



(51) International Patent Classification:

G06N 5/04 (2006.01) G06F 21/55 (2013.01)
G06N 99/00 (2010.01)

(21) International Application Number:

PCT/US2018/018332

(22) International Filing Date:

15 February 2018 (15.02.2018)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/473,694	20 March 2017 (20.03.2017)	US
62/474,141	21 March 2017 (21.03.2017)	US
62/537,015	26 July 2017 (26.07.2017)	US
15/889,666	06 February 2018 (06.02.2018)	US

(71) Applicant: NEC LABORATORIES AMERICA, INC
[US/US]; 4 Independence Way, Suite 200, Princeton, New
Jersey 08540 (US).

(72) Inventors: ZHANG, Hui; 69 Southfield Road, Prince-
ton Junction, New Jersey 08550 (US). XU, Jianwu; 95
Oneill Court, Lawrenceville, New Jersey 08648 (US). DEB-
NATH, Biplob; 36G Needham Way, Princeton, New Jer-
sey 08540 (US).

(74) Agent: KOLODKA, Joseph; NEC Laboratories America,
Inc., 4 Independence Way, Suite 200, Princeton, New Jersey
08540 (US).

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

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

(54) Title: METHOD AND SYSTEM FOR INCREMENTALLY LEARNING LOG PATTERNS ON HETEROGENEOUS LOGS

(57) Abstract: Systems and methods for enabling automated log analysis with
controllable resource requirements are provided. A training set for log pattern
learning is generated based on heterogeneous logs generated by a computer sys-
tem. An incremental learning process is implemented to generate a set of log
patterns from the training set. The heterogeneous logs are parsed using the set of
log patterns. A set of applications is applied to the parsed logs.

100

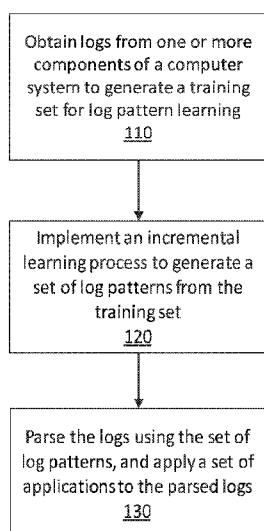


FIG. 1

Published:

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

METHOD AND SYSTEM FOR INCREMENTALLY LEARNING LOG PATTERNS ON HETEROGENEOUS LOGS

RELATED APPLICATION INFORMATION

[0001] This application claims priority to U.S. Provisional Patent Application No. 62/473,694, filed on March 20, 2017, U.S. Provisional Patent Application No. 62/474,141, filed on March 21, 2017, and U.S. Provisional Patent Application No. 62/537,015, filed on July 25, 2017, all incorporated by reference herein in their entirety.

[0002] This application is related to applications entitled “Automatic and Scalable Log Pattern Learning in Security Log Analysis,” having attorney docket number 16092A, and “Security System Using Automatic and Scalable Log Pattern Learning in Security Log Analysis,” having attorney docket number 16092B, all incorporated by reference herein in their entirety.

BACKGROUND

Technical Field

[0003] The present invention relates to log analysis, and more particularly to enabling automated log analysis with controllable resource requirements.

Description of the Related Art

[0004] A computer-generated record, referred to as a log or a log record, is data output generated from source code associated with a computer system component (e.g., network device, operating system, application and programmable devices) that documents activities performed by the computer system component. A log can include one or more

lines of sentences that record system/application events or status at a time corresponding to a time stamp. Log analytics can be performed for a variety of practical applications, such as software debugging, compliance, computer system troubleshooting, forensics and investigations, security incident response and web analytics (e.g., search engine optimization).

[0005] Computer systems can generate a vast number of logs from a variety of components, such as in the form of heterogeneous logs. Heterogeneous logs can provide rich contextual information describing system status and are critical sources for system monitoring and diagnosis. However, manually interpreting heterogeneous logs is not effective due to large volume and complicated domain-specific syntax and semantic knowledge.

SUMMARY

[0006] According to an aspect of the present principles, a computer-implemented method is provided for enabling automated log analysis with controllable resource requirements. The method includes a processor operatively coupled to a memory generating a training set for log pattern learning based on heterogeneous logs generated by a computer system. The method further includes the processor implementing an incremental learning process to generate a set of log patterns from the training set. Implementing the incremental learning process includes: defining a first set as the training set, a second set as a set of log patterns that have been generated, and a third set as a set of logs of the training set that lack a matching pattern in the second set; sampling the third set to generate a fourth set having a size corresponding to a parameter

controlling a maximum resource requirement for the incremental learning process; and performing automatic log pattern recognition to generate a fifth set. The method further includes the processor parsing the heterogeneous logs using the set of log patterns, and applying a set of applications to the parsed logs.

[0007] According to another aspect of the present principles, a computer program product having program instructions embodied therewith is provided. The program instructions are executable by a computer to cause the computer to perform a method for enabling automated log analysis with controllable resource requirements. The method includes generating a training set for log pattern learning based on heterogeneous logs generated by a computer system. An incremental learning process is implemented to generate a set of log patterns from the training set. Implementing the incremental learning process includes: defining a first set as the training set, a second set as a set of log patterns that have been generated, and a third set as a set of logs of the training set that lack a matching pattern in the second set; sampling the third set to generate a fourth set having a size corresponding to a parameter controlling a maximum resource requirement for the incremental learning process; and performing automatic log pattern recognition to generate a fifth set. The heterogeneous logs are parsed using the set of log patterns. A set of applications is applied to the parsed logs.

[0008] According to another aspect of the present principles, a system is provided for enabling automated log analysis with controllable resource requirements. The system includes a computer system configured to generate heterogeneous logs, and at least one processor operatively coupled to a memory. The at least one processor is configured to obtain the heterogeneous logs from the computer system, generate a training set for log

pattern learning based on the obtained logs, implement an incremental learning process to generate a set of log patterns from the training set, parse the one or more logs using the set of log patterns, and apply a set of applications to the parsed logs. The incremental learning process is implemented by: defining a first set as the training set, a second set as a set of log patterns that have been generated, and a third set as a set of logs of the training set that lack a matching pattern in the second set; sampling the third set to generate a fourth set having a size corresponding to a parameter controlling a maximum resource requirement for the incremental learning process; and performing automatic log pattern recognition to generate a fifth set.

[0009] These and other features and advantages will become apparent from the following detailed description of illustrative embodiments thereof, which is to be read in connection with the accompanying drawings.

BRIEF DESCRIPTION OF DRAWINGS

[0010] The disclosure will provide details in the following description of preferred embodiments with reference to the following figures wherein:

[0011] FIG. 1 is a block/flow diagram illustrating a system/method for generating a set of log patterns, in accordance with the present principles;

[0012] FIG. 2 is a block/flow diagram illustrating a system/method for implementing an incremental learning process to generate a set of log patterns, in accordance with the present principles.

[0013] FIG. 3 illustrates exemplary pseudocode for implementing an incremental learning process to generate a set of log patterns, in accordance with the present principles.

[0014] FIG. 4 is a block/flow diagram illustrating a system/method for security log analysis, in accordance with the present principles.

[0015] FIG. 5 is a block/flow diagram illustrating a system/method for applying a security information and event management service system, in accordance with the present principles.

[0016] FIG. 6 is a block/flow diagram illustrating a computer system, in accordance with the present principles.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0017] Numerous log messages are generated every day by social networks, banking systems, health care systems, smart cities, Internet of Things (IoT) devices etc. Administrators are interested in analyzing these log messages in order to, for example, predict future events, detect anomalies or debug performance problems. To analyze logs, the logs first need to be parsed understand their content. One traditional way to parse logs is to come up with some regular expressions (regex) manually. This is difficult to do, since logs have different formats, and there is no prior knowledge about how many unique formats exist in the logs. In addition, web servers and IoT devices can generate millions of logs every day. To keep up with this rate, the log parser should be able to handle millions of logs in few hours.

[0018] Heterogeneous log analysis (HLA) is a machine learning approach that can provide for log pattern analysis, and can perform heterogeneous log categorization, automatic log format recognition, and heterogeneous log indexing. Additionally, HLA can support queries and outlier detection on heterogeneous logs. HLA can provide a framework which is purely data-oriented and thus general enough to adapt to arbitrary log formats, applications or systems.

[0019] An illustrative implementation of HLA is Heterogeneous Log Analyzer (HLAer). HLAer generates clusters implementing OPTICS (Ordering Points to Identify the Clustering Structure) from the input logs based two parameters: ε and MinPts. Then, for every cluster, HLAer generates a pattern using an UPGMA (Unweighted Pair Group Method with Arithmetic Mean) algorithm.

[0020] HLA can suffer from scalability and flexibility problems due to operational complexity, making it hard to scale-up to a large training log set. In the particular context of HLAer, for a given dataset having n logs with average length of l , the time complexity of OPTICS is $O(n^2 \cdot l)$, and the time complexity of UPGMA is $O(m \cdot \left(\frac{n}{m}\right)^2 \cdot l^2)$ considering m clusters each having $\frac{n}{m}$ logs. In addition, OPTICS requires $O(n^2)$ memory space. For a real dataset, there can be 100,000,000 logs with the average length of 100 fields. Accordingly, considering these numbers, HLAer particularly suffers from a severe scalability problem.

[0021] One manner of dealing with machine learning on a large data set is by using a scale-out Big-Data platform such as Apache™ Hadoop®. In Hadoop®, a MapReduce programming model is implemented so that multiple replicas of the same machine learning processes are executed concurrently on a cluster of machines, in which all the

data set will be loaded and analyzed in the memory of multiple machines. This is a brute-force approach and requires the availability of abundant computing and networking resources.

[0022] The embodiments described herein address the above-noted scalability problem by implementing an incremental learning approach. When abundant computing and networking resources (e.g., a large cluster of networked servers) are not available for the log pattern learning task, the incremental learning approach according to the present principles takes a divide-and-conquer approach, and integrates learning and log filtering to overcome the above-noted scalability issue. This divide-and-conquer approach will continue until the procedure has processed all of the training logs.

[0023] The incremental learning scheme provided herein solves the aforementioned problems by applying and integrating advanced machine learning techniques with recursive algorithm design principles to enable automated log analysis with controllable resource requirements. The incremental learning approach recursively learns a subset of log patterns based on a log sample, and uses the learned patterns to filter out logs before proceeding to the next iteration. Multiple iterations are executed sequentially until all log patterns are discovered.

[0024] The embodiments described herein can significantly reduce the computing resource cost of learning log patterns within a vast amount of heterogeneous log data from various system components and/or applications, even when prior knowledge about the system might not be available. By integrating advanced machine learning techniques with recursive algorithm design principles, the incremental learning approach described herein constructs an automatic sequential pattern recognition method for heterogeneous

logs in a principled way that allows for faster operation upon system updates. Accordingly, the incremental learning mechanism described herein improves the performance and operation of a computer system.

[0025] Although the embodiments described herein will be illustratively applied within the context of HLA (e.g., HLAer), the scalability of other similar techniques can be improved in accordance with the embodiments described herein.

[0026] Referring now in detail to the figures in which like numerals represent the same or similar elements and initially to FIG. 1, a system/method 100 for incremental learning on log pattern recognition is illustratively depicted in accordance with one embodiment of the present principles.

[0027] At block 110, logs are obtained from one or more sources to generate a training set for log pattern learning. Each log can include a time stamp and text content with one or more fields. In one embodiment, the logs include heterogeneous logs. The one or more sources can include arbitrary and/or unknown systems or applications.

[0028] At block 120, an incremental learning process is implemented to generate a set of log patterns from the training set. Each pattern of the set of log patterns can be in terms of regular expression (regex). Regex is a sequence of characters that defines a search pattern. Such a regex pattern can be used by a search algorithm (e.g., string searching algorithm) for performing one or more operations on strings (e.g., find operation). Further details regarding the incremental learning process performed at block 120 will now be described with reference to FIG. 2.

[0029] Referring now to FIG. 2, a system/method 200 for implementing an incremental learning process to generate a set of log patterns is illustratively depicted in accordance with one embodiment of the present principles.

[0030] At block 210, sets L, P and U are defined. L is defined as a training set of logs, such as the training set output at block 110 of FIG. 1. P is a set of log patterns that have been generated so far. U is defined as a set of logs of the training set that have no matching patterns in P. Accordingly, initially, P is defined as an empty set, and U is equivalent to L.

[0031] At block 220, a set of logs S is generated by sampling U. For example, S can be generated from U by sequential sampling. In one embodiment, S has size K, where K is a parameter that controls a maximum resource requirement for the incremental learning process (e.g., K=5000). K can be a user-specified value. Once S reaches K number of logs, the sampling procedure stops. If the size of U is less than K, then S is equivalent to U.

[0032] At block 230, automatic log pattern recognition is performed using S to generate a set of log patterns N. In one embodiment, performing the automatic log pattern recognition includes applying one or more machine learning techniques on set S. An example of such a machine learning technique is HLA (e.g., HLAer).

[0033] The following is an example of one possible automatic log pattern recognition procedure that can be performed at block 230 to generate N.

[0034] The logs of S are tokenized to generate semantically meaningful tokens from the logs. Any tokenization procedure suitable for generating tokens from the logs can be used in accordance with the embodiments described herein. After the logs are tokenized,

a similarity measurement on the logs is applied. This similarity measurement leverages both log layout information and log content information, and is specially tailored to arbitrary logs. Once the similarities among the logs are captured, a hierarchical clustering method can be applied to generate and output a log cluster hierarchy. Any hierarchical clustering method can be implemented in accordance with the embodiments described herein. Once the log cluster hierarchies are received, the logs are aligned within each cluster that is on the lowest level in the log cluster hierarchy. The log alignment is designed to preserve the unknown layouts of the logs in order to facilitate log pattern recognition.

[0035] Once the logs are aligned, log motif discovery is conducted to find log motifs (e.g., the most representative layouts and log fields). Pattern recognition is then performed from such motifs. To do this, fields including time stamps, Internet Protocol (IP) addresses and universal resource locators (URLs) are recognized. Next, other fields which are highly conserved in the logs are recognized and organized in a data structure in the cluster hierarchy. Note that the log motif discovery and pattern recognition are initially performed in the lowest level of the log cluster hierarchy. All of the information is then backpropagated up to higher levels in the log cluster hierarchy, and fused with their local patterns to provide format pattern representations from the logs.

[0036] Blocks 240-290 collectively represent a log filtering process performed based on the set of log patterns generated at block 230 (i.e., N).

[0037] At block 240, it is determined if N is empty.

[0038] If N is determined to be empty, U is updated so that $U = U - S$ at block 250.

[0039] If N is not determined to be empty at block 240, then P is updated so that $P = P + N$ at block 260. That is, the log patterns of N are added to P.

[0040] At block 270, logs in U that have no matching patterns in P are identified, and U is updated with the identified logs. In one embodiment, a log parser is used to identify the logs. An example of a log parser that can be used to identify the logs is Logstash®.

[0041] After U is updated either at block 250 or at block 270, it is determined if U is empty at block 280. If U is determined not to be empty, then the process reverts back to block 220. However, if U is determined to be empty, P is output at block 290 and the incremental learning process ends.

[0042] Referring now to FIG. 3, exemplary pseudocode 300 for implementing the incremental learning process of FIG. 2 is illustratively depicted in accordance with one embodiment of the present principles.

[0043] As shown, IncrementalLearning inputs are defined, which include sets L, P and U, as described above in FIG. 2. P is returned after the training if U is empty or there are no more patterns that HLA can recognize on U. A set of logs S is generated by sampling U, as described above in FIG. 2. Automatic log pattern recognition is performed using S to generate a set of log patterns N (e.g., by calling HLA to learn N), as described above in FIG. 2. Patterns generated/learned so far are accumulated into P, as described above in FIG. 2. A log parser (e.g., Logstash®) is called to update U with the logs identified from N, as described above in FIG. 2. IncrementalLearning(L, P, U) is returned.

[0044] Referring back to FIG. 1, at block 130, the logs are parsed using the set of log patterns, and a set of applications are applied to the parsed logs. The set of applications can include one or more management applications. For example, a log-based fault

management application can be applied to detect abnormal logs that do not match any log pattern of the set of log patterns.

[0045] If the logs being analyzed are security logs or alerts, security log analysis can be performed to provide computer system security. One implementation of security log analysis is provided by security information and event management (SIEM) software products and services, which combine security information management (SIM) and security event management (SEM). These products and services provide real-time analysis of security logs generated by network hardware and applications. One challenge regarding SIEM log analytics is the scalability of log parsing. For example, during a Distributed Denial-of-Service (DDoS) attack, in which multiple computer systems flood the bandwidth and/or resources of a target computer system with traffic, the number of logs generated by the log-in process could be huge. The system administrator of the target computer system needs to quickly determine the origins of the DDoS attack by analyzing the logs. Therefore, it is imperative to produce log patterns (e.g., regex) in a timely manner to aid in the analysis. On the other hand, the large scale of logs may slow down the log pattern generation procedure if the resources are not managed properly. Thus, the incremental learning procedure in accordance with the embodiments described herein can handle the aforementioned scalability problem without sacrificing the accuracy and the quality of the log patterns.

[0046] Referring to FIG. 4, a system/method 400 for implementing a security log analysis service is illustratively depicted in accordance with one embodiment of the present principles.

[0047] In this illustrative embodiment, the system/method 400 is implemented utilizing a SIEM service that includes SEM 410 and SIM 420. SEM 410 can illustratively include one or more firewalls 412, one or more proxy servers 414, one or more virtual private networks (VPNs) 416, and one or more perimeter devices 418. SIM 420 can illustratively include one or more desktops 422, one or more servers 424, one or more routers 426, one or more switches 428, one or more applications 430 and one or more syslog devices 432.

[0048] Security logs from SEM 410 and SIM 420 are analyzed by Security Log Analysis (SLA) 440.

[0049] At block 442, the security logs are collected from SEM 410 and SIM 420. A vast number of log messages can be generated every day by SIEM network hardware and applications.

[0050] At block 444, the collected security logs are parsed as structured data. One traditional way to parse the logs is to come up with some regular expressions (regex) manually. This can be hard to do because logs can have different formats, and there may be no prior knowledge about how many unique formats exist in the logs. Thus, an automatic and scalable log pattern generation approach is highly desired for parsing the security logs, such as the approach described herein above with reference to FIGs. 1-3.

[0051] At block 446, the parsed security logs are archived and, at block 448, are further analyzed for different security applications. Examples of such security applications include intrusion detection and forensic analysis.

[0052] Referring to FIG. 5, a system/method 500 for implementing a security management service 510 is illustratively depicted in accordance with one embodiment of

the present principles. In one embodiment, the security management service 510 utilizes SIEM, as described above with reference to FIG. 4. The security management service 510 can provide security against a multitude of potential computer system security threats, including but not limited to malicious e-mail 520, fraud 530, data theft 540 and compromised websites 550. In addition, the security management service 510 can provide security for a multitude of applications, including but not limited to IT systems 560 (e.g., data centers), cyber-physical systems 570 (e.g., power plant control systems) and social infrastructures 580 (e.g., smart cities).

[0053] Embodiments described herein may be entirely hardware, entirely software or including both hardware and software elements. In a preferred embodiment, aspects of the present invention are implemented in software, which includes but is not limited to firmware, resident software, microcode, etc.

[0054] Embodiments may include a computer program product accessible from a computer-usable or computer-readable medium providing program code for use by or in connection with a computer or any instruction execution system. A computer-usable or computer readable medium may include any apparatus that stores, communicates, propagates, or transports the program for use by or in connection with the instruction execution system, apparatus, or device. The medium can be magnetic, optical, electronic, electromagnetic, infrared, or semiconductor system (or apparatus or device) or a propagation medium. The medium may include a computer-readable storage medium such as a semiconductor or solid state memory, magnetic tape, a removable computer diskette, a random access memory (RAM), a read-only memory (ROM), a rigid magnetic disk and an optical disk, etc.

[0055] Each computer program may be tangibly stored in a machine-readable storage media or device (e.g., program memory or magnetic disk) readable by a general or special purpose programmable computer, for configuring and controlling operation of a computer when the storage media or device is read by the computer to perform the procedures described herein. The inventive system may also be considered to be embodied in a computer-readable storage medium, configured with a computer program, where the storage medium so configured causes a computer to operate in a specific and predefined manner to perform the functions described herein.

[0056] A data processing system suitable for storing and/or executing program code may include at least one processor coupled directly or indirectly to memory elements through a system bus. The memory elements can include local memory employed during actual execution of the program code, bulk storage, and cache memories which provide temporary storage of at least some program code to reduce the number of times code is retrieved from bulk storage during execution. Input/output or I/O devices (including but not limited to keyboards, displays, pointing devices, etc.) may be coupled to the system either directly or through intervening I/O controllers.

[0057] Network adapters may also be coupled to the system to enable the data processing system to become coupled to other data processing systems or remote printers or storage devices through intervening private or public networks. Modems, cable modem and Ethernet cards are just a few of the currently available types of network adapters.

[0058] Referring now to FIG. 6, an exemplary computer system 600 is shown which may represent a server or a network device, in accordance with an embodiment of the

present invention. The computer system 600 includes at least one processor (CPU) 605 operatively coupled to other components via a system bus 602. A cache 606, a Read Only Memory (ROM) 608, a Random-Access Memory (RAM) 610, an input/output (I/O) adapter 620, a sound adapter 630, a network adapter 690, a user interface adapter 650, and a display adapter 660, are operatively coupled to the system bus 602.

[0059] A first storage device 622 and a second storage device 629 are operatively coupled to system bus 602 by the I/O adapter 620. The storage devices 622 and 629 can be any of a disk storage device (e.g., a magnetic or optical disk storage device), a solid state magnetic device, and so forth. The storage devices 622 and 629 can be the same type of storage device or different types of storage devices.

[0060] A speaker 632 may be operatively coupled to system bus 602 by the sound adapter 630. A transceiver 695 is operatively coupled to system bus 602 by network adapter 690. A display device 662 is operatively coupled to system bus 602 by display adapter 660.

[0061] A first user input device 652, a second user input device 659, and a third user input device 656 are operatively coupled to system bus 602 by user interface adapter 650. The user input devices 652, 659, and 656 can be any of a sensor, a keyboard, a mouse, a keypad, a joystick, an image capture device, a motion sensing device, a power measurement device, a microphone, a device incorporating the functionality of at least two of the preceding devices, and so forth. Of course, other types of input devices can also be used, while maintaining the spirit of the present invention. The user input devices 652, 659, and 656 can be the same type of user input device or different types of user

input devices. The user input devices 652, 659, and 656 are used to input and output information to and from system 600.

[0062] Of course, the computer system 600 may also include other elements (not shown), as readily contemplated by one of skill in the art, as well as omit certain elements. For example, various other input devices and/or output devices can be included in computer system 600, depending upon the particular implementation of the same, as readily understood by one of ordinary skill in the art. For example, various types of wireless and/or wired input and/or output devices can be used. Moreover, additional processors, controllers, memories, and so forth, in various configurations can also be utilized as readily appreciated by one of ordinary skill in the art. These and other variations of the computer system 600 are readily contemplated by one of ordinary skill in the art given the teachings of the present invention provided herein.

[0063] The foregoing is to be understood as being in every respect illustrative and exemplary, but not restrictive, and the scope of the invention disclosed herein is not to be determined from the Detailed Description, but rather from the claims as interpreted according to the full breadth permitted by the patent laws. It is to be understood that the embodiments shown and described herein are only illustrative of the principles of the present invention and that those skilled in the art may implement various modifications without departing from the scope and spirit of the invention. Those skilled in the art could implement various other feature combinations without departing from the scope and spirit of the invention. Having thus described aspects of the invention, with the details and particularity required by the patent laws, what is claimed and desired protected by Letters Patent is set forth in the appended claims.

WHAT IS CLAIMED IS:

1. A computer-implemented method for enabling automated log analysis with controllable resource requirements, comprising:

generating, by a processor operatively coupled to a memory, a training set for log pattern learning based on heterogeneous logs generated by a computer system;

implementing, by the processor, an incremental learning process to generate a set of log patterns from the training set, wherein implementing the incremental learning process comprises:

defining a first set as the training set, a second set as a set of log patterns that have been generated, and a third set as a set of logs of the training set that lack a matching pattern in the second set;

sampling the third set to generate a fourth set having a size corresponding to a parameter controlling a maximum resource requirement for the incremental learning process; and

performing automatic log pattern recognition to generate a fifth set;
parsing, by the processor, the heterogeneous logs using the set of log patterns; and
applying, by the processor, a set of applications to the parsed logs.

2. The method as recited in claim 1, further comprising obtaining the heterogeneous logs from the computer system.

3. The method as recited in claim 1, wherein performing the automatic log pattern recognition further comprises:
- tokenizing logs of the fourth set to generate tokens from the logs of the fourth set;
 - applying a similarity measurement on the logs to capture similarities among the logs of the fourth set;
 - implementing a hierarchical clustering algorithm to generate a log cluster hierarchy for the logs of the fourth set;
 - aligning the logs of the fourth set within each cluster associated with a given level of the log cluster hierarchy;
 - conducting log motif discovery on the aligned logs to find log motifs; and
 - performing pattern recognition from the log motifs by recognizing one or more log fields.
4. The method as recited in claim 3, wherein the one or more log fields include one or more of time stamps, Internet Protocol (IP) addresses and universal resource locators (URLs).
5. The method as recited in claim 3, further comprising backpropagating, by the processor, information to higher levels of the log cluster hierarchy.
6. The method as recited in claim 1, further comprising performing, by the processor, a log filtering process based on the fifth set.

7. The method as recited in claim 6, wherein performing the log filter process comprises determining if the fifth set is empty.

8. The method as recited in claim 7, wherein, in response to determining that the fifth set is empty, performing the log filtering process further comprises:

updating the third set by subtracting the fourth set from the third set;

in response to determining that the updated third set is empty, outputting the second set; and

in response to determining that the updated third set includes at least one log, updating the fourth set by sampling the updated third set, and updating the fifth set by performing automatic pattern recognition using the updated fourth set.

9. The method as recited in claim 7, wherein, in response to determining that the fifth set includes at least one log pattern, performing the log filtering process further comprises:

updating the second set by adding the fifth set to the second set;

updating the third set with logs identified in the third set that lack a matching pattern in the second set;

in response to determining that the updated third set is empty, outputting the updated second set; and

in response to determining that the updated third set includes at least one log, updating the fourth set by sampling the updated third set, and updating the fifth set by performing automatic pattern recognition using the updated fourth set.

10. The method as recited in claim 9, further comprising using a log parser to identify the logs in the third set that lack a matching pattern in the second set.
11. The method as recited in claim 1, wherein the set of applications include a log-based fault management application to detect abnormal logs that do not match any log pattern of the set of the log patterns.
12. The method as recited in claim 1, wherein each pattern of the set of log patterns is in terms of regular expression.
13. A computer program product comprising a non-transitory computer readable storage medium having program instructions embodied therewith, the program instructions executable by a computer to cause the computer to perform a method for enabling automated log analysis with controllable resource requirements comprising:
- generating a training set for log pattern learning based on heterogeneous logs generated by a computer system;
 - implementing an incremental learning process to generate a set of log patterns from the training set, wherein implementing the incremental learning process comprises:
 - defining a first set as the training set, a second set as a set of log patterns that have been generated, and a third set as a set of logs of the training set that lack a matching pattern in the second set;

sampling the third set to generate a fourth set having a size corresponding to a parameter controlling a maximum resource requirement for the incremental learning process; and

performing automatic log pattern recognition to generate a fifth set;
parsing the heterogeneous logs using the set of log patterns; and
applying a set of applications to the parsed logs.

14. The computer program product as recited in claim 13, wherein performing the automatic log pattern recognition further comprises:

tokenizing logs of the fourth set to generate tokens from the logs of the fourth set;
applying a similarity measurement on the logs to capture similarities among the logs of the fourth set;

implementing a hierarchical clustering algorithm to generate a log cluster hierarchy for the logs of the fourth set;

aligning the logs of the fourth set within each cluster associated with a given level of the log cluster hierarchy;

conducting log motif discovery on the aligned logs to find log motifs; and
performing pattern recognition from the log motifs by recognizing one or more log fields.

15. The computer program product of claim 14, further comprising performing a log filtering process based on the fifth set.

16. The computer program product of claim 15, wherein performing the log filtering process comprises determining if the fifth set is empty.

17. The computer program product of claim 16, wherein, in response to determining that the fifth set is empty, performing the log filtering process further comprises:

updating the third set by subtracting the fourth set from the third set;

in response to determining that the updated third set is empty, outputting the second set; and

in response to determining that the updated third set includes at least one log, updating the fourth set by sampling the updated third set, and updating the fifth set by performing automatic pattern recognition using the updated fourth set.

18. The computer program product of claim 16, wherein, in response to determining that the fifth set includes at least one log pattern, performing the log filtering process further comprises:

updating the second set by adding the fifth set to the second set;

updating the third set with logs identified in the third set that lack a matching pattern in the second set;

in response to determining that the updated third set is empty, outputting the updated second set; and

in response to determining that the updated third set includes at least one log, updating the fourth set by sampling the updated third set, and updating the fifth set by performing automatic pattern recognition using the updated fourth set.

19. The computer program product as recited in claim 18, further comprising using a log parser to identify the logs in the third set that lack a matching pattern in the second set.

20. A system for enabling automated log analysis with controllable resource requirements, comprising:

a computer system configured to generate heterogeneous logs; and

at least one processor operatively coupled to a memory, wherein the at least one processor is configured to:

obtain the heterogeneous logs from the computer system, and generate a training set for log pattern learning based on the obtained logs;

implement an incremental learning process to generate a set of log patterns from the training set by:

defining a first set as the training set, a second set as a set of log patterns that have been generated, and a third set as a set of logs of the training set that lack a matching pattern in the second set;

sampling the third set to generate a fourth set having a size corresponding to a parameter controlling a maximum resource requirement for the incremental learning process; and

performing automatic log pattern recognition to generate a fifth set;

parse the heterogeneous logs using the set of log patterns; and
apply a set of applications to the parsed logs.

1/6
16091

100

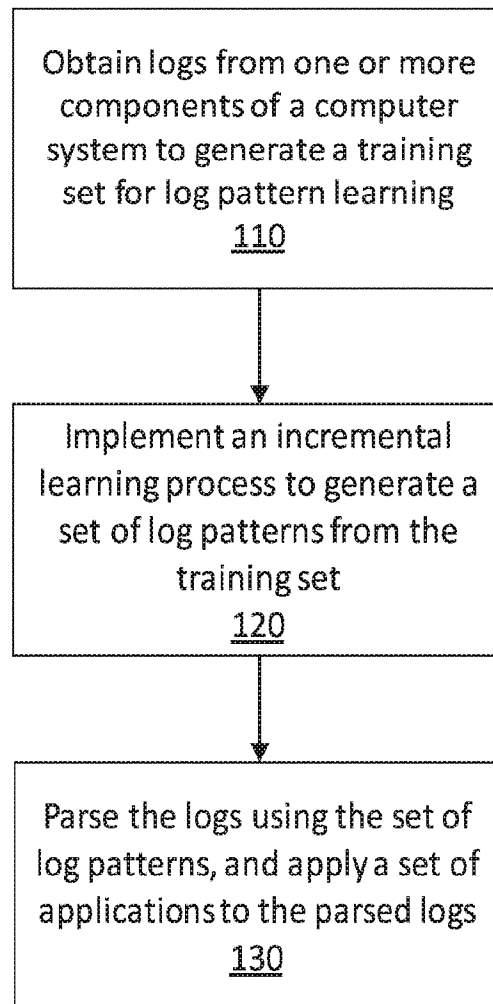


FIG. 1

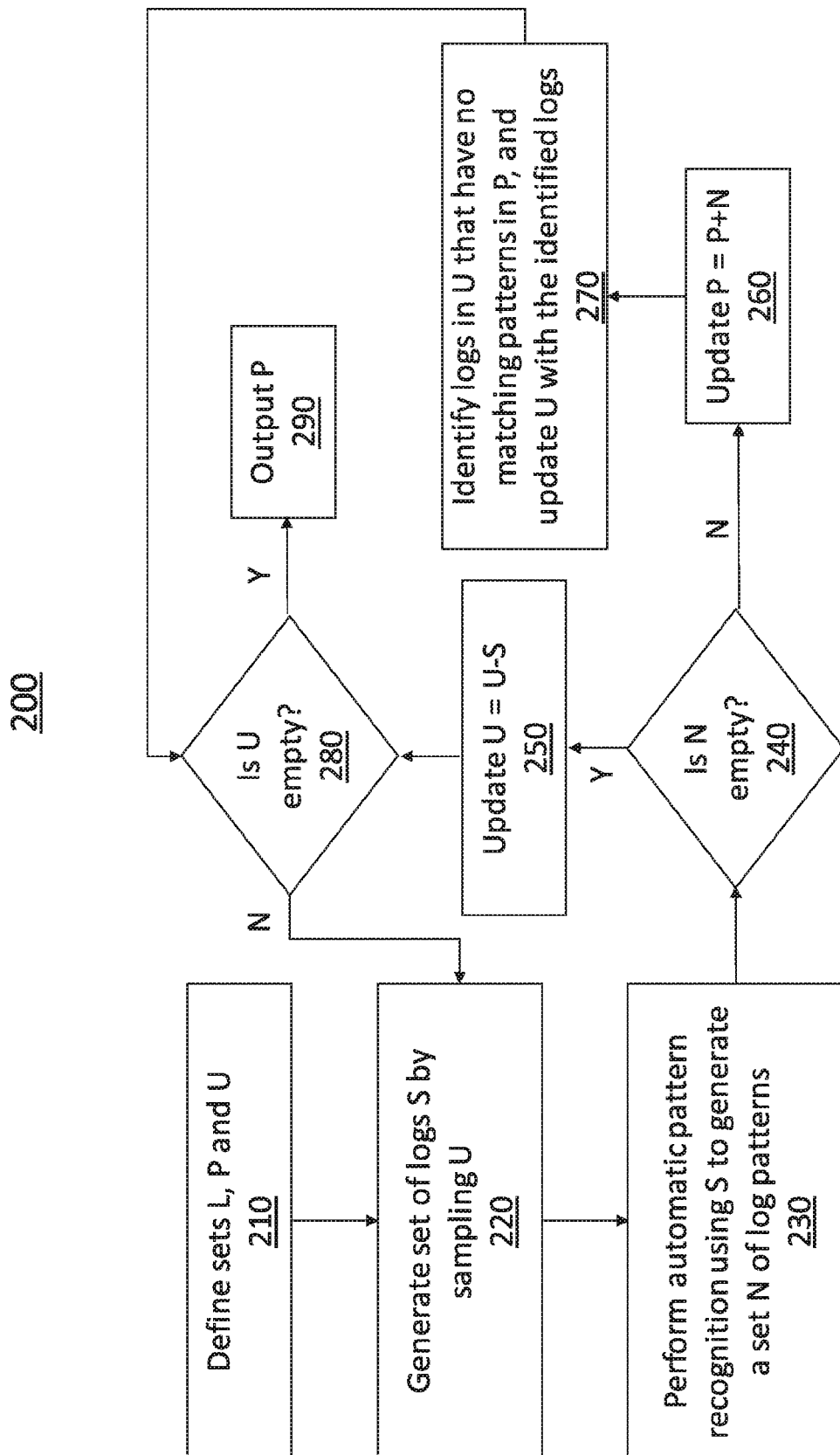
2/6
16091

FIG. 2

3/6
16091

300

```

IncrementalLearning(training logs L, known ptnn patterns P, logs with unknown patterns U)
    L ← the total training log set;
    P ← set of ptnn files generated so far, initially set as empty;
    U ← logs that have no matching patterns in the set P;

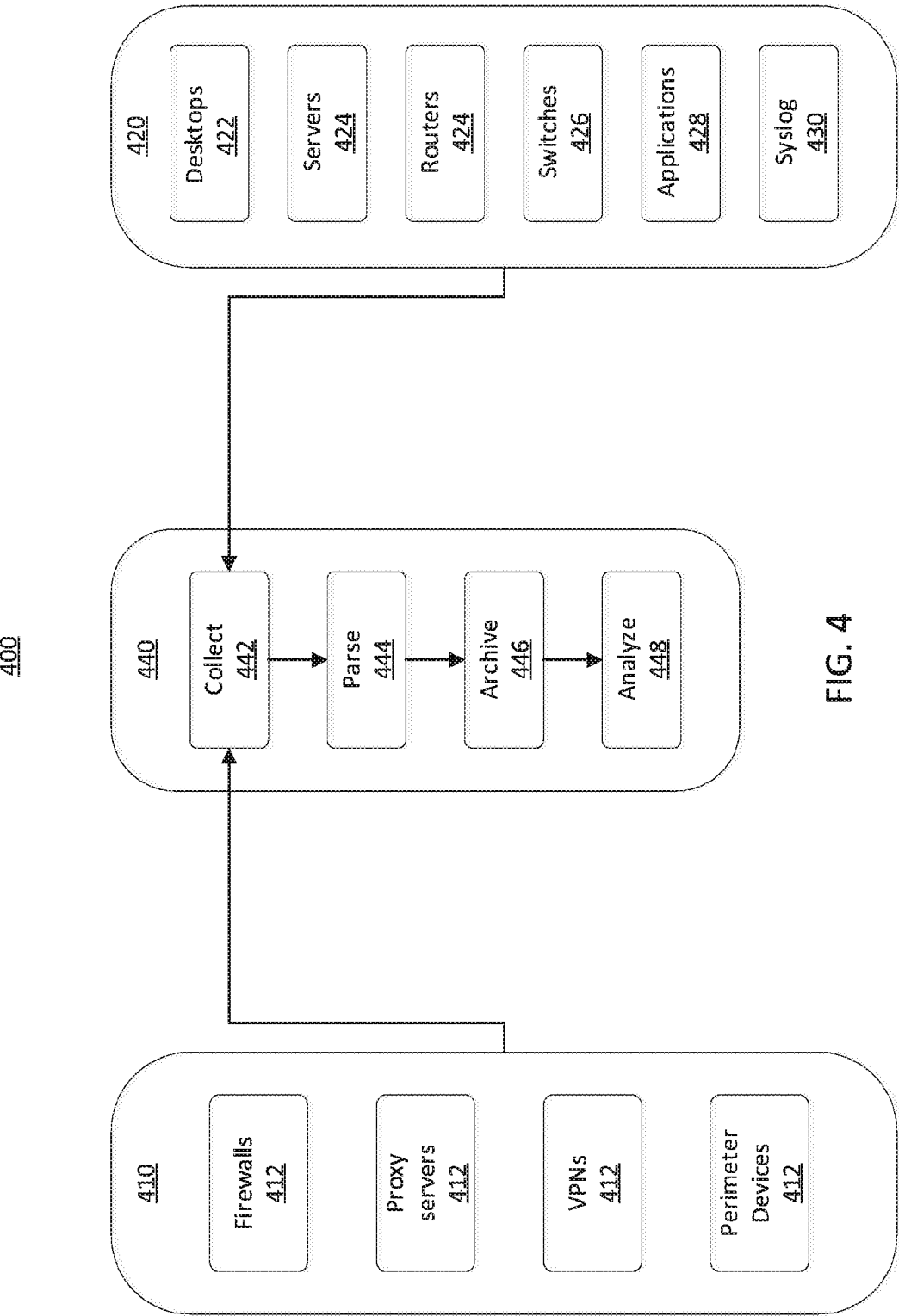
    {
        if(U is empty || no more patterns HLA can recognize on U)
            return buildIndex(P); //return the final grok pattern set after training
        Sampling(U) -> S; //generate a log set sampled from U
        HLA(S) -> N; //call HLA to learn the new log pattern set N
        P += N; //accumulate the ptnn patterns learnt so far
        Logstash(N, U) -> U; //call Logstash to update the log set with
                               //unknown patterns

        return IncrementalLearning(L, P, U);
    }

```

FIG. 3

4/6
16091



5/6
16091

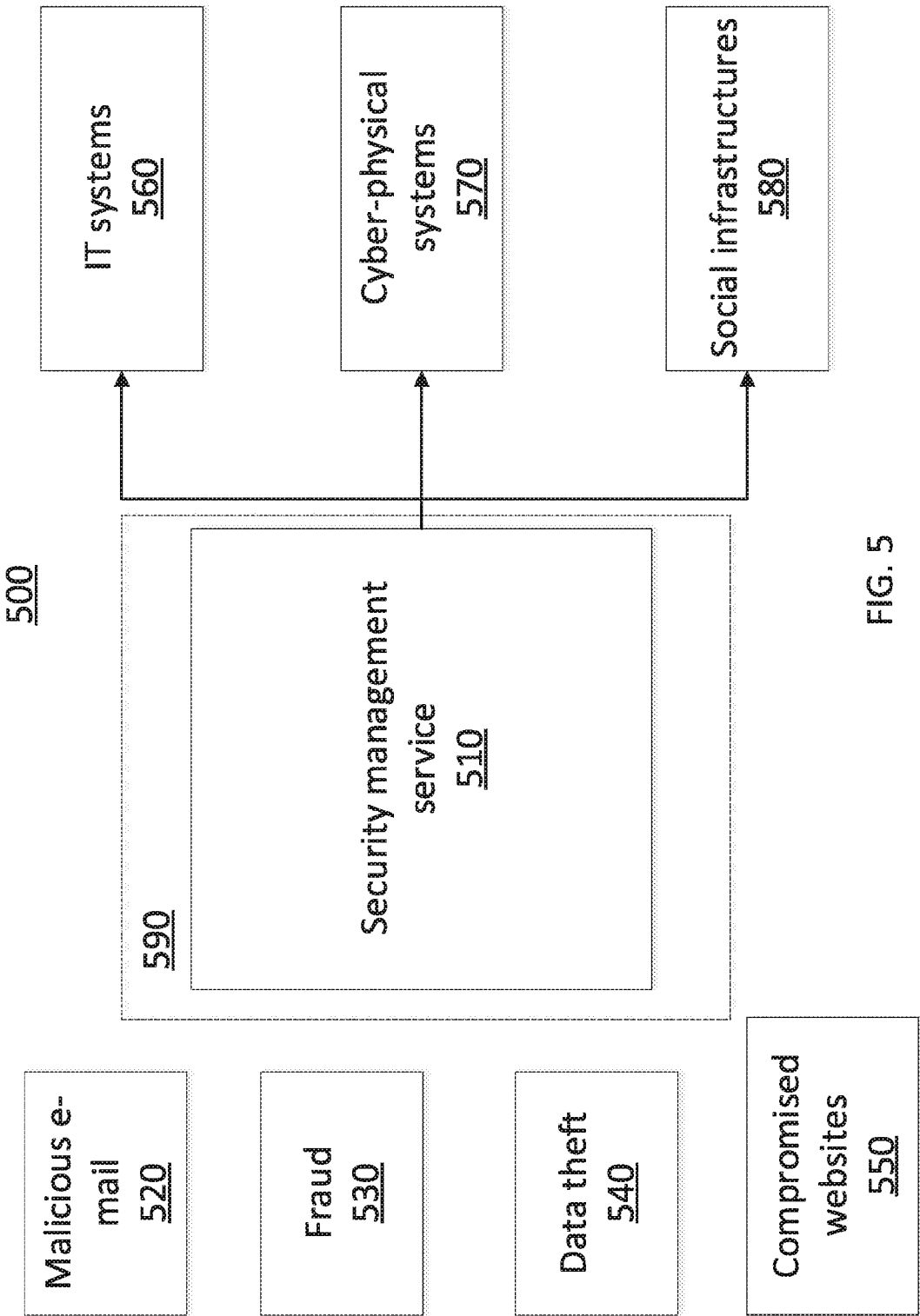


FIG. 5

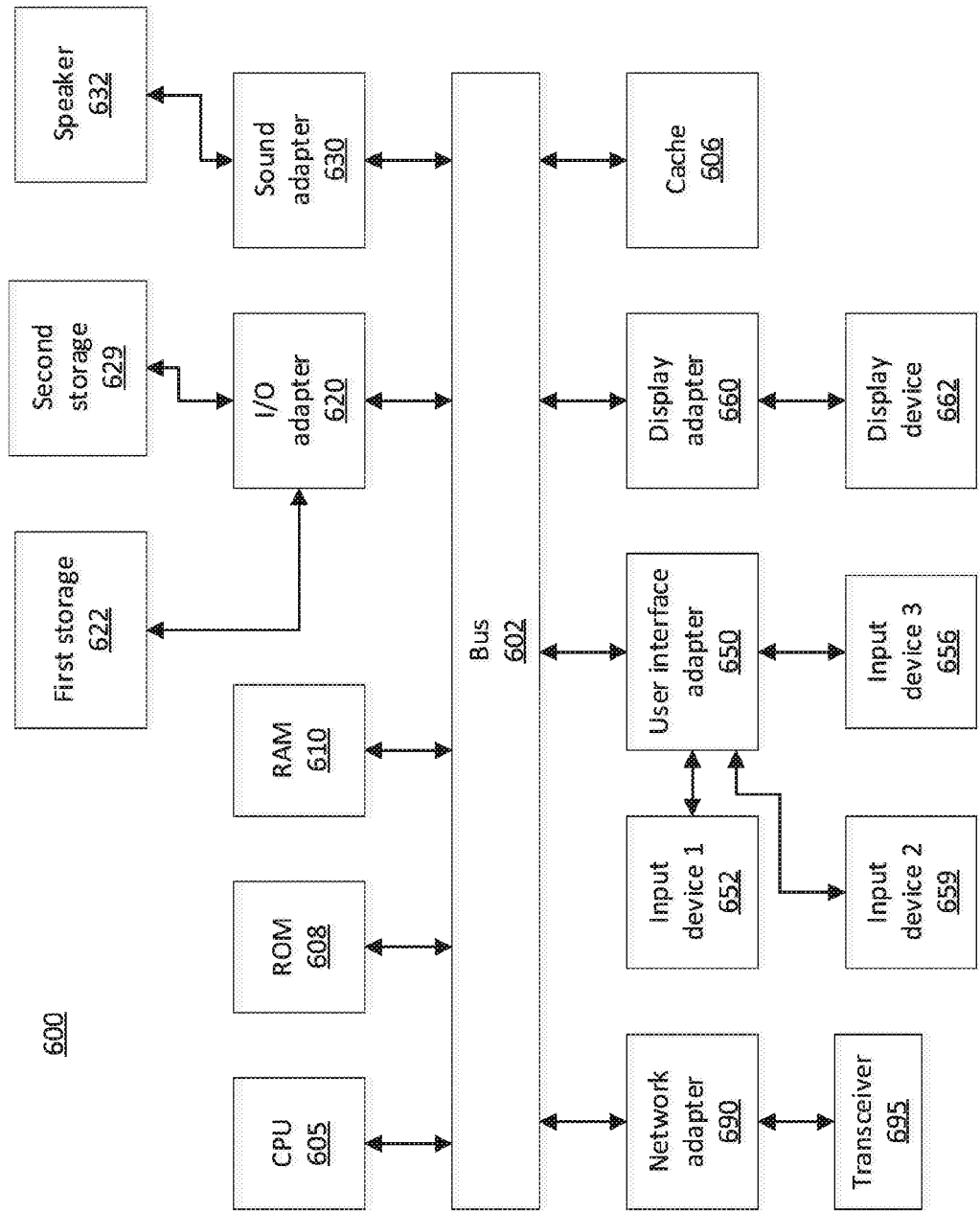


FIG. 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2018/018332**A. CLASSIFICATION OF SUBJECT MATTER****G06N 5/04(2006.01)i, G06N 99/00(2010.01)i, G06F 21/55(2013.01)i**

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)

G06N 5/04; G06F 11/07; G06F 17/00; G06F 11/28; G06F 17/30; G06F 155/00; G06F 15/18; G06N 99/00; G06F 21/55

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

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & Keywords: incremental learning, heterogeneous logs, training set, pattern, matching, size, maximum resource, recognition, parsing

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2011-0131156 A1 (KATHLEEN FISHER et al.) 02 June 2011 See paragraphs [0028]-[0045]; claim 1; and figures 1-9.	1-20
A	US 2014-0006861 A1 (NAVENDU JAIN et al.) 02 January 2014 See paragraphs [0018]-[0039]; claim 1; and figures 1-4.	1-20
A	US 2003-0050719 A1 (LU BAO-LIANG et al.) 13 March 2003 See paragraphs [0044]-[0062]; claims 1-3; and figures 1-2.	1-20
A	US 2011-0296244 A1 (QIANG FU et al.) 01 December 2011 See paragraphs [0026]-[0076]; claim 1; and figures 1-7.	1-20
A	US 2014-0324865 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 30 October 2014 See paragraphs [0041]-[0053]; claim 1; and figures 1-2.	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

04 June 2018 (04.06.2018)

Date of mailing of the international search report

04 June 2018 (04.06.2018)

Name and mailing address of the ISA/KR

International Application Division

Korean Intellectual Property Office

189 Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

NHO, Ji Myong

Telephone No. +82-42-481-8528



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2018/018332

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2011-0131156 A1	02/06/2011	US 2014-0149323 A1 US 8341096 B2 US 9208448 B2	29/05/2014 25/12/2012 08/12/2015
US 2014-0006861 A1	02/01/2014	US 9229800 B2	05/01/2016
US 2003-0050719 A1	13/03/2003	JP 2003-030656 A	31/01/2003
US 2011-0296244 A1	01/12/2011	US 8495429 B2	23/07/2013
US 2014-0324865 A1	30/10/2014	JP 05946423 B2 JP 2014-215883 A	06/07/2016 17/11/2014