



(51) International Patent Classification:

G06F 11/36 (2006.01) G06F 16/901 (2019.01)

G06F 16/36 (2019.01)

(21) International Application Number:

PCT/US2021/021063

(22) International Filing Date:

05 March 2021 (05.03.2021)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/987,928 11 March 2020 (11.03.2020) US

17/191,964 04 March 2021 (04.03.2021) US

(71) Applicant: NEC LABORATORIES AMERICA, INC.

[US/US]; 4 Independence Way, Suite 200, Princeton, NJ 08540 (US).

(72) Inventors: ZONG, Bo; 28 Hawthorne Drive, West Wind-

sor, NJ 08550 (US). CHEN, Haifeng; 11 Blackhawk Court, West Windsor, NJ 08550 (US). ZHANG, Xunchao; 7300 Summit Rock Road, Elkridge, MD 21075 (US).

(74) Agent: BITETTO, James, J.; Tutunjian & Bitetto, P.C.,

401 Broadhollow Road, Suite 402, Melville, NY 11747 (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, IT, 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: GRAPH-BASED METHOD FOR INDUCTIVE BUG LOCALIZATION

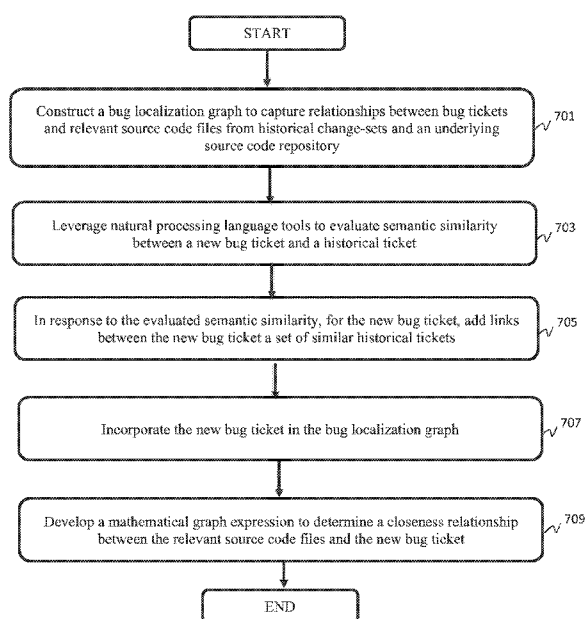


FIG. 7

(57) Abstract: A computer-implemented method executed by at least one processor for software bug localization is presented. The method includes constructing (701) a bug localization graph to capture relationships between bug tickets and relevant source code files from historical change-sets and an underlying source code repository, leveraging (703) natural processing language tools to evaluate semantic similarity between a new bug ticket and a historical ticket, in response to the evaluated semantic similarity, for the new bug ticket, adding (705) links between the new bug ticket a set of similar historical tickets, incorporating (707) the new bug ticket in the bug localization graph, and developing (709) a mathematical graph expression to determine a closeness relationship between the relevant source code files and the new bug ticket.

Published:

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

GRAPH-BASED METHOD FOR INDUCTIVE BUG LOCALIZATION

RELATED APPLICATION INFORMATION

[0001] This application claims priority to Provisional Application No. 62/987,928, filed on March 11, 2020, and U.S. Patent Application No. 17/191,964, filed on March 4, 2021, each incorporated herein by reference herein in their entirety.

BACKGROUND

Technical Field

[0002] The present invention relates to software development and, more particularly, to a graph-based method for inductive bug localization.

Description of the Related Art

[0003] A software bug is generally an error in software which may cause a program to stop its execution or to execute improperly. The bug may be introduced into the program due to a flaw in the software design or unintentionally by the program's developer while writing the program or code. If the bug is not located and corrected prior to the program being released to a user, or possibly installed in a user device, the program or device may malfunction with potentially negative repercussions to the user.

[0004] Software testing is generally conducted during a program's development stage to identify bugs. The test may check all the areas in the program which, depending on the size of the program, may be costly and time consuming. Alternatively, in order to reduce the cost and duration of the test, areas thought to have bugs, hereinafter referred to as bug-prone areas, may be selected for testing.

[0005] In some cases, testing may be performed manually, semi-automatically, automatically, or the like. The testing efforts may be focused on the entire program, on bug-prone areas, on specific aspects of the program, or the like.

SUMMARY

[0006] A computer-implemented method executed by at least one processor for software bug localization is presented. The method includes constructing a bug localization graph to capture relationships between bug tickets and relevant source code files from historical change-sets and an underlying source code repository, leveraging natural processing language tools to evaluate semantic similarity between a new bug ticket and a historical ticket, in response to the evaluated semantic similarity, for the new bug ticket, adding links between the new bug ticket a set of similar historical tickets, incorporating the new bug ticket in the bug localization graph, and developing a mathematical graph expression to determine a closeness relationship between the relevant source code files and the new bug ticket.

[0007] A system for software bug localization is also presented. The system includes a memory and a processor in communication with the memory, wherein the processor is configured to construct a bug localization graph to capture relationships between bug tickets and relevant source code files from historical change-sets and an underlying source code repository, leverage natural processing language tools to evaluate semantic similarity between a new bug ticket and a historical ticket, in response to the evaluated semantic similarity, for the new bug ticket, add links between the new bug ticket a set of similar historical tickets, incorporate the new bug ticket in the bug localization graph, and develop a mathematical graph expression to determine a closeness relationship between the relevant source code files and the new bug ticket.

[0008] A non-transitory computer-readable storage medium comprising a computer-readable program is presented for software bug localization, wherein the computer-readable program when executed on a computer causes the computer to perform the steps of constructing a bug localization graph to capture relationships between bug tickets and relevant source code files from historical change-sets and an underlying source code repository, leveraging natural processing language tools to evaluate semantic similarity between a new bug ticket and a historical ticket, in response to the evaluated semantic similarity, for the new bug ticket, adding links between the new bug ticket a set of similar historical tickets, incorporating the new bug ticket in the bug localization graph, and developing a mathematical graph expression to determine a closeness relationship between the relevant source code files and the new bug ticket.

[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 an exemplary procedure for graph-based bug localization (GBL), in accordance with embodiments of the present invention;

[0012] FIG. 2 is a block/flow diagram illustrating an exemplary procedure of bug localization graph construction, in accordance with embodiments of the present invention;

[0013] FIG. 3 is a block/flow diagram illustrating an exemplary procedure of connecting a new ticket with its historical similar tickets, in accordance with embodiments of the present invention;

[0014] FIG. 4 is a block/flow diagram illustrating an exemplary procedure of evaluating closeness between source code files and the new ticket, in accordance with embodiments of the present invention;

[0015] FIG. 5 is a block/flow diagram of practical applications for the graph-based bug localization (GBL), in accordance with embodiments of the present invention;

[0016] FIG. 6 is an exemplary processing system for GBL, in accordance with embodiments of the present invention;

[0017] FIG. 7 is a block/flow diagram of a method for GBL, in accordance with embodiments of the present invention; and

[0018] FIG. 8 are closeness score equations for GBL, in accordance with embodiments of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0019] In software development, bug localization is the process of finding portions of source code associated to a submitted bug report. This task has been modeled as an information retrieval task at the source code file, where the report is a query.

[0020] The exemplary embodiments deal with the problem of bug localization in software engineering. Given a source code repository and a bug ticket that describes an encountered problem in natural language, the goal is to locate source code files (e.g., from the code repository) that are related to the problem stated in the bug ticket.

[0021] The exemplary embodiments focus on the following inductive setting.

[0022] Historical bug tickets and the change-sets (e.g., relevant source code files) for individual bug tickets are available.

[0023] The goal is to provide bug localization for new incoming bug tickets.

[0024] The exemplary embodiments introduce Graph-based Bug Localization (GBL) to address the bug localization problem under the setting stated above.

[0025] The exemplary embodiments solve the problem as follows:

[0026] First, the exemplary methods construct a bug localization graph to capture the relationships among bug tickets and relevant source code files from historical change-sets and the underlying source code repository.

[0027] Then, the exemplary methods leverage natural language processing tools to evaluate semantic similarity between a new bug ticket and a historical bug ticket.

[0028] After that, given the similarity evaluated, for a new bug ticket, the exemplary methods add links between the new ticket and its top similar historical bug tickets so that the new ticket is also included in the bug localization graph.

[0029] Finally, the exemplary methods develop a graph algorithm to discover or determine top close source code files for the new bug ticket.

[0030] The features of GBL include at least the following:

[0031] Bug localization graph construction, which captures the relationships between bug tickets and source code files.

[0032] A natural-language-processing based method to identify similar historical bug tickets so that the exemplary methods can connect a new bug ticket with its similar tickets and bring or incorporate the new ticket into the bug localization graph.

[0033] A graph algorithm for evaluating the closeness between new tickets and source code files.

[0034] FIG. 1 is a block/flow diagram 100 illustrating an exemplary procedure for graph-based bug localization (GBL), in accordance with embodiments of the present invention.

[0035] Construction of bug localization graphs from historical data includes blocks 101-103 and bug source code file recommendations for a new bug ticket includes blocks 104-108.

[0036] Regarding block 101, historical data, historical data include three types of data.

[0037] The first type of data is bug tickets. Bug tickets are documents created by software engineers, reporting encountered problems/bugs in a source code repository. In particular, the bugs are described in natural language.

[0038] The second type of data is change-sets. A change-set of a bug ticket refers to a set of source code files where a software engineer made changes to fix the reported bug.

[0039] The third type of data is source code repository. Source code repository includes all source code files under management.

[0040] Regarding block 102, bug localization graph construction, the exemplary embodiments construct a bug localization graph from change-sets and the source code repository. In the graph, a node denotes or represents either a source code file or a bug ticket. An edge between two nodes indicates the two nodes are related. If a source code file is in a bug ticket's change-set, they are related. If the implementation in one source code file refers to data types defined in another source code file, the two source code files are related.

[0041] Regarding block 103, storing an established bug localization graph, the exemplary method stores the bug localization graph established in block 102 into a database.

[0042] Regarding block 104, a new bug ticket, when a new bug ticket arrives, its natural-language based description states the reported bug, however, its change-set is unknown.

[0043] Regarding block 105, ticket semantic similarity evaluation, the exemplary methods evaluate semantic similarity between the new ticket and historical bug tickets using natural language processing tools.

[0044] Regarding block 106, connecting the new bug ticket with its similar historical bug tickets, based on the similarity evaluation provided by block 105, the exemplary methods select similar historical bug tickets and build edges between the new ticket and its similar tickets.

[0045] Regarding block 107, evaluating closeness between the new ticket and source code files, the exemplary methods leverage the graph algorithm to evaluate closeness between the ticket and the source code files.

[0046] Regarding block 108, recommending top close source code files, based on the closeness provided by block 107, the exemplary methods make recommendations to end users, where top close source code files are recommended with respect to the new ticket.

[0047] FIG. 2 is a block/flow diagram 200 illustrating an exemplary procedure of bug localization graph construction, in accordance with embodiments of the present invention.

[0048] Regarding block 201, extracting reference and change-set relations, the exemplary methods extract reference and change-set relations from a source code repository and historical change-set data.

[0049] With respect to reference relations, across different programming languages, in one source code file, the following data types are usually implemented: classes, structures, functions, and so on. When a data type in one source code file is referred to or used in another source code file, it is said that these two source code files have reference relations. Given a source code repository, the exemplary methods extract possible reference relations between source code files.

[0050] With respect to change-set relations, if a source code file is in the change-set of a bug ticket, the exemplary methods indicate that there is a change-set relation between the source code file and the bug ticket.

[0051] Regarding block 202, building nodes and edges, given a source code repository and historical change-sets for tickets, a bug localization graph is constructed by including a plurality of nodes and a plurality of edges connecting certain nodes.

[0052] With respect to nodes, single nodes are used to represent source code files or tickets.

[0053] With respect to edges, if two nodes have reference relations, there is an edge between the two nodes. If two nodes have change-set relations, there is an edge between the two nodes.

[0054] The output of block 202 is a bug localization graph that captures relations among tickets and source code files.

[0055] FIG. 3 is a block/flow diagram 300 illustrating an exemplary procedure of connecting a new ticket with its historical similar bug tickets, in accordance with embodiments of the present invention.

[0056] FIG. 3 depicts the detailed procedure of connecting the new ticket with its historical similar bug tickets. For the ease of discussion, it is supposed that $S(t_1, t_2)$ is a function that evaluates semantic similarity between t_1 and t_2 provided by block 105.

[0057] Regarding block 301, semantic similarity evaluation and semantic close relation extraction, given a new ticket t , the exemplary methods leverage the function S to compare t against all the historical bug tickets. The output of the function S is usually a numeric value r , where r is larger if two tickets are more similar. Using a pre-defined threshold θ , the exemplary methods can filter dissimilar tickets.

[0058] If $r > \theta$, two tickets are similar, otherwise, the two tickets are dissimilar.

[0059] Regarding block 302, adding extra edges, for a given new ticket t , the exemplary methods obtain its similar tickets from block 301. If t and t_1 are similar, the exemplary methods add an edge between t and t_1 into the bug localization graph from block 103. If t has m similar historical bug tickets, in total m edges are added into the bug localization graph.

[0060] FIG. 4 is a block/flow diagram 400 illustrating an exemplary procedure of evaluating closeness between source code files and the new ticket, in accordance with embodiments of the present invention.

[0061] Regarding block 401, computing closeness from the new ticket to source code files, the exemplary methods compute the closeness parameters as follows.

[0062] With respect to initialization, let $c[i]$ be the closeness score from the new ticket node to node v_i .

[0063] The exemplary methods initialize $c[i]$ as 0.

[0064] Then let v_0 be the node that denotes the new ticket and assign $c[0] = 1$.

[0065] With respect to the closeness score update, for each node v_i , its closeness score is updated by the following equation:

$$[0066] \quad c^{(k+1)}[i] = \lambda \sum_{v_j \in N_i} \frac{1}{d_j} c^{(k)}[j],$$

[0067] where N_i is a set of one-hop neighbors of v_i in the bug localization graph, d_j is a node degree of neighbor v_j , k is an iteration number, and λ is a pre-defined numerical value between 0 and 1.

[0068] For the new ticket node v_0 , the exemplary methods make additional updates as follows:

$$[0069] \quad c^{(k+1)}[0] = c^{(k+1)}[0] + (1 - \lambda) \sum_{j \neq 0} c^{(k)}[j].$$

[0070] With respect to stop criteria, closeness score updates are performed iteratively. The updates are stopped if a pre-defined or predetermined maximum iteration number is reached.

[0071] Regarding block 402, ranking closeness among source code files, for source code files, the exemplary methods rank them based on their closeness score computed from block 401. If one source code file has a higher closeness score, its rank gets closer to the top of a ranking list.

[0072] FIG. 5 is a block/flow diagram 500 of practical applications for GBL, in accordance with embodiments of the present invention.

[0073] Given source code files 502 and a bug ticket 504, a graph-based bug localization 670 is constructed to capture relationships between source code files 502 and the bug ticket 504. The relationship is based on historical change-sets and the underlying source code repository. Existing natural language processing tools 506 are leveraged to evaluate semantic similarity between a new bug ticket and a historical bug ticket. Links are then added between the new ticket and its top historical bug tickets so that a new ticket (new node) is added to the bug localization graph 510. A graph algorithm providing a closeness score 672 is developed to uncover top close source

code files for the new bug ticket. The bug localization graph 510 can be provided or displayed on a user interface 512 handled by a user 514.

[0074] FIG. 6 is an exemplary processing system for GBL, in accordance with embodiments of the present invention.

[0075] The processing system includes at least one processor (CPU) 604 operatively coupled to other components via a system bus 602. A GPU 605, a cache 606, a Read Only Memory (ROM) 608, a Random Access Memory (RAM) 610, an input/output (I/O) adapter 620, a network adapter 630, a user interface adapter 640, and a display adapter 650, are operatively coupled to the system bus 602. Additionally, graph-based bug localization 670 can be employed to execute the graph algorithm or mathematical graph expression of the closeness function 672.

[0076] A storage device 622 is operatively coupled to system bus 602 by the I/O adapter 620. The storage device 622 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.

[0077] A transceiver 632 is operatively coupled to system bus 602 by network adapter 630.

[0078] User input devices 642 are operatively coupled to system bus 602 by user interface adapter 640. The user input devices 642 can be any of a keyboard, a mouse, a keypad, an image capture device, a motion sensing 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 642 can be the same type of user input device or different types of user input devices. The user input devices 642 are used to input and output information to and from the processing system.

[0079] A display device 652 is operatively coupled to system bus 602 by display adapter 650.

[0080] Of course, the processing system 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 the system, 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 processing system are readily contemplated by one of ordinary skill in the art given the teachings of the present invention provided herein.

[0081] FIG. 7 is a block/flow diagram of a method for GBL, in accordance with embodiments of the present invention.

[0082] At block 701, construct a bug localization graph to capture relationships between bug tickets and relevant source code files from historical change-sets and an underlying source code repository.

[0083] At block 703, leverage natural processing language tools to evaluate semantic similarity between a new bug ticket and a historical bug ticket.

[0084] At block 705, in response to the evaluated semantic similarity, for the new bug ticket, add links between the new bug ticket a set of similar historical bug tickets.

[0085] At block 707, incorporate the new bug ticket in the bug localization graph.

[0086] At block 709, develop a mathematical graph expression to determine a closeness relationship between the relevant source code files and the new bug ticket.

[0087] FIG. 8 is an equation for closeness scores 800 for GBL, in accordance with embodiments of the present invention.

[0088] The closeness scores 800 can be implemented by a graph algorithm or a mathematical graph expression.

[0089] In the present disclosure a “code” may refer to a software representation of the software undergoing testing (SUT). The code may be, for example, a binary code of low-level instructions, Java byte code, a high-level code, or the like. The SUT may include system software, application software, or any other type of computer program, as a whole or in components of the whole program.

[0090] In the present disclosure, a “developer” may refer to a person participating in the development of the SUT. The developer may be a program designer, a code writer, software tester, quality assurance (QA) personnel, or other personnel which may be involved in the development of the software from the conceptual stage until release to client following testing. The developer may also be a user of the SUT during the development stage of the SUT. The term “developer” may also refer to one a person with past experience in development of software programs.

[0091] In the present disclosure a “testing strategy” may refer to any decision regarding how testing is performed, such as which tools/resources are being used (e.g., test generators, formal verification tools, manual testers, or the like) and how such resources are used. For example, one testing strategy may be testing only of a sub-portion of the modules of the SUT. Another testing strategy may be testing different modules using different resources, such as using formal verification tools in testing one module and using a test suite to test a second module. Yet another testing strategy may be to test different aspects of the SUT, such as to test security, test concurrency, or the like.

[0092] As will be appreciated by one skilled in the art, aspects of the present invention may be embodied as a system, method or computer program product. Accordingly, aspects of the present invention may take the form of an entirely hardware embodiment, an entirely software embodiment (including firmware, resident software, micro-code, etc.) or an embodiment combining software and hardware aspects that may all generally be referred to herein as a “circuit,” “module” or “system.” Furthermore, aspects of the present invention may take the form of a computer program product embodied in one or more computer readable medium(s) having computer readable program code embodied thereon.

[0093] Any combination of one or more computer readable medium(s) may be utilized. The computer readable medium may be a computer readable signal medium or a computer readable storage medium. A computer readable storage medium may be, for example, but not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable combination of the foregoing. More specific examples (a non-exhaustive list) of the computer readable storage medium would include the following: an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical data storage device, a magnetic data storage device, or any suitable combination of the foregoing. In the context of this document, a computer readable storage medium may be any tangible medium that can include or store a program for use by or in connection with an instruction execution system, apparatus, or device.

[0094] A computer readable signal medium may include a propagated data signal with computer readable program code embodied therein, for example, in baseband or

as part of a carrier wave. Such a propagated signal may take any of a variety of forms, including, but not limited to, electro-magnetic, optical, or any suitable combination thereof. A computer readable signal medium may be any computer readable medium that is not a computer readable storage medium and that can communicate, propagate, or transport a program for use by or in connection with an instruction execution system, apparatus, or device.

[0095] Program code embodied on a computer readable medium may be transmitted using any appropriate medium, including but not limited to wireless, wireline, optical fiber cable, RF, etc., or any suitable combination of the foregoing.

[0096] Computer program code for carrying out operations for aspects of the present invention may be written in any combination of one or more programming languages, including an object-oriented programming language such as Java, Smalltalk, C++ or the like and conventional procedural programming languages, such as the “C” programming language or similar programming languages. The program code may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider).

[0097] Aspects of the present invention are described below with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems) and computer program products according to embodiments of the present invention. It will be understood that each block of the flowchart illustrations and/or block diagrams, and

combinations of blocks in the flowchart illustrations and/or block diagrams, can be implemented by computer program instructions. These computer program instructions may be provided to a processor of a general-purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks or modules.

[0098] These computer program instructions may also be stored in a computer readable medium that can direct a computer, other programmable data processing apparatus, or other devices to function in a particular manner, such that the instructions stored in the computer readable medium produce an article of manufacture including instructions which implement the function/act specified in the flowchart and/or block diagram block or blocks or modules.

[0099] The computer program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other devices to cause a series of operational steps to be performed on the computer, other programmable apparatus or other devices to produce a computer implemented process such that the instructions which execute on the computer or other programmable apparatus provide processes for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks or modules.

[00100] It is to be appreciated that the term “processor” as used herein is intended to include any processing device, such as, for example, one that includes a CPU (central processing unit) and/or other processing circuitry. It is also to be understood that the term “processor” may refer to more than one processing device and that various

elements associated with a processing device may be shared by other processing devices.

[00101] The term “memory” as used herein is intended to include memory associated with a processor or CPU, such as, for example, RAM, ROM, a fixed memory device (e.g., hard drive), a removable memory device (e.g., diskette), flash memory, etc. Such memory may be considered a computer readable storage medium.

[00102] In addition, the phrase “input/output devices” or “I/O devices” as used herein is intended to include, for example, one or more input devices (e.g., keyboard, mouse, scanner, etc.) for entering data to the processing unit, and/or one or more output devices (e.g., speaker, display, printer, etc.) for presenting results associated with the processing unit.

[00103] 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 executed by at least one processor for software bug localization, the method comprising:

constructing (701) a bug localization graph to capture relationships between bug tickets and relevant source code files from historical change-sets and an underlying source code repository;

leveraging (703) natural processing language tools to evaluate semantic similarity between a new bug ticket and a historical ticket;

in response to the evaluated semantic similarity, for the new bug ticket, adding (705) links between the new bug ticket a set of similar historical tickets;

incorporating (707) the new bug ticket in the bug localization graph; and

developing (709) a mathematical graph expression to determine a closeness relationship between the relevant source code files and the new bug ticket.

2. The method of claim 1, wherein constructing the bug localization graph includes extracting reference relations and change-set relations.

3. The method of claim 1, wherein the bug localization graph includes a plurality of nodes and a plurality of edges connecting the nodes, each node representing either a relevant source code file or a bug ticket, and wherein two nodes are connected if they have reference relations.

4. The method of claim 1, wherein the mathematical graph expression generates a closeness score for each new bug ticket node.

5. The method of claim 4, wherein the closeness score is updated by:

$$c^{(k+1)}[i] = \lambda \sum_{v_j \in N_i} \frac{1}{d_j} c^{(k)}[j]$$

where N_i is a set of one-hop neighbors of v_i in the bug localization graph, d_j is a node degree of neighbor v_j , k is an iteration number, and λ is a pre-defined numerical value between 0 and 1.

6. The method of claim 4, wherein the closeness score for the new ticket bug is updated by:

$$c^{(k+1)}[0] = c^{(k+1)}[0] + (1 - \lambda) \sum_{j \neq 0} c^{(k)}[j]$$

where k is an iteration number and λ is a pre-defined numerical value between 0 and 1.

7. The method of claim 4, wherein updates for the closeness score are performed iteratively.

8. A system for software bug localization, the system comprising:

a memory; and

a processor in communication with the memory, wherein the processor runs program code to:

construct (701) a bug localization graph to capture relationships between bug tickets and relevant source code files from historical change-sets and an underlying source code repository;

leverage (703) natural processing language tools to evaluate semantic similarity between a new bug ticket and a historical ticket;

in response to the evaluated semantic similarity, for the new bug ticket, add (705) links between the new bug ticket a set of similar historical tickets;

incorporate (707) the new bug ticket in the bug localization graph; and

develop (709) a mathematical graph expression to determine a closeness relationship between the relevant source code files and the new bug ticket.

9. The system of claim 8, wherein constructing the bug localization graph includes extracting reference relations and change-set relations.

10. The system of claim 8, wherein the bug localization graph includes a plurality of nodes and a plurality of edges connecting the nodes, each node representing either a relevant source code file or a bug ticket, and wherein two nodes are connected if they have reference relations.

11. The system of claim 8, wherein the mathematical graph expression generates a closeness score for each new bug ticket node.

12. The system of claim 11, wherein the closeness score is updated by:

$$c^{(k+1)}[i] = \lambda \sum_{v_j \in N_i} \frac{1}{d_j} c^{(k)}[j]$$

where N_i is a set of one-hop neighbors of v_i in the bug localization graph, d_j is a node degree of neighbor v_j , k is an iteration number, and λ is a pre-defined numerical value between 0 and 1.

13. The system of claim 11, wherein the closeness score for the new ticket bug is updated by:

$$c^{(k+1)}[0] = c^{(k+1)}[0] + (1 - \lambda) \sum_{j \neq 0} c^{(k)}[j]$$

where k is an iteration number and λ is a pre-defined numerical value between 0 and 1.

14. The system of claim 11, wherein updates for the closeness score are performed iteratively.

15. A non-transitory computer-readable storage medium comprising a computer-readable program for software bug localization, wherein the computer-readable program when executed on a computer causes the computer to perform the steps of:

constructing (701) a bug localization graph to capture relationships between bug tickets and relevant source code files from historical change-sets and an underlying source code repository;

leveraging (703) natural processing language tools to evaluate semantic similarity between a new bug ticket and a historical ticket;

in response to the evaluated semantic similarity, for the new bug ticket, adding (705) links between the new bug ticket a set of similar historical tickets;

incorporating (707) the new bug ticket in the bug localization graph; and
 developing (709) a mathematical graph expression to determine a closeness
 relationship between the relevant source code files and the new bug ticket.

16. The non-transitory computer-readable storage medium of claim 15, wherein
 constructing the bug localization graph includes extracting reference relations and
 change-set relations.

17. The non-transitory computer-readable storage medium of claim 15, wherein
 the bug localization graph includes a plurality of nodes and a plurality of edges
 connecting the nodes, each node representing either a relevant source code file or a bug
 ticket, and wherein two nodes are connected if they have reference relations.

18. The non-transitory computer-readable storage medium of claim 15, wherein
 the mathematical graph expression generates a closeness score for each new bug ticket
 node.

19. The non-transitory computer-readable storage medium of claim 18, wherein
 the closeness score is updated by:

$$c^{(k+1)}[i] = \lambda \sum_{v_j \in N_i} \frac{1}{d_j} c^{(k)}[j]$$

where N_i is a set of one-hop neighbors of v_i in the bug localization graph, d_j
 is a node degree of neighbor v_j , k is an iteration number, and λ is a pre-defined
 numerical value between 0 and 1.

20. The non-transitory computer-readable storage medium of claim 18, wherein the closeness score for the new ticket bug is updated by:

$$c^{(k+1)}[0] = c^{(k+1)}[0] + (1 - \lambda) \sum_{j \neq 0} c^{(k)}[j]$$

where k is an iteration number and λ is a pre-defined numerical value between 0 and 1.

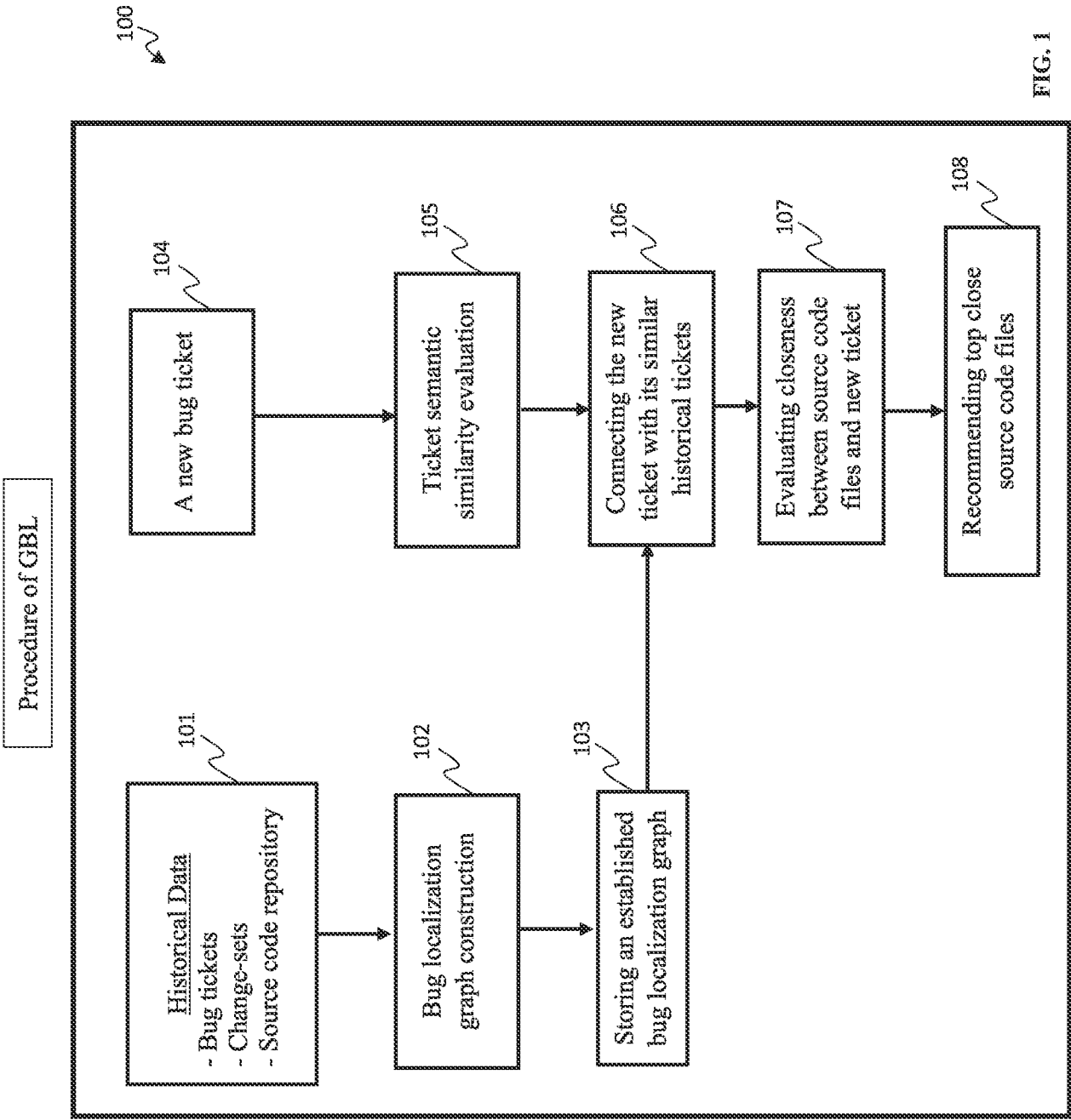


FIG. 1

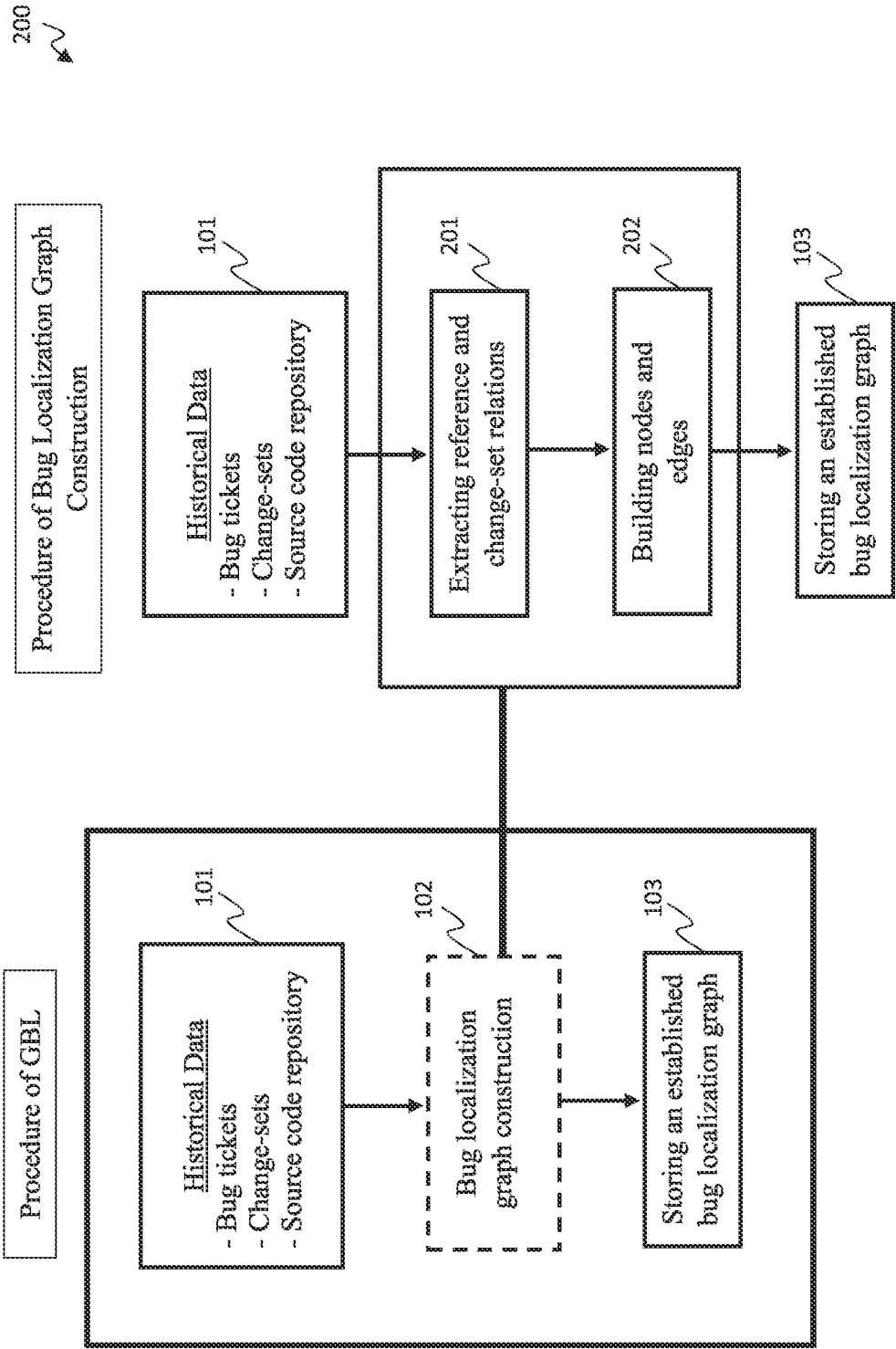


FIG. 2

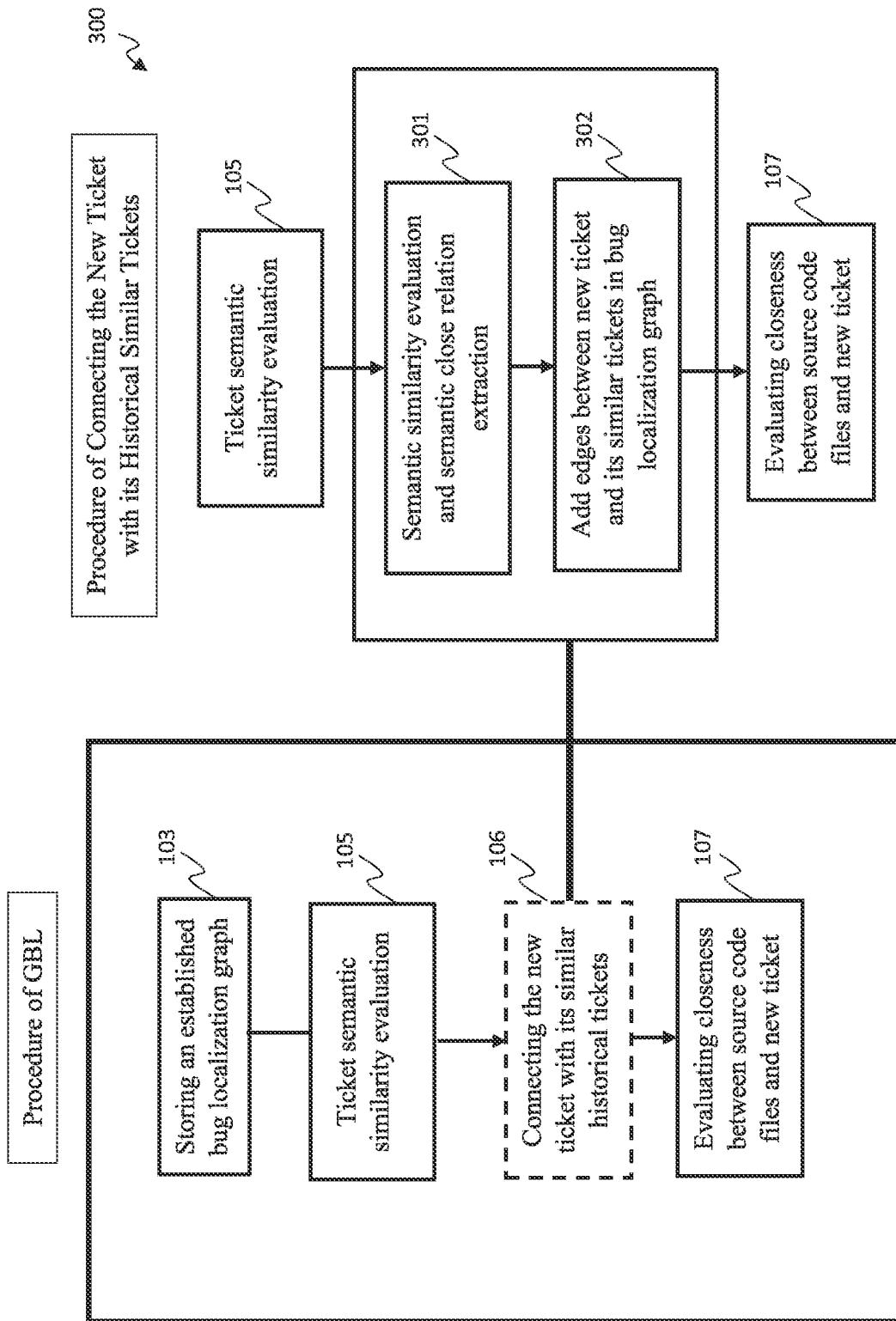


FIG. 3

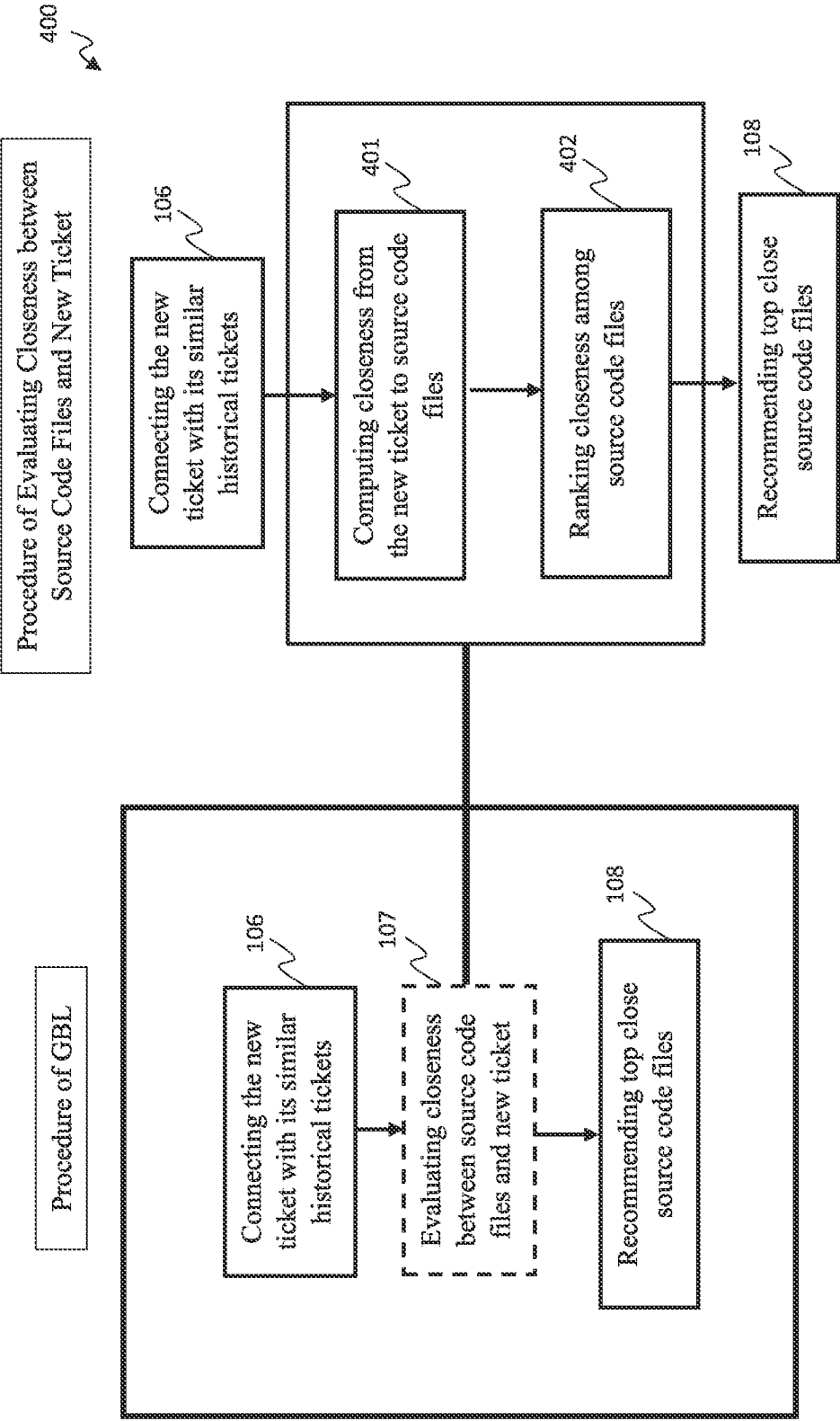


FIG. 4

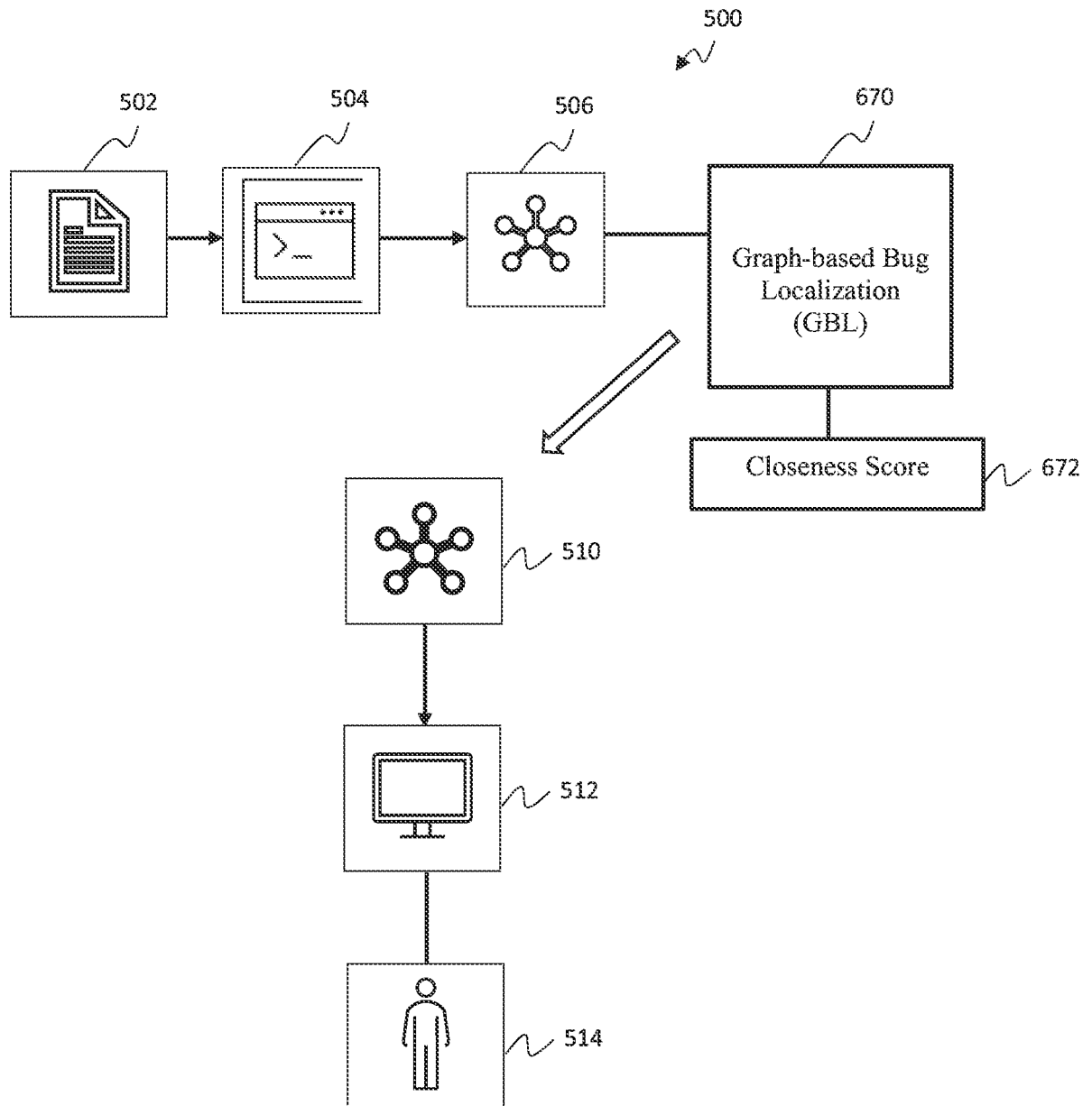


FIG. 5

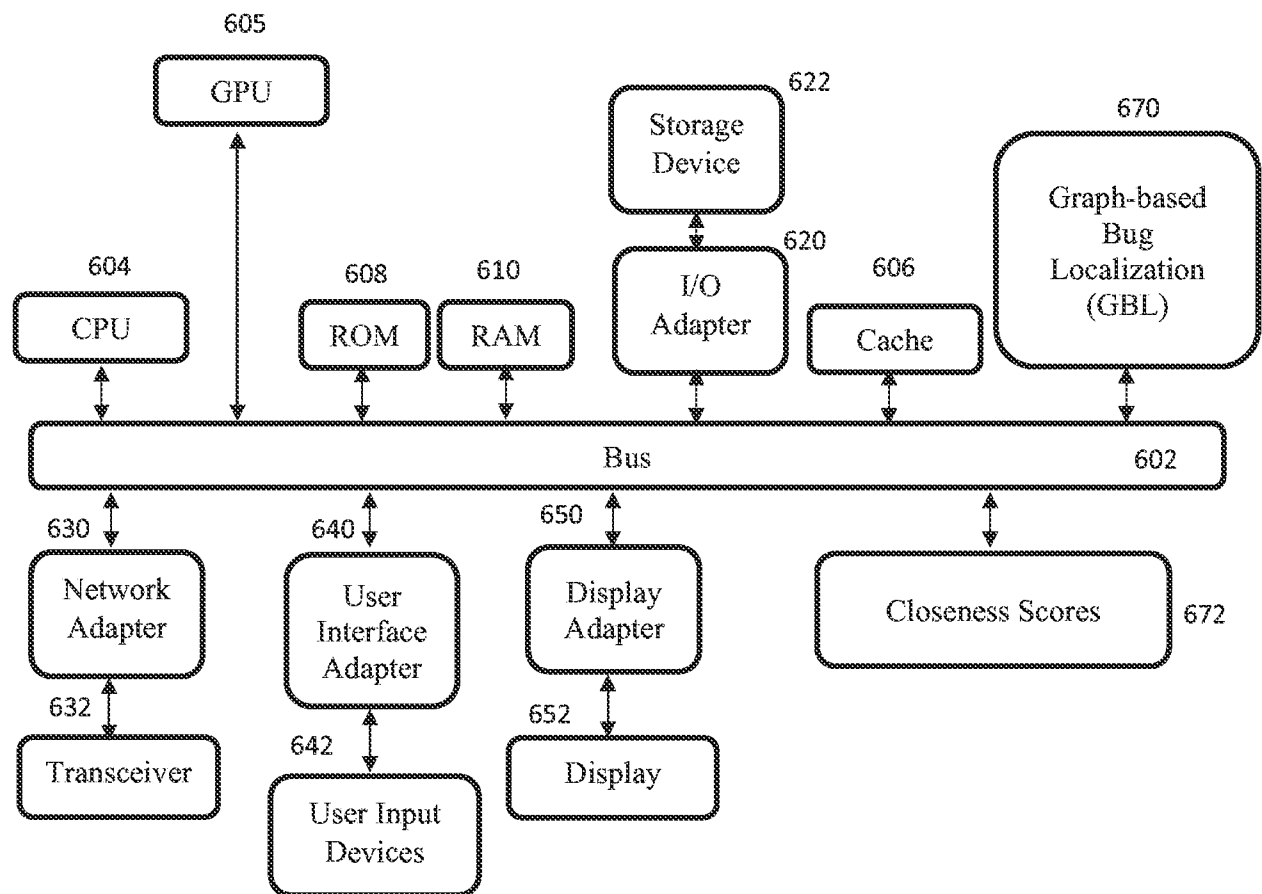


FIG. 6

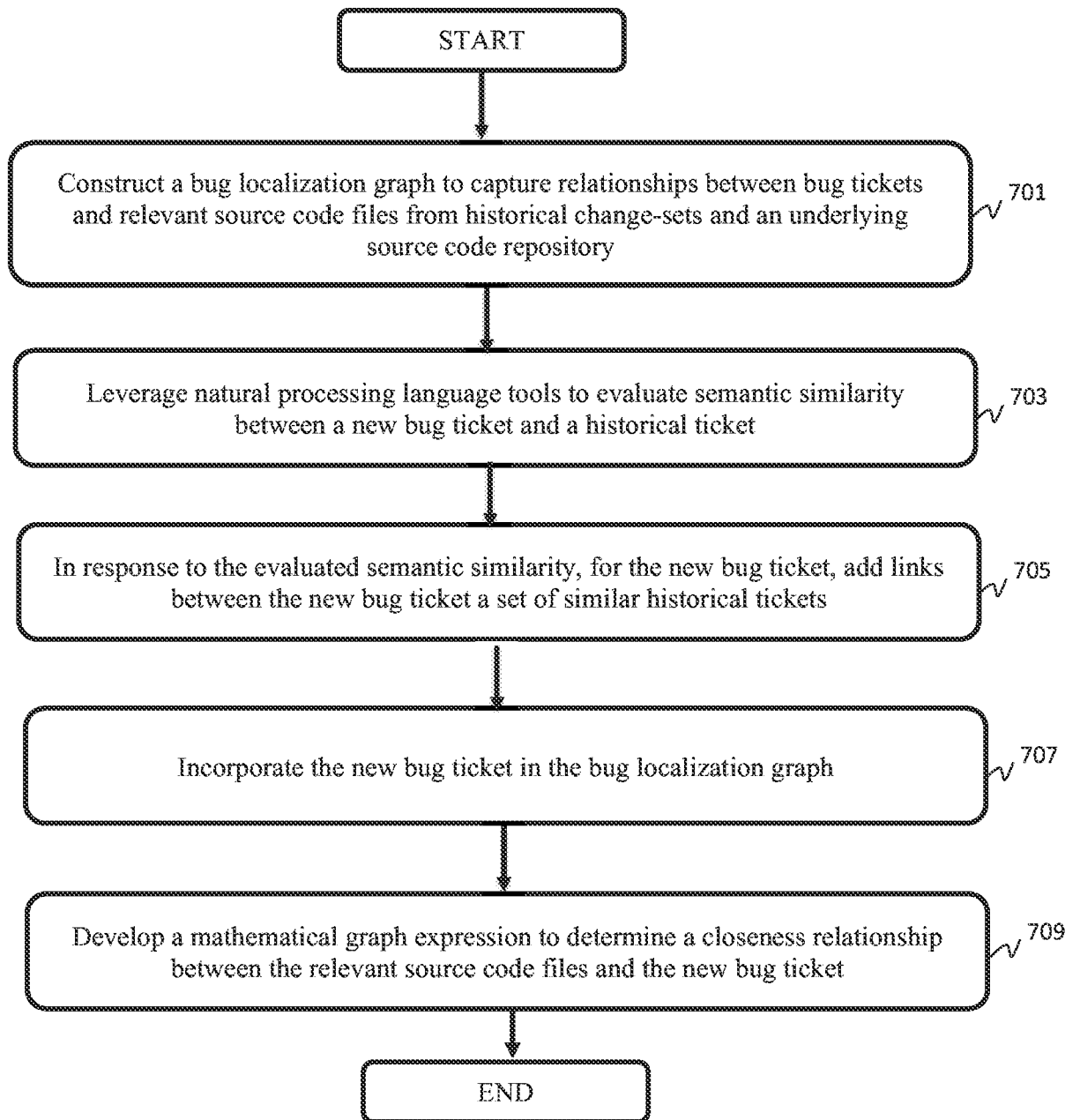


FIG. 7

800

Closeness score for each node

$$c^{(k+1)}[i] = \lambda \sum_{v_j \in N_i} \frac{1}{d_j} c^{(k)}[j]$$

Closeness score for each new ticket node

$$c^{(k+1)}[0] = c^{(k+1)}[0] + (1 - \lambda) \sum_{j \neq 0} c^{(k)}[j]$$

FIG. 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2021/021063

A. CLASSIFICATION OF SUBJECT MATTER

G06F 11/36(2006.01)i; G06F 16/36(2019.01)i; G06F 16/901(2019.01)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)

G06F 11/36(2006.01); G06F 9/44(2006.01)

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: graph-based, bug localization, semantic similarity, bug ticket, historical ticket

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2019-0196934 A1 (PURDUE RESEARCH FOUNDATION) 27 June 2019 (2019-06-27) paragraphs [0211]-[0213], [0218]; and figure 22	1-20
A	US 2015-0089478 A1 (INFOSYS LIMITED) 26 March 2015 (2015-03-26) paragraphs [0018]-[0033]; and figures 1-4	1-20
A	US 2010-0088683 A1 (VALERY GOLENDER et al.) 08 April 2010 (2010-04-08) paragraphs [0043]-[0290]; and figures 1A-17	1-20
A	MARWA GABER ABD EL-WAHAB et al., Graph Mining for Software Fault Localization: An Edge Ranking based Approach, Journal of Communications Software and Systems, Vol. 13, No. 4, PP. 178-188, 21 December 2017. [retrieved on 2021.05.12] Retrieved from the Internet: <URL: https://www.researchgate.net/publication/322434313_Graph_Mining_for_Software_Fault_Localization_An_Edge_Ranking_based_Approach > pages 178-186; and figures 1-5	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

“D” document cited by the applicant in the international application

“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

18 June 2021

Date of mailing of the international search report

21 June 2021

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
189 Cheongsa-ro, Seo-gu, Daejeon
35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

YANG, JEONG ROK

Telephone No. +82-42-481-5709

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2021/021063**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	A. Adhiselvam et al., An Enhanced Approach for Software Bug Localization using Map Reduce Technique based Apriori (MRTBA) Algorithm, Indian Journal of Science and Technology, Vol 8(35), PP. 1-9. December 2015. [retrieved on 2021.05.12] Retrieved from the Internet: <URL: https://indjst.org/download-article.php?Article_Unique_Id=INDJST7895&Full_Text_Pdf_Download=True > pages 1-8; and figures 1-6	1-20

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/US2021/021063

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)		Publication date (day/month/year)
US	2019-0196934	A1	27 June 2019	US	10108526 B2	23 October 2018
				US	2014-0149435 A1	29 May 2014
US	2015-0089478	A1	26 March 2015	US	10402194 B2	03 September 2019
US	2010-0088683	A1	08 April 2010	US	2002-0087949 A1	04 July 2002
				US	8504994 B2	06 August 2013