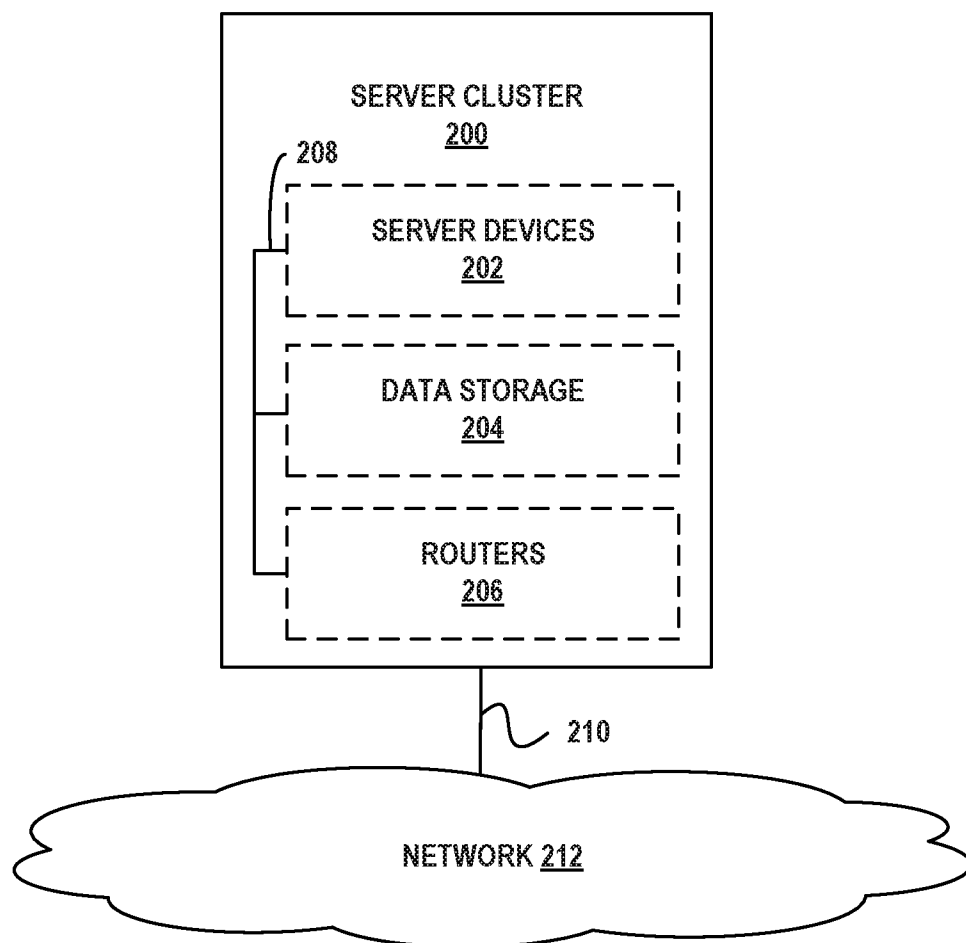


FIG. 2

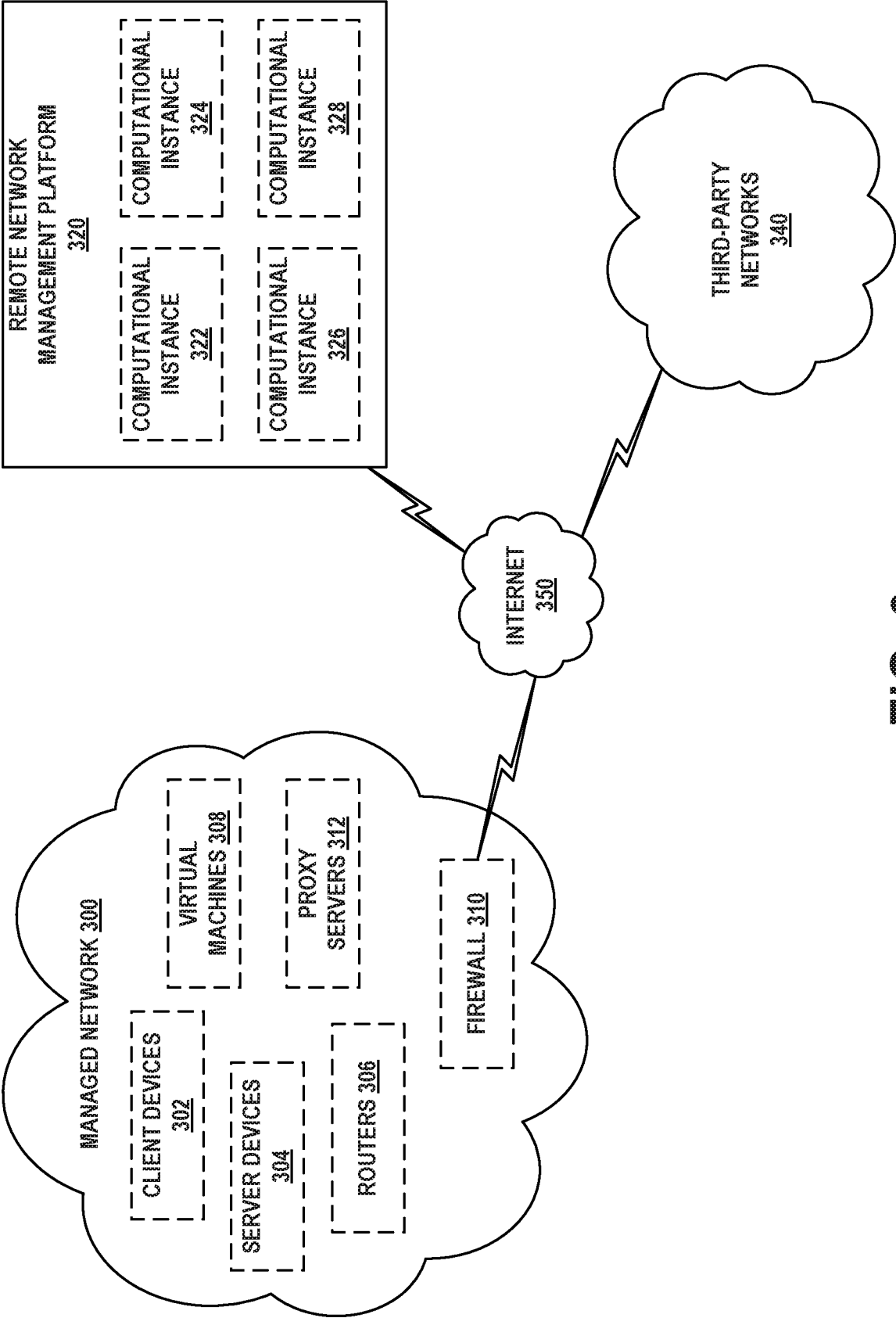


FIG. 3

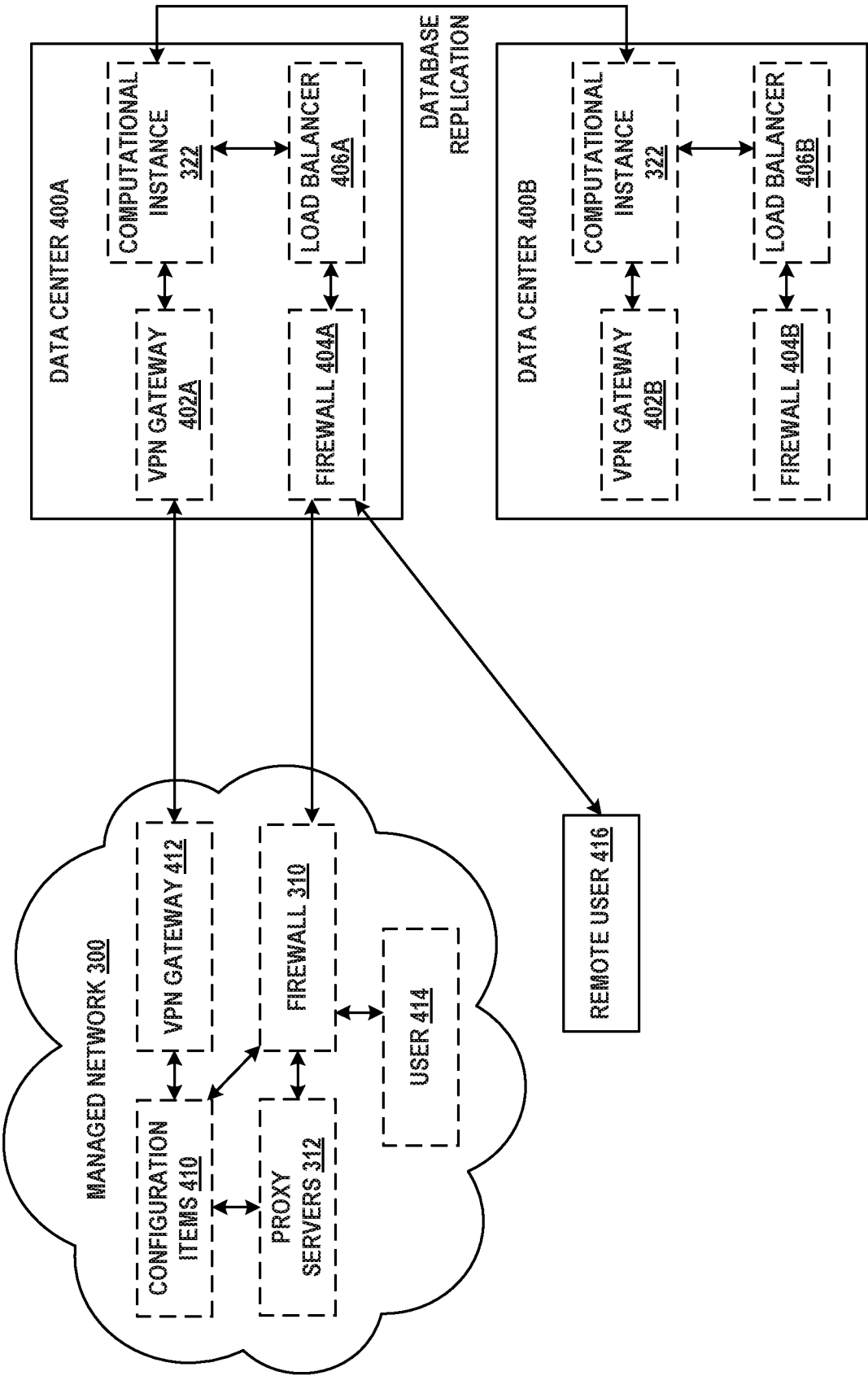


FIG. 4

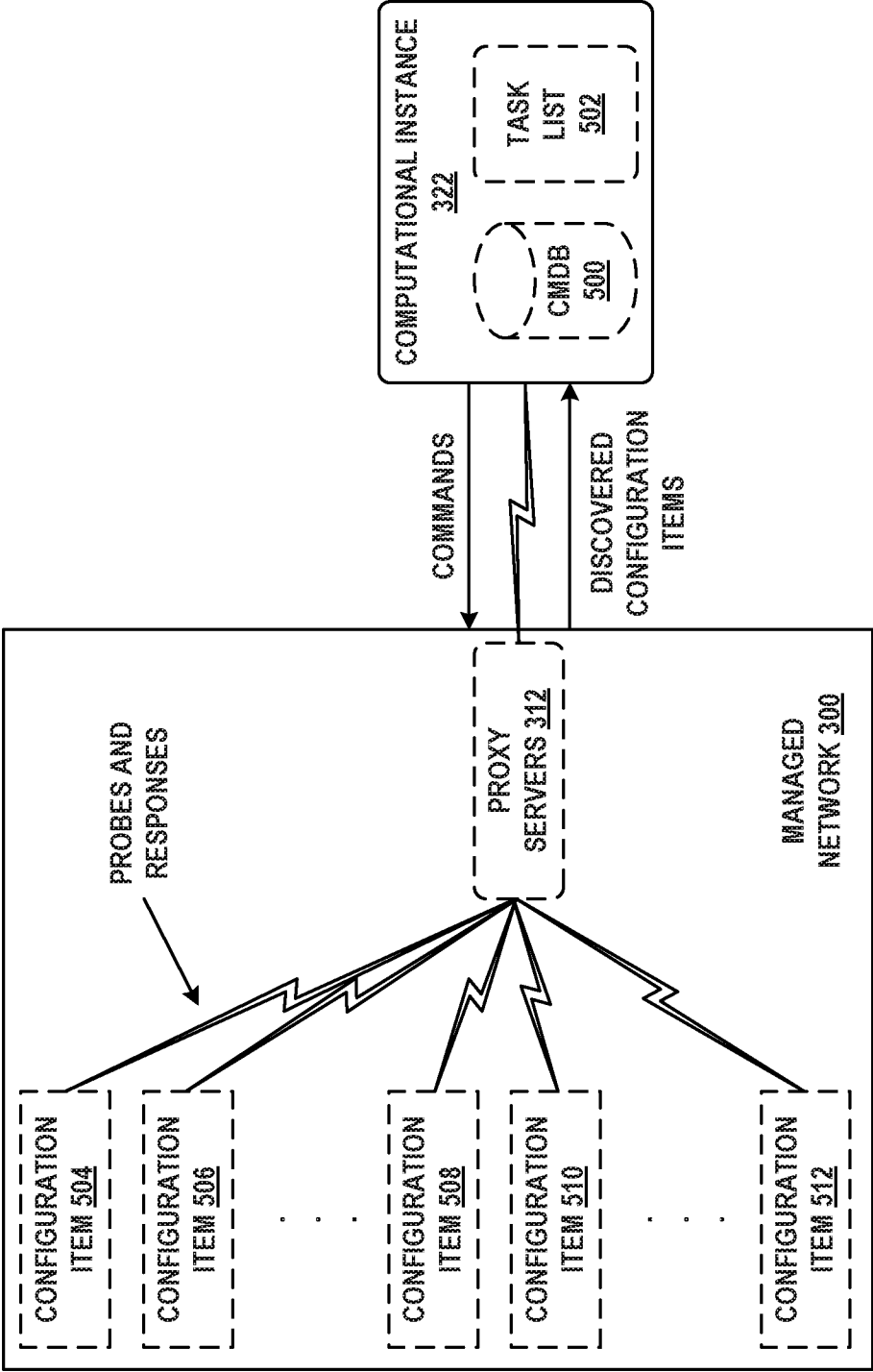
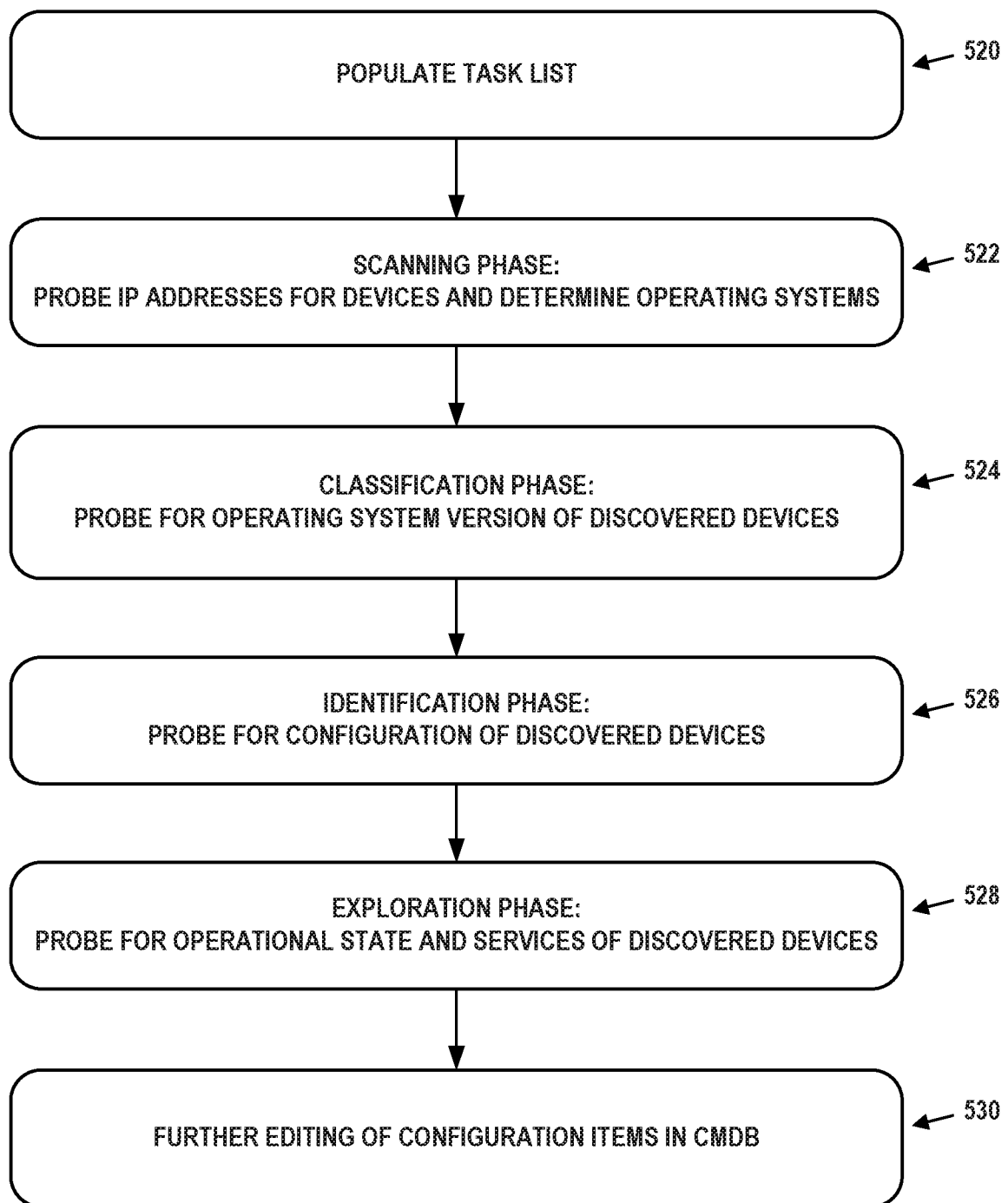


FIG. 5A

6 of 16

**FIG. 5B**

7 of 16

600
↓

ORIGINATOR	BOB SMITH	← 602
CREATED	2018-02-07 9:56AM	← 604
SHORT DESCRIPTION	CAN'T CONNECT TO EMAIL	← 605
PROBLEM DESCRIPTION	MY EMAIL CLIENT IS NOT DOWNLOADING NEW EMAILS. I AM CONNECTED TO THE WIRELESS NETWORK AND CAN ACCESS THE INTERNET, BUT MY CLIENT KEEPS REPORTING THAT IT IS UNABLE TO ACCESS THE SERVER.	← 606
CATEGORY	EMAIL	← 608
ASSIGNED TO	ALICE JONES	← 610
STATUS	RESOLVED	← 612
RESOLUTION	BOB'S LAPTOP WAS CONNECTED TO THE "GUEST" WIRELESS NETWORK, WHICH IS OUTSIDE OF THE FIREWALL. I TOLD HIM THAT HE COULD EITHER USE THE VPN TO ACCESS HIS EMAIL, OR CONNECT TO THE "EMPLOYEE" WIRELESS NETWORK, WHICH REQUIRES THE PASSWORD BUT IS INSIDE THE FIREWALL.	← 614
RESOLVED TIME	2018-02-07 10:10AM	← 616
CLOSURE CODE	SOLVED (PERMANENTLY)	← 618
NOTES	THIS IS THE THIRD TIME IN A WEEK THAT I'VE HAD TO ADDRESS A SIMILAR ISSUE. CAN WE POST WIRELESS NETWORKING INSTRUCTIONS IN ALL CONFERENCE ROOMS?	← 620
LINK TO KNOWLEDGE BASE	ITKB/WIRELESS/CONFIGURATION.HTML	← 622

FIG. 6

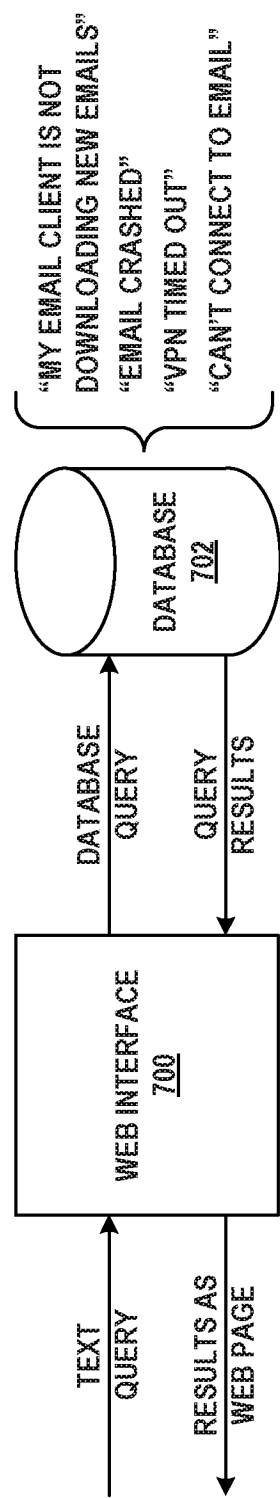


FIG. 7

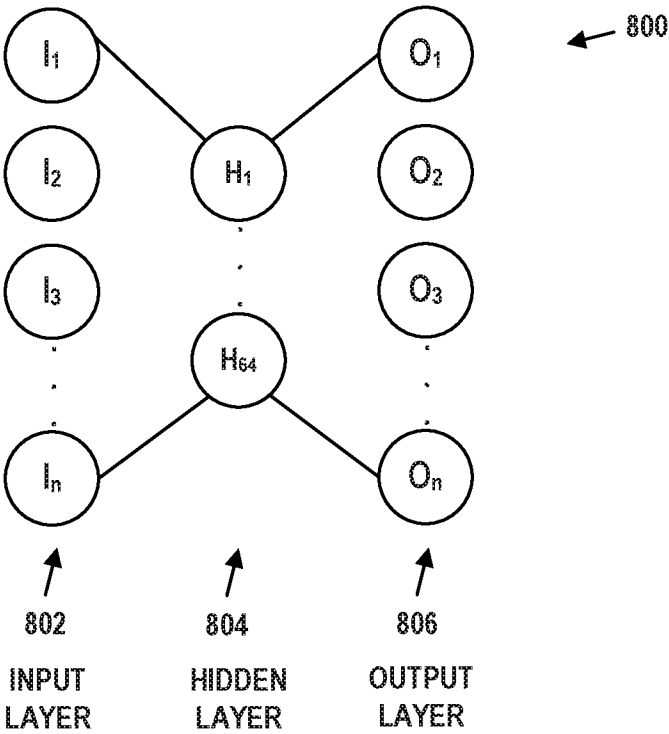


FIG. 8A

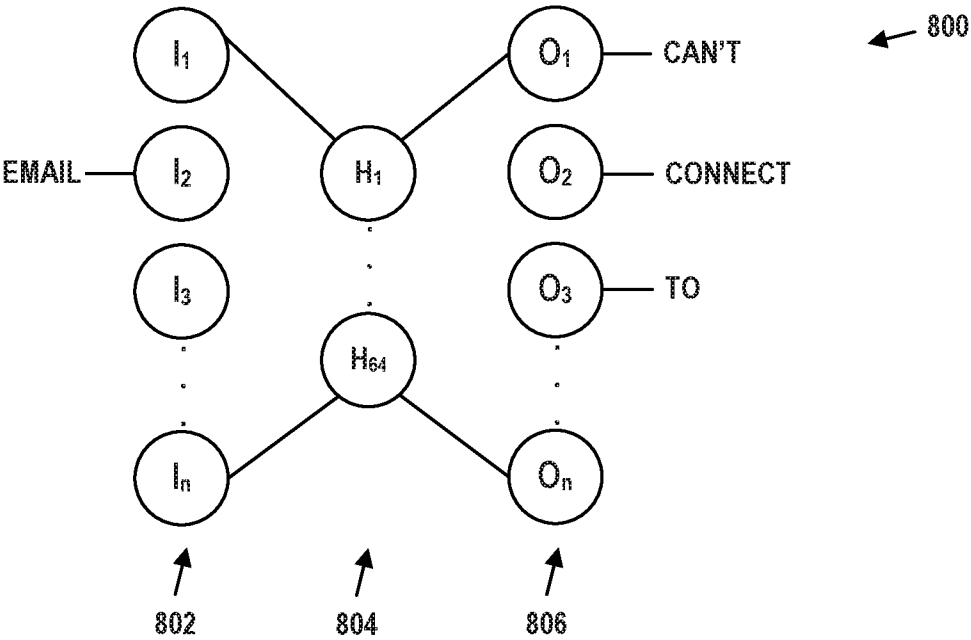


FIG. 8B

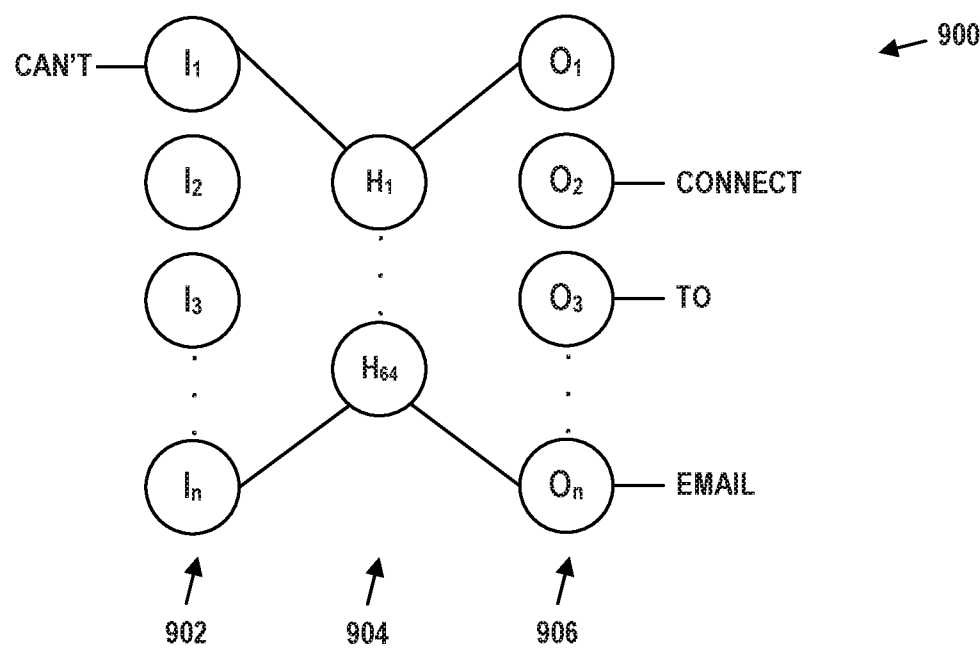


FIG. 9A

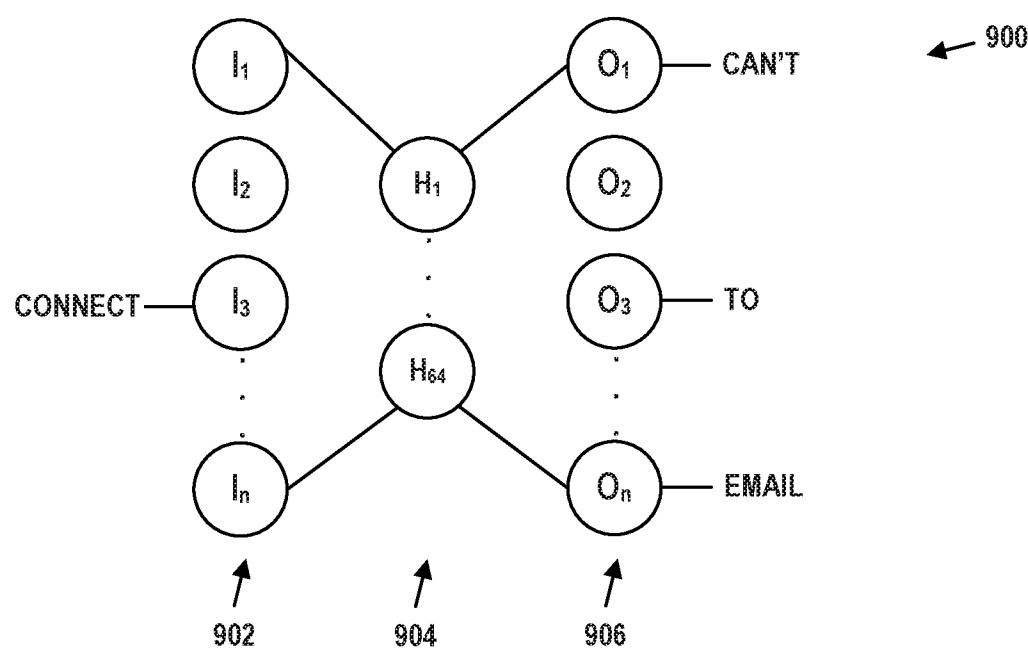


FIG. 9B

11 of 16

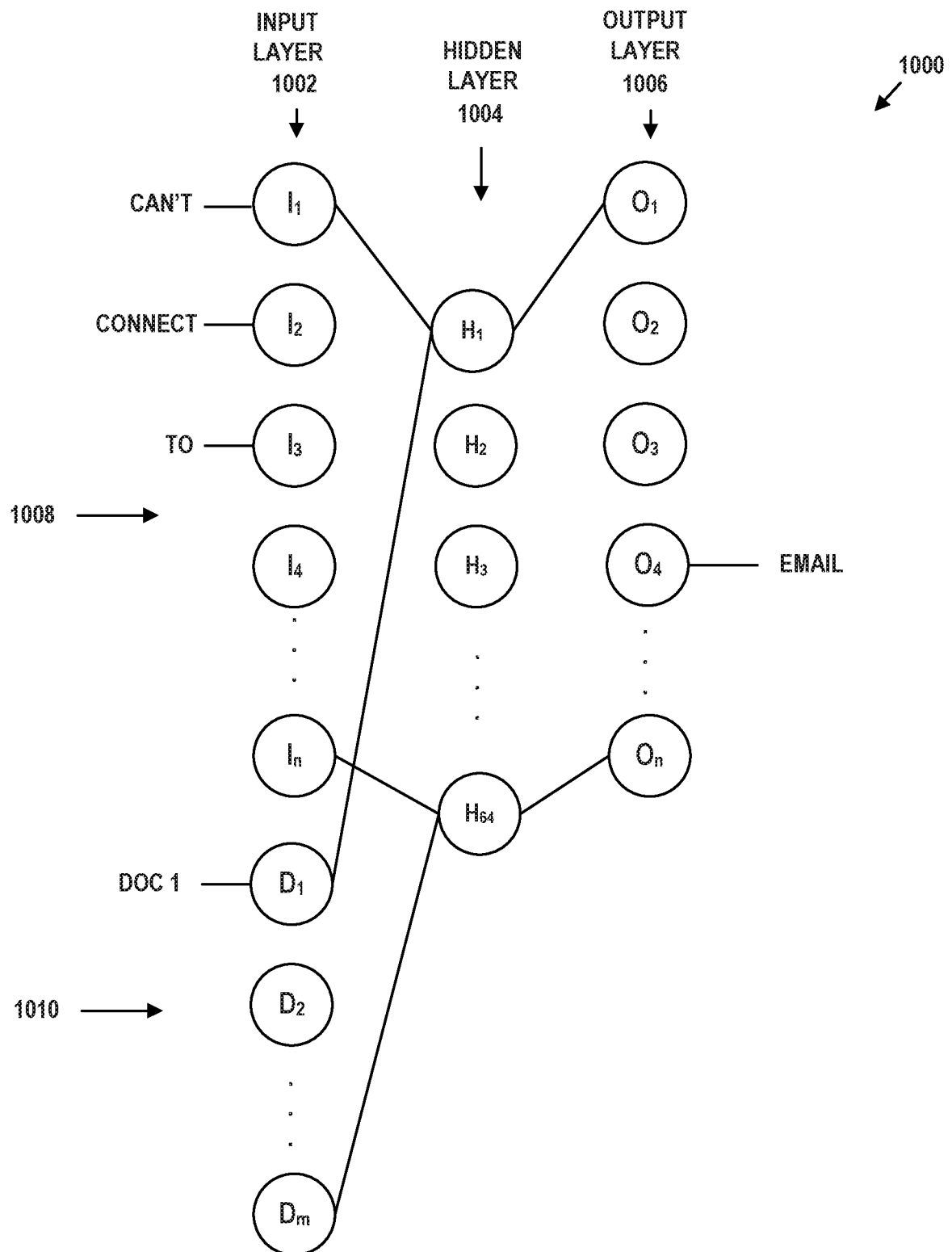
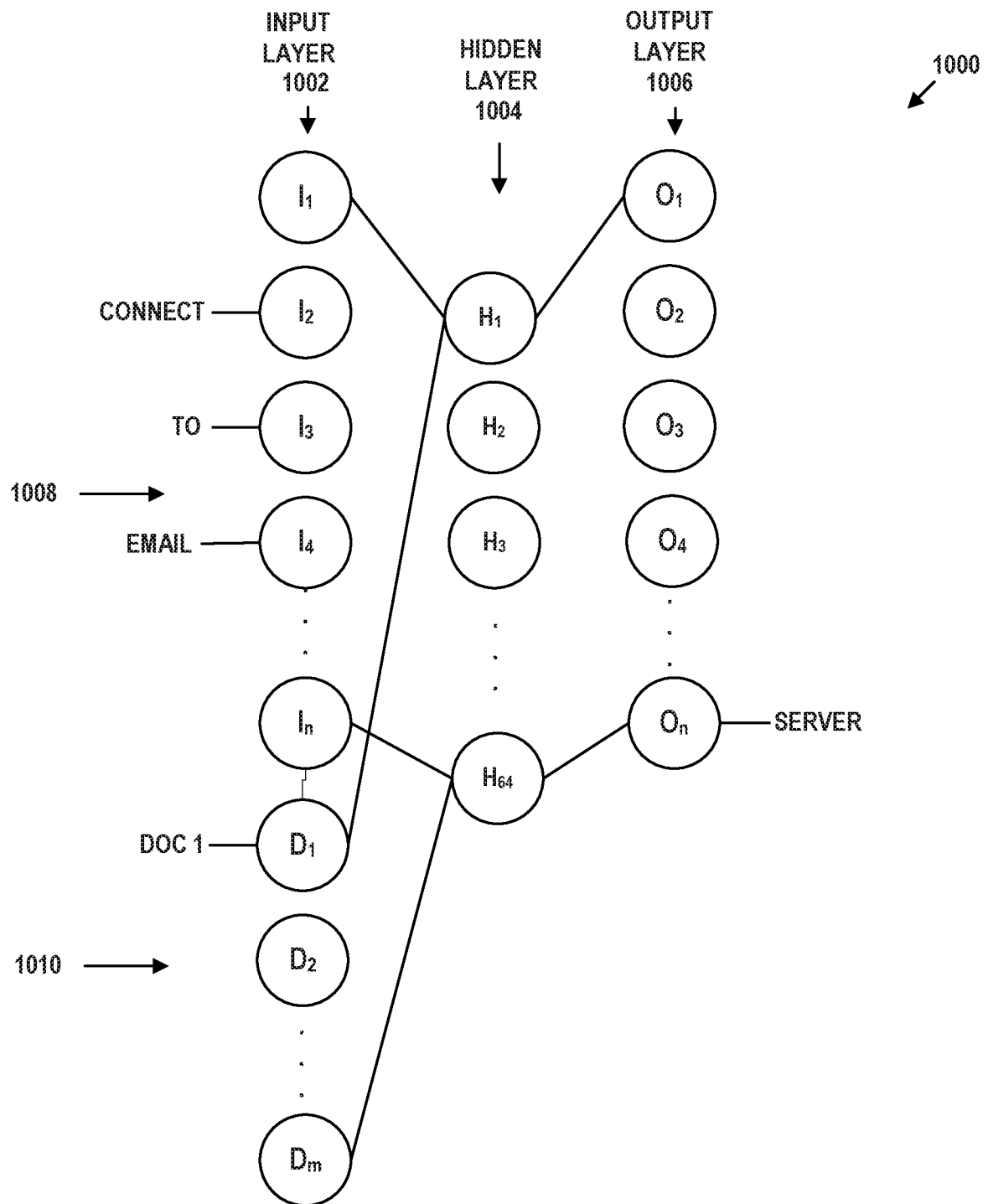


FIG. 10A

12 of 16

**FIG. 10B**

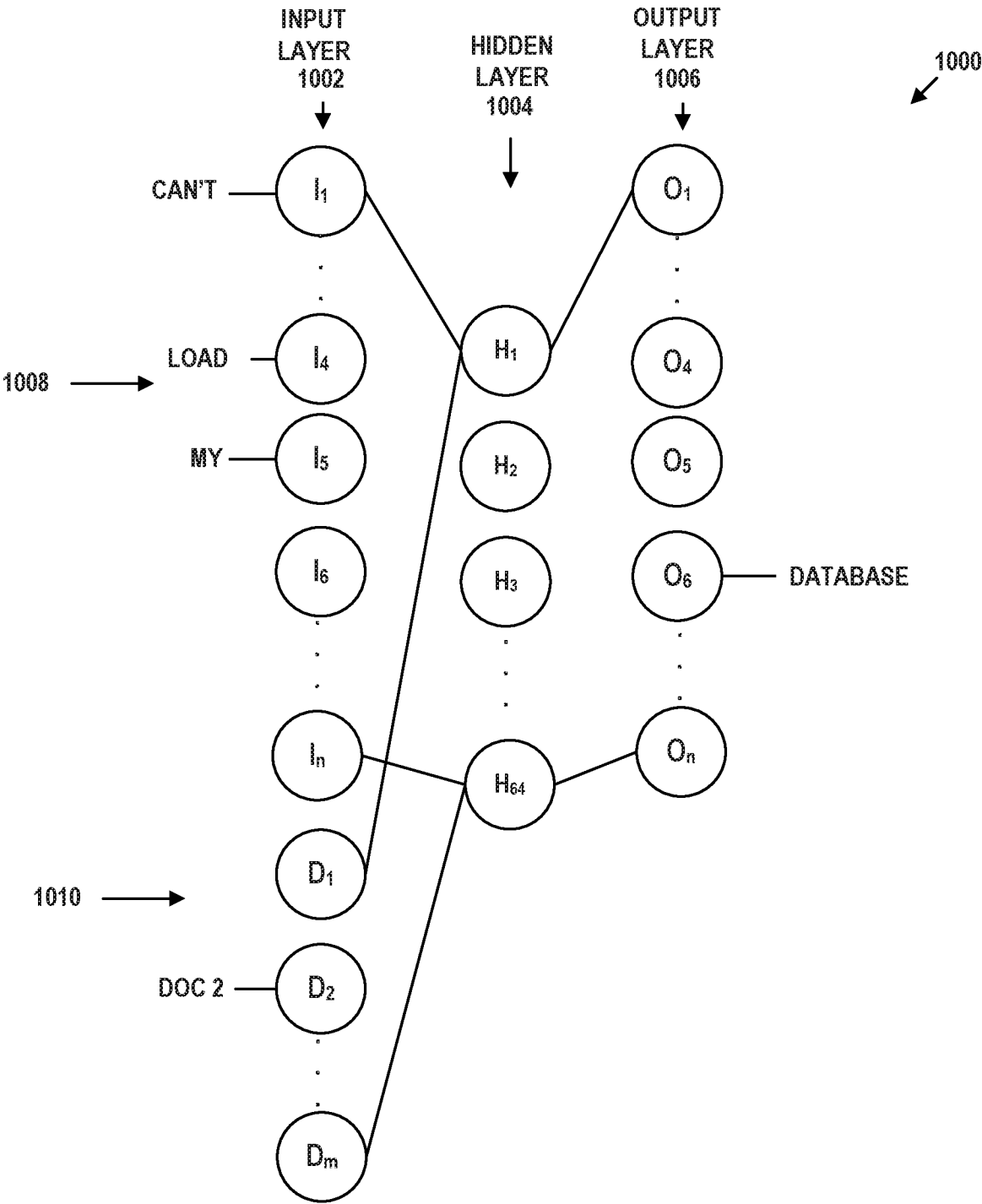


FIG. 10C

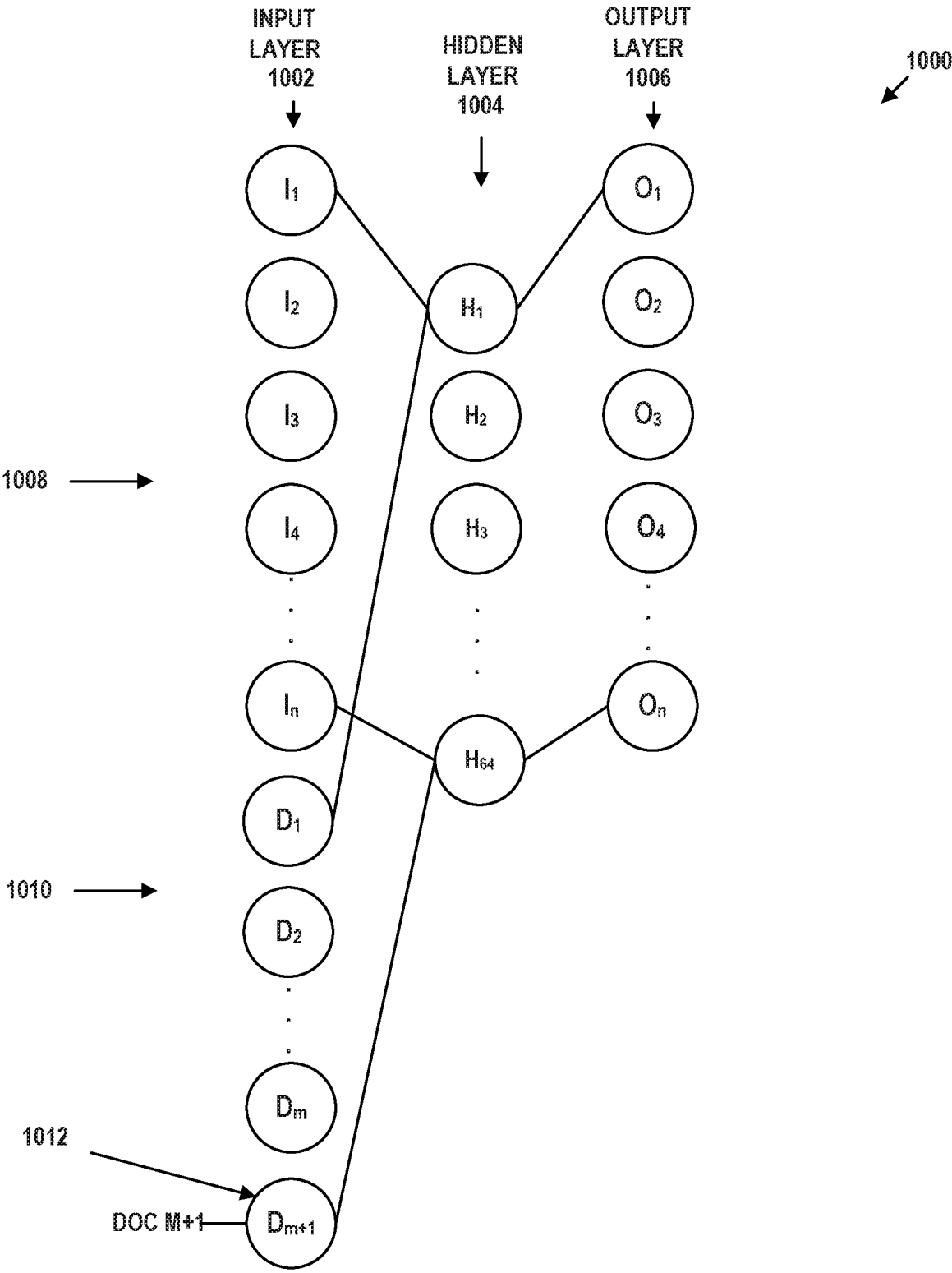


FIG. 10D

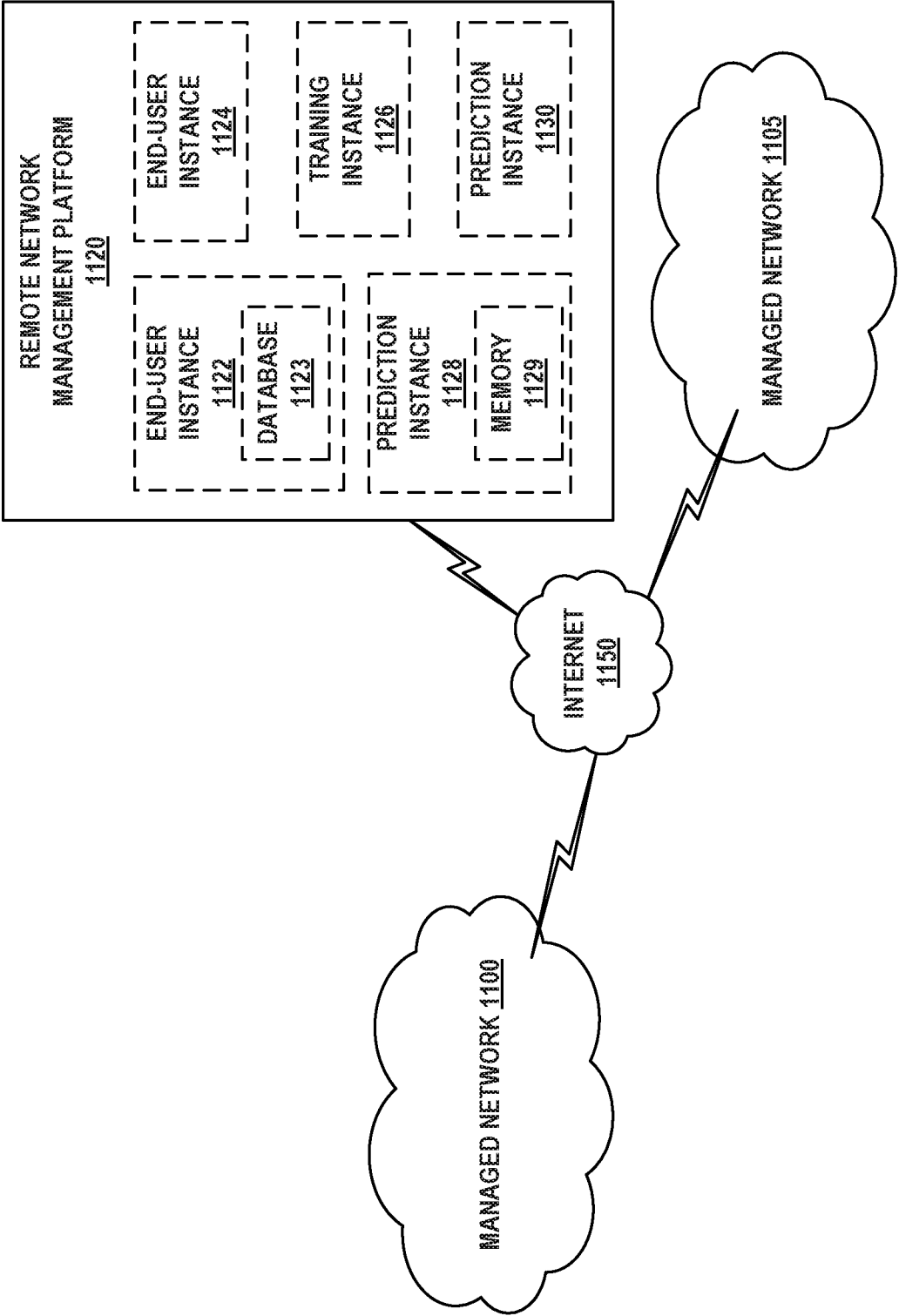


FIG. 11

16 of 16

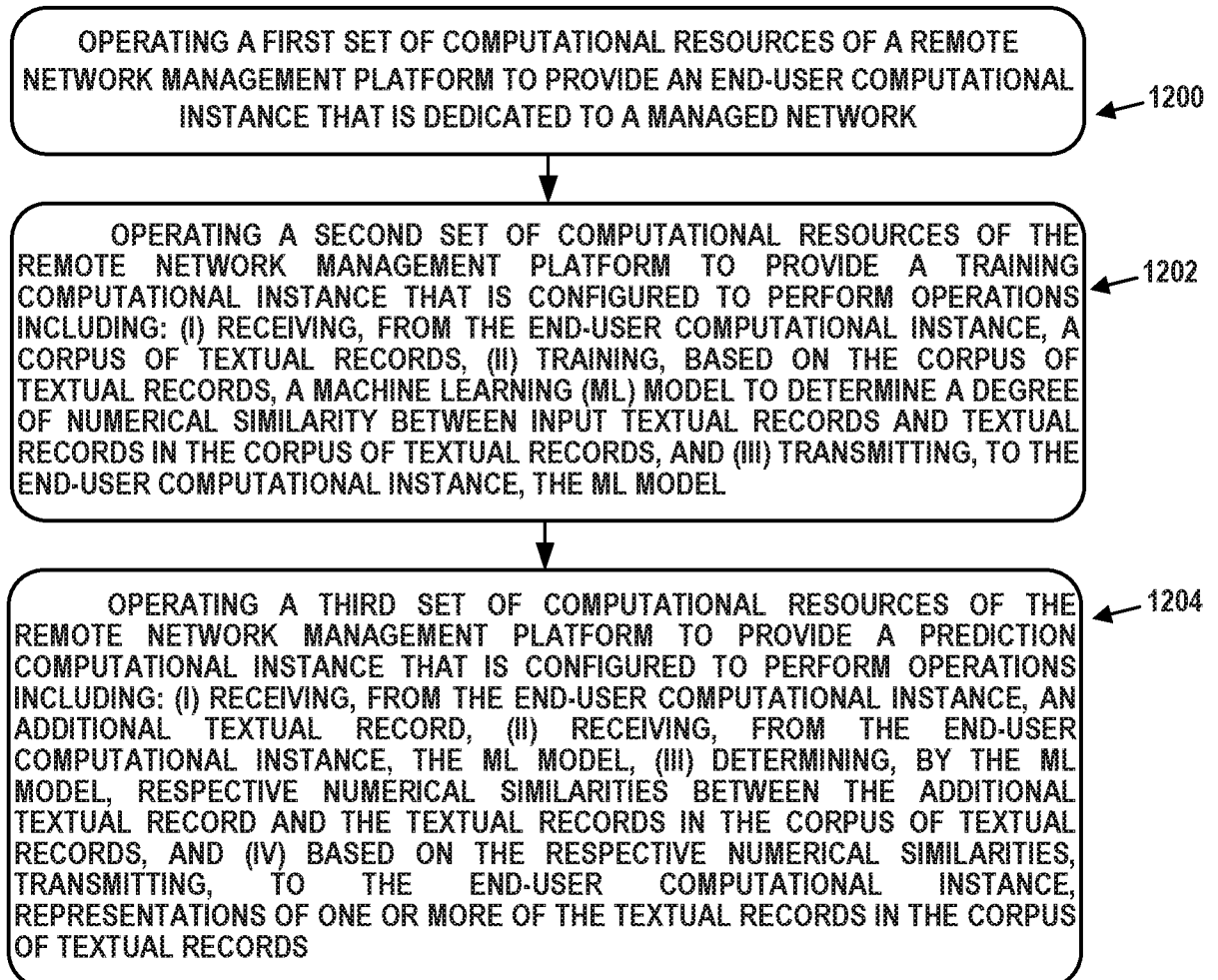


FIG. 12

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2020/031170

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F16/35
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	EP 3 462 387 A1 (SERVICENOW INC [US]) 3 April 2019 (2019-04-03) abstract paragraphs [0001], [0092] - [0097] -----	1-20
X	US 2017/243112 A1 (EKAMBARAM VIJAY [IN] ET AL) 24 August 2017 (2017-08-24) abstract pages 1,3,9-13; figure 3B ----- -/--	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search

9 July 2020

Date of mailing of the international search report

17/07/2020

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Siódmok, Wojciech

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2020/031170

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	GUPTA R ET AL: "Multi-dimensional Knowledge Integration for Efficient Incident Management in a Services Cloud", SERVICES COMPUTING, 2009. SCC '09. IEEE INTERNATIONAL CONFERENCE ON, IEEE, PISCATAWAY, NJ, USA, 21 September 2009 (2009-09-21), pages 57-64, XP031544499, ISBN: 978-1-4244-5183-8 abstract section "III. Automated incident classification" -----	1-20

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2020/031170

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 3462387	A1	03-04-2019	
		AU 2018200014 A1	18-04-2019
		AU 2019210675 A1	22-08-2019
		CA 2990452 A1	02-04-2019
		EP 3462387 A1	03-04-2019
		US 2019102682 A1	04-04-2019
		US 2019102683 A1	04-04-2019

US 2017243112	A1	24-08-2017	NONE
