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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 December 2016 (15.12.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/200373 Al

(51) International Patent Classification:
H04L 12/22 (2006.01) *G06K 9/00* (2006.01)
H04L 12/26 (2006.01)

(21) International Application Number:
PCT/US20 15/034906

(22) International Filing Date:
9 June 2015 (09.06.2015)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **HEWLETT-PACKARD DEVELOPMENT COMPANY, L.P.** [US/US]; 11445 Compaq Center Drive W., Houston, Texas 77070 (US).

(72) Inventors: **HAO, Ming C ;** 1501 Page Mill Road, Palo Alto, California 94304-1 100 (US). **HAMBORG, Felix;** UniversitatsstraBe 10, 78457 Konstanz (DE). **CHANG, Nelson L.;** 1501 Page Mill Road, Palo Alto, California 94304-1 100 (US). **SCAGGS, Justin Aaron;** 5400 Legacy, Piano, Texas 75024 (US). **KEIM, Daniel;** UniversitatsstraBe 10, 78457 Konstanz (DE).

(74) Agents: **KIRCHEV, Ivan Tomov** et al; Hewlett-packard Company, Intellectual Property Administration, Mail Stop 35 3404 E. Harmony Road, Fort Collins, Colorado 80528 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, 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).

Declarations under Rule 4.17 :

— *as to the identity of the inventor (Rule 4.1 7(i))*

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.1 7(H))*

Published:

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

WO 2016/200373 A1

(54) Title: GENERATING FURTHER GROUPS OF EVENTS BASED ON SIMILARITY VALUES AND BEHAVIOR MATCHING USING A REPRESENTATION OF BEHAVIOR

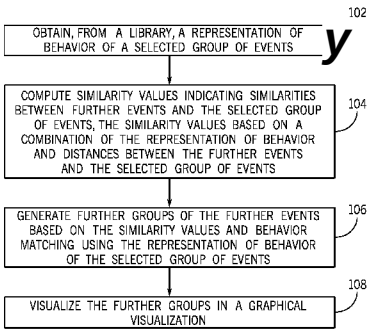

FIG. 1

(57) Abstract: A representation of behavior of a selected group of events is obtained from a library, the representation including values corresponding to respective distributions of dimensions of the events. Similarity values indicating similarities between further events and the selected group of events are computed, the similarity values based on a combination of the representation of behavior and distances between the further events and the selected group of events. Further groups of the further events are computed based on the similarity values and behavior matching using the representation of behavior of the selected group of events. The further groups are visualized in a visualization.

GENERATING FURTHER GROUPS OF EVENTS BASED ON SIMILARITY VALUES AND BEHAVIOR MATCHING USING A REPRESENTATION OF BEHAVIOR

Background

[0001] A large amount of data (such as network traffic and so forth) can be produced or received in an environment, such as a network environment that includes many machines (*e.g.* computers, storage devices, communication nodes, etc.), or other types of environments. As examples, data can be acquired by sensors or collected by applications. Other types of data can include security data, financial data, health-related data, sales data, human resources data, and so forth.

Brief Description Of The Drawings

[0002] Some implementations are described with respect to the following figures.

[0003] Fig. 1 is a flow diagram of an example pattern detection process, according to some implementations.

[0004] Fig. 2 is a flow diagram of an example process for an event behavior representation generating (profiling) process for a selected group of events, according to some implementations.

[0005] Fig. 3 is a flow diagram of an example process of computing similarity values, according to some implementations.

[0006] Fig. 4 is a graph including example data points representing respective events, which can be used for determining similarities between events and a group of events, according to some implementations.

[0007] Fig. 5 is a flow diagram of an example process of generating further groups of events, according to some implementations.

[0008] Fig. 6 illustrates an example graphical visualization including an input data set, a selected group of events, and further groups of events determined according to some implementations .

[0009] Fig. 7 is a block diagram of an example computer system according to some implementations .

Detailed Description

[0010] Activity occurring within an environment can give rise to events. An environment can include a collection of machines and/or program code, where the machines can include computers, storage devices, communication nodes, and so forth. Events that can occur within a network environment can include receipt of data packets that contain corresponding addresses and/or ports, monitored measurements of specific operations (such as metrics relating to usage of processing resources, storage resources, communication resources, and so forth), or other events. Although reference is made to activity of a network environment in some examples, it is noted that techniques or mechanisms according to the present disclosure can be applied to other types of events in other environments, where such events can relate to financial events, health-related events, human resources events, sales events, and so forth.

[001 1] Generally, an event can be generated in response to occurrence of a respective activity. An event can be represented as a data point (also referred to as a data record).

[0012] Each data point can include multiple dimensions (also referred to as attributes), where an dimension can refer to a feature or characteristic of an event represented by the data point. More specifically, each data point can include a respective collection of values for the multiple dimensions. In the context of a network environment, examples of dimensions of an event include a network address dimension *(e.g.* a source network address and/or a destination network address), a network subnet dimension *(e.g.* an identifier of a subnet), a port dimension *(e.g.* source port number and/or destination port number), and so forth. Data points that include a relatively large number of dimensions (dimensions) can be considered to be part of a high-dimensional data set.

[0013] Finding patterns (such as patterns relating to failure or fault, unauthorized access, or other issues) in data points representing respective events can be difficult when there is a very large number of data points. For example, some patterns can indicate an attack on a network environment by hackers, or can indicate other security issues. Other patterns can indicate other issues that may have to be addressed.

[0014] For example, to identify security attack patterns in a high-dimensional data set collected for a network environment, analysts can use scatter plots for identifying patterns associated with security attacks. A scatter plot includes graphical elements representing data points, where positions of the data points in the scatter plot depend on values of a first dimension corresponding to an *x* axis of the scatter plot, and values of a second dimension corresponding to a *y* axis. In some examples, the first dimension can be time, while the second dimension can include a value of a port *(e.g.* destination port) that is being accessed.

[0015] If ports are scanned (accessed) sequentially by security attacks, the security attacks can be manifested as a visible diagonal pattern in the scatter plot. If the ports are accessed in randomized order, however, the network traffic (i.e., port scan) may not be visible in the scatter plot.

[0016] In accordance with some implementations according to the present disclosure, techniques or mechanisms are provided to allow users to identify patterns associated with issues of interest to the users, such as occurrence of security attacks in a network environment, or other issues in other environments.

[0017] In some examples, a user can be presented with a graphical visualization including data points that represent respective events. Within the graphical visualization of the events, the user may see a pattern of interests, and can thus make an interactive selection of pixels (and more specifically, of a pattern of pixels) representing the data points associated with the pattern of interests. A "pattern of pixels" can refer to any collection of pixels that may be of interest to a user. The selected data points make up a selected group of events.

[0018] Fig. 1 is a flow diagram of an example process of identifying and visualizing further patterns that are similar to the pattern represented by the selected group of events. In

some implementations, the further groups of events are identified based on both a distance criterion and a behavior matching criterion (which are discussed further below).

[0019] The process of Fig. 1 obtains (at 102) a representation of behavior of the selected group of events, where the representation can include values corresponding to respective distributions of dimensions of the events in the selected group. The representation of behavior of the selected group of events can be retrieved from a profile library, where the representation of behavior was previously stored.

[0020] Each data point representing a respective event can include multiple dimensions (also referred to as attributes). For example, a data point representing an event associated with data communications in a network can include the following dimensions (or attributes): source Internet Protocol (IP) address, source port number, destination IP address, destination port number, and so forth. For each given dimension, the different events of the selected group can have respective values of the given dimension. For example, if the given dimension is the source IP address, then the events of the selected group can have different values of the source IP address.

[0021] For the given dimension, the representation of behavior of the selected group of events can include a respective diversity value that represents the distribution of values of the given dimension of the events of the selected group. Generally, a diversity value provides an indication of the distribution of values of a given dimension (or of multiple dimensions). As an example, assume that there are $N (N > 1)$ events in the selected group of events, and each data point representing a corresponding event of the selected group has M dimensions $(x_1, \ldots, x_M)$. Then for each given dimension x_j $(j = 1 \text{ to } M)$, the N events have N respective values of x_j. Note that at least some of the N values of x_j can be different and/or at least some of the N values of x_j can be the same.

[0022] In some implementations of the present disclosure, a distribution of the values of x_i in the selected group of events can be determined, and this distribution of values of x_i can be used to produce the corresponding diversity value in the representation of behavior of the selected group of events. The representation of behavior of the selected group of events

includes *M* diversity values, where each of the *M* diversity values represents the distribution of values of the dimension *xj (j* = 1 to *M).*

[0023] In some examples, the diversity values in the representation of the behavior of the selected group of events can include entropy values, where each entropy value is calculated for each respective dimension *xi*. The entropy value (or more specifically, a Shannon entropy value) can represent an expected value of information contained in each event. Entropy is zero when only a certain outcome is expected. As an example, for vertical network traffic (which can be an example of an attack against a network performed by a malicious entity in which a single IP address is scanned against multiple ports), the value of the destination IP address dimension stays the same across the events of the selected group, while the values of the destination port number dimension vary across the events of the selected group. Entropy characterizes uncertainty about a source (or sources) of information that give rise to the events; entropy increases for sources of greater randomness. In some examples, reference to "entropy" is the present discussion can be a reference to "normalized entropy," which can be entropy divided by information length.

[0024] In other examples, instead of using entropy values, the representation of the behavior of the selected group of events can include diversity values computed using other probability distribution statistical techniques, where values in the representation of the behavior are derived from statistical distributions of the dimensions.

[0025] In some examples of the present disclosure, the representation of behavior of the selected group of events can be in the form of a behavior feature vector, such as according to Table 1 below:

Table 1

ATTRIBUTE	SRC	SPT	DST	DPT
BEHAVIOR PROFILE	0.66	0	0	1.0

[0026] In Table 1 above, four dimensions are included in data points representing the selected group of events. The four dimensions are: SRC (source IP address), SPT (source

port number), DST (destination IP address), and DPT (destination port number). Each dimension is associated with a respective entropy value (0.66 for SRC, 0 for SPT, 0 for DST, and 1.0 for DPT). The foregoing entropy values are computed for each respective dimension based on the distribution of values of the respective dimension in the data points representing the events of the selected group. The four entropy values of the respective four dimensions make up a behavior profile, and these four values form a behavior feature vector in the described example.

[0027] In some examples, the behavior feature vector can be stored in a profile library for later retrieval for the purpose of identifying further groups of events that are similar to the selected group of events. A "profile library" or more generally a "library" can refer to any collection of information that can be stored in a storage medium (or storage media).

[0028] The process of Fig. 1 further computes (at 104) similarity values indicating similarities between further events and the selected group of events. The further events can include events that are not part of the selected group of events. The similarity values are based on a combination of the representation of behavior of the selected group of events that is obtained at 102, and on distances between the further events and the selected group of events. Details regarding computation of the similarity values are provided below.

[0029] The process of Fig. 1 generates (at 106) further groups of the further events based on the similarity values and also based on behavior matching using the representation of behavior of the selected group of events. In some implementations, behavior matching includes a process of determining whether adding a further event to a particular further group would result in deviation from the representation of the selected group of events. Details regarding task 106 are discussed further below.

[0030] The process of Fig. 1 then visualizes (at 108) the further groups in a graphical visualization. A "graphical visualization" can refer to any viewable representation of information that can be displayed by a display device. More generally, the further groups can be visualized in a visualization such as in files or other entities.

[003 1] Fig. 2 is a flow diagram of an example process for determining the representation of behavior of the selected group of events. The process of Fig. 2 extracts (at 202) behavior for each respective dimension of the multiple dimensions associated with the events. Extracting the behavior can include extracting the values of the respective dimension across the different events in the selected group.

[0032] The process of Fig. 2 analyzes (at 204) the dimension distribution for each dimension, to determine diversity of the respective dimension. As part of the analysis, the process of Fig. 2 can compute a value (such as an entropy value or other type of value). In some examples, an entropy value for a given dimension can be computed according to Eq. 1 below:

$$\textbf{ENTROPY(C)} = \frac{\sum i P\{C_i\} \text{ -log(p(q))}}{\sum_i c_i}. \qquad \text{(Eq. 1)}$$

[0033] In Eq. 1, the entropy value is represented as ENTROPY(C) , where C is a frequency distribution, q is the frequency of the ż-th value of a given dimension, and p(q) represents is the probability of q . Stated differently, p(q) represents the relative frequency of q , where q is a specific value out of all values in C. As a specific example, assume there are 1,000 values for a given dimension, and let c_{20} of a specific value *{e.g.* "100.1.150.2") be 30 *{i.e.* the value "100.1.150.2" occurs 30 times for the given dimension). Then, in this example, $p(c_{20})$ = 30 / 1,000.

[0034] Moreover, $\sum_i$ p(q) represents a sum over all p(q) values of a given dimension in the events of the selected group, and $\sum_i$ q represents a sum over all q values of a given dimension in the events of the selected group.

[0035] From the values computed by the analyzing performed at 204, a behavior feature vector is constructed (at 206), such as the behavior feature vector of Table 1 discussed above. The behavior feature vector is stored (at 208) in the profile library.

[0036] Fig. 3 is a flow diagram of an example process to compute the similarity values as performed at 104 in Fig. 1, according to some implementations. In the following discussion, reference is also made to Fig. 4, which is a graph illustrating various data points

that can be displayed in a graphical visualization. The horizontal axis of the graph of Fig. 4 can represent a first dimension (such as time), while the vertical axis can represent a second dimension (such as port number). In other examples, the horizontal and vertical axes can represent different dimensions.

[0037] In Fig. 4, a box 402 represents the selected group of events, as selected by a user or another entity (such as an application or a machine). Fig. 4 also depicts data points PI, P2, and P3 that represent further events. The calculation of the similarity values is between the further events represented by the data points PI, P2, and P3, and the events of the data points in the selected group 402.

[0038] The process of Fig. 3 calculates (at 302) distances between the further events and the selected group of events. As noted above, the similarity values computed at 104 in Fig. 1 are based on distances between the further events and the selected group of events, and also based on the representation of behavior *(e.g.* behavior feature vector) of the selected group of events.

[0039] The distance for dimension i between each data point Px_i representing a further event and each data point Py_i in the selected group 402 is calculated at 302. The distance can be computed in one of several different ways based on the type of dimension. If the dimension is a categorical dimension (a dimension that does not have numerical values, but rather, has values in different categories), then the distance is a categorical distance measure $d(x, y)$, where $d(x,y) = 0$ if x does not equal y, and $d(x, y) = 1$ if x equals y. For other types of dimensions, other techniques for computing the distance can be used, such as based on a difference between values of x and y.

[0040] The average distance, $\boldsymbol{dm}(\boldsymbol{i})(Px_i, Py_i)$, per dimension i is then calculated (at 304) across each data point representing a further event and the data points in the selected group 402. A weighted distance, DISTANCE_WEIGHTED(Px, Py), for a data point Px representing a further event is then calculated (at 306) according to Eq. 2, with the entropy values across the dimensions used as weighting factors.

$$\text{DISTANCE\_WEIGHTED}(Px, Py) = \sum_i w_i \cdot dm(i)(Px_i, Py_i). \quad \text{(Eq. 2)}$$

[0041] In Eq. 2, w_i is the weight for dimension *i*, where w_i is derived from the entropy value in the behavior feature vector for dimension *i*. In some examples, w_i can be set equal to the entropy value for dimension *i*. In other examples, w_i can be calculated based on the entropy value and at least another factor. There is one weighted distance calculated for each data point *(e.g.* PI, P2, or P3 in Fig. 4) representing a further event.

[0042] In some examples, the weighted distance, DISTANCE_WEIGHTED*(Px, Py),* can be normalized to a value between 0 and 1 by dividing the weighted distance by the sum of all weights w_i. In some examples, the normalized weighted distance can be used as the similarity value computed at 104. Alternatively, the non-normalized weighted distance can be used as the similarity value computed at 104.

[0043] The further events are sorted (at 308) by the similarity values (the normalized or non-normalized weighted distances). The sorted data points are thus arranged in order of their similarity values.

[0044] Fig. 5 is a flow diagram of an example process to generate the further groups at 106 in Fig. 1, using the sorted data points that have been sorted according to the similarity values. The process of Fig. 5 has multiple iterations for multiple further groups and multiple further events that are to be added to a respective further group.

[0045] The process of Fig. 5 can start with an initial further group (which can be empty to start with) and can iterate to identify additional further groups until no further groups can be generated. For a current further group (502), the process of Fig. 5 considers a current further event (504). The current further event can start with the further event that is most similar to the selected group of events *(i.e.* the further event with the smallest weighted distance to the selected group of events). The process performs (at 506) a distance check for the current further event with respect to the current further group. More specifically, the distance check includes checking if the distance between the current further event and the last event of the current further group is less than a specified threshold. The "last" event in the

current further group is the event of the current further group that is farthest away from the selected group of events.

[0046] If the distance check does not pass (as determined at 508), in other words, the distance between the current further event and the last event of the current further group is not less than the specified threshold, then the process of Fig. 5 iterates (at 510) to the next current further group and re-iterates through the process of Fig. 5. Note that the current further event is not added to the current further group in this scenario.

[0047] However, if the distance check passes (as determined at 508), then the process of Fig. 5 performs (at 512) behavior matching using the diversity values of the representation of behavior of the selected group of events ***(e.g.*** entropy values of the behavior feature vector discussed above).

[0048] The following describes an example using entropy values, although other types of diversity values can be used in other examples.

[0049] The behavior matching is performed for each dimension of multiple dimensions of the events. For each dimension ***i,*** the behavior matching sets a target entropy, TE_i, equal to the entropy value of the behavior feature vector for dimension ***i.***

[0050] In addition, the behavior matching calculates a current entropy, CE_i, for dimension ***i,*** for the current further group without the current further event included in the current further group. In addition, the behavior matching calculates a new entropy, NE_i, for dimension ***i*** for {current further group, current further event}, in other words, the expanded current further group with the current further event added. The computation of the new entropy or the current entropy is based on use of Eq. 1.

[005 1] The behavior matching then computes a current entropy difference according to Eq. 3:

$$\textbf{CURRENT\_ENTR0PY\_DIFFERENCE}(7\mathfrak{L}'_i, CE_i) = \backslash TE_i - CE_i \backslash. \quad \textbf{(Eq. 3)}$$

[0052] The behavior matching then computes a new entropy difference according to Eq. 4:

$$\text{NEW\_ENTROPY\_DIFFERENCE}(TE_i, NE_i) = |TE_i - NE_i|. \quad \textbf{(Eq. 4)}$$

[0053] If the new entropy difference is less than the current entropy difference, then the behavior matching is determined (at 514) to have passed. However, if the new entropy difference is not less than the current entropy difference, then the behavior matching is determined (at 514) to not have passed.

[0054] If the behavior matching is determined to not have passed, then the process of Fig. 5 iterates to the next further group. However, if the behavior matching is determined to have passed, then the process adds (at 516) the current further event to the current further group, and iterates (at 518) to the next current further event.

[0055] More generally, the behavior matching determines a first representation of behavior *(e.g.* behavior feature vector of entropy values) (the first representation containing first diversity values) of the current further group including the current further event, and based on at least one value of the representation of behavior of the current further group and at least one corresponding value of the representation of behavior of the selected group of events, the behavior matching decides whether to add the current further event to the further group.

[0056] In addition, a second representation of behavior *(e.g.* behavior feature vector of entropy values) (the second representation containing second diversity values) of the current further group without the given further event is determined. A first difference *(e.g.* according to Eq. 4) between a value of the first representation of behavior and a corresponding value of the representation of behavior of the selected group of events is computed. A second difference *(e.g.* according to Eq. 3) between a value of the second representation of behavior and a corresponding value of the representation of behavior of the selected group of events is computed. The first difference and the second difference are compared, where deciding whether to add the current further event to the current further group is in response to the comparing.

[0057] Once the further groups of further events have been determined using techniques or mechanisms according to some implementations, at least two of the further groups of

further events may be combined based on a relationship between the at least two further groups.

[0058] Fig. 6 illustrates an example graphical visualization that includes a representation 602 of the entire data set of data points. Since there are a large number of data points in the representation 602, it can be difficult to identify patterns that are similar to a particular pattern. Within the representation 602 of the entire data set, a user can notice a particular pattern of interest, and can make an interactive selection of the data points of the particular pattern—this interactive selection produces the selected group of events discussed above.

[0059] In Fig. 6, a representation 604 of the selected group of events of the particular pattern of interest is depicted. Using techniques or mechanisms according to the present disclosure as discussed above, further groups can be identified and represented with respective representations 606, 608, 610, 612, and 614 in the graphical visualization of Fig. 6.

[0060] It is noted that data points can be represented as pixels in each of the representations 602 to 614. A pixel can be assigned a visual indicator, such as a respective color, based on the value of a particular dimension of the event represented by the pixel. Each of the representations 602 to 614 has a respective color scale *(e.g.* 616) that has different colors to indicate different values of the particular dimension.

[0061] In some examples, interactive control elements can be provided in the graphical visualization to allow a user to interactively combine or split further groups identified using techniques or mechanisms according to some implementations.

[0062] In some examples, the representations 606, 608, 610, 612, and 614 can be marked with visual indicators *(e.g.* different colors) to rank how similar the respective further groups are to the selected group of events (represented by representation 604).

[0063] Using techniques or mechanisms according to the present disclosure, a user can more easily identify multiple groups of events that may be similar to a selected group of events that is of interest to the user.

[0064] Fig. 7 is a block diagram of an example computer system 700 according to some implementations. The computer system 700 can include one computer, or a distributed arrangement of multiple computers (where different tasks of techniques according to some implementations can be executed in different computers of the distributed arrangement). The computer system 700 includes a processor (or multiple processors) 702. A processor can include a microprocessor, a microcontroller, a physical processor module or subsystem, a programmable integrated circuit, a programmable gate array, or a physical control or computing device.

[0065] The processor(s) 702 can be coupled to a non-transitory machine-readable or computer-readable storage medium (or storage media) 704, which can store various machine-executable instructions, including instructions 706 to generate a representation of behavior of a selected group of events (such as according to tasks 202-206 of Fig. 2), instructions 708 to compute similarity values (such as according to task 104 in Fig. 1 and tasks 302-308 of Fig. 3), and instructions 710 to generate further groups of further events (such as according to task 106 in Fig. 1 and tasks 502-518 of Fig. 5), and other machine-readable instructions to perform other tasks as discussed above.

[0066] The storage medium (or storage media) 704 can include one or multiple different forms of memory including semiconductor memory devices such as dynamic or static random access memories (DRAMs or SRAMs), erasable and programmable read-only memories (EPROMs), electrically erasable and programmable read-only memories (EEPROMs) and flash memories; magnetic disks such as fixed, floppy and removable disks; other magnetic media including tape; optical media such as compact disks (CDs) or digital video disks (DVDs); or other types of storage devices. Note that the instructions discussed above can be provided on one computer-readable or machine-readable storage medium, or alternatively, can be provided on multiple computer-readable or machine-readable storage media distributed in a large system having possibly plural nodes. Such computer-readable or machine-readable storage medium or media is (are) considered to be part of an article (or article of manufacture). An article or article of manufacture can refer to any manufactured single component or multiple components. The storage medium or media can be located

either in the machine running the machine-readable instructions, or located at a remote site from which machine-readable instructions can be downloaded over a network for execution.

[0067] In the foregoing description, numerous details are set forth to provide an understanding of the subject disclosed herein. However, implementations may be practiced without some of these details. Other implementations may include modifications and variations from the details discussed above. It is intended that the appended claims cover such modifications and variations.

What is claimed is:

1. A method comprising:

obtaining, by a system comprising a processor from a library, a representation of behavior of a selected group of events, the representation including values corresponding to respective distributions of dimensions of the events;

computing, by the system, similarity values indicating similarities between further events and the selected group of events, the similarity values based on a combination of the representation of behavior and distances between the further events and the selected group of events;

generating, by the system, further groups of the further events based on the similarity values and behavior matching using the representation of behavior of the selected group of events; and

visualizing, by the system, the further groups in a visualization.

2. The method of claim 1, wherein the generating of a first further group of the further groups comprises, for a given further event:

determining a first representation of behavior of the first further group including the given further event;

based on at least one a value of the first representation of behavior of the first further group and at least one corresponding value of the representation of behavior of the selected group of events, decide whether to add the given further event to the first further group.

3. The method of claim 2, wherein the generating of the first further group further comprises, for the given further event:

determining a second representation of behavior of the first further group without the given further event;

computing a first difference between a value of the first representation of behavior and a corresponding value of the representation of behavior of the selected group of events;

computing a second difference between a value of the second representation of behavior and a corresponding value of the representation of behavior of the selected group of events;

comparing the first difference and the second difference,

wherein deciding whether to add the given further event to the first further group is in response to the comparing.

4. The method of claim 1, wherein the generating of a first further group of the further groups comprises, for a given further event:

determining whether a distance between the given further event and an event in the first further group is less than a specified threshold; and

adding the given further event to the first further group in response to the determining.

5. The method of claim 1, further comprising sorting the further events according to the similarity values, wherein the generating of the further groups uses the sorted further events.

6. The method of claim 1, wherein the values in the representation of behavior of the selected group of events comprises entropy values.

7. The method of claim 1, wherein the values in the representation of behavior of the selected group of events are based on statistical distributions of the dimensions.

8. The method of claim 1, further comprising receiving interactive user selection in a visualization of a pattern of pixels representing events in the selected group of events.

9. The method of claim 1, wherein computing the similarity values comprises:

computing distances between the further events and the selected group of events; and

applying weights to the computed distances, the weights including values of the representation of behavior of the selected group of events.

10. The method of claim 1, further comprising combining at least two of the further groups according to relationships between the at least two further groups.

11. A system comprising:

at least one processor to:

determine a distribution of values of each dimension of a plurality of dimensions of events in a selected group of events;

generate diversity values in a representation of behavior of the selected group of events based on the distributions of values of the respective dimensions;

compute similarity values indicating similarities between further events and the selected group of events, the similarity values computed using the diversity values and distances between the further events and the selected group of events; and

generate further groups of the further events based on the similarity values and behavior matching using the representation of the behavior of the selected group of events.

12. The system of claim 11, wherein the behavior matching comprises using the representation of behavior of the selected group of events to decide whether to add a respective further event to one of the further groups.

13. The system of claim 12, wherein the behavior matching further comprises:

generating first diversity values for a given further group of the further groups with a given further event included in the given further group;

generating second diversity values for the given further group without the given further event included in the given further group; and

deciding whether or not to add the given further event to the given further group using the first and second diversity values and using the diversity values of the representation of behavior of the selected group of events.

14. The system of claim 13, wherein the at least one processor is to further:

compute a distance between the given further event and the given further group,

wherein deciding whether or not to add the given further event to the given further group is further based on the computed distance.

15. An article comprising at least one non-transitory machine-readable storage medium storing instructions that upon execution cause a system to:

obtain, from a library, a representation of behavior of a selected group of events interactively selected in a visualization, the representation including values corresponding to respective distributions of dimensions of the events;

compute similarity values indicating similarities between further events and the selected group of events, the similarity values based on distances between the further events and the selected group of events, and on the representation of behavior;

sort the further events according to the similarity values; and

generate further groups of the further events using the sorted further groups and according to behavior matching using the representation of behavior of the selected group of events; and

cause visualization of the further groups.

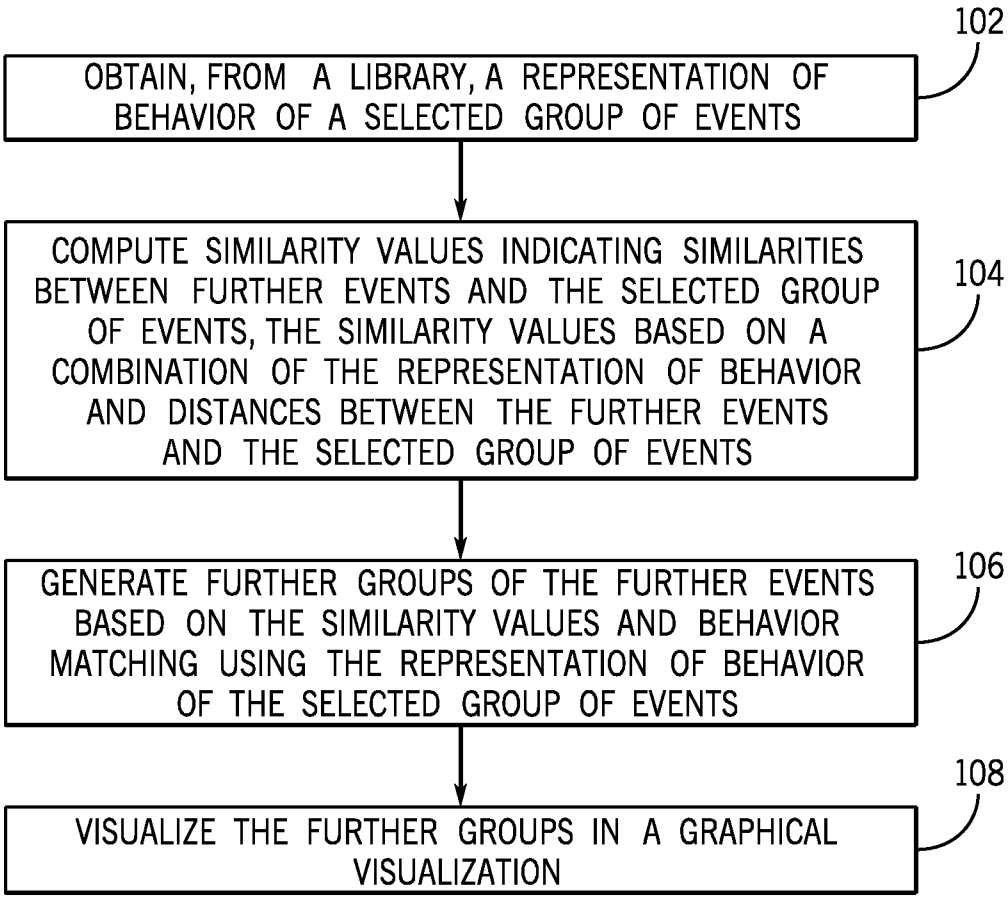
102
OBTAIN, FROM A LIBRARY, A REPRESENTATION OF BEHAVIOR OF A SELECTED GROUP OF EVENTS
104
COMPUTE SIMILARITY VALUES INDICATING SIMILARITIES BETWEEN FURTHER EVENTS AND THE SELECTED GROUP OF EVENTS, THE SIMILARITY VALUES BASED ON A COMBINATION OF THE REPRESENTATION OF BEHAVIOR AND DISTANCES BETWEEN THE FURTHER EVENTS AND THE SELECTED GROUP OF EVENTS
106
GENERATE FURTHER GROUPS OF THE FURTHER EVENTS BASED ON THE SIMILARITY VALUES AND BEHAVIOR MATCHING USING THE REPRESENTATION OF BEHAVIOR OF THE SELECTED GROUP OF EVENTS
108
VISUALIZE THE FURTHER GROUPS IN A GRAPHICAL VISUALIZATION

FIG. 1

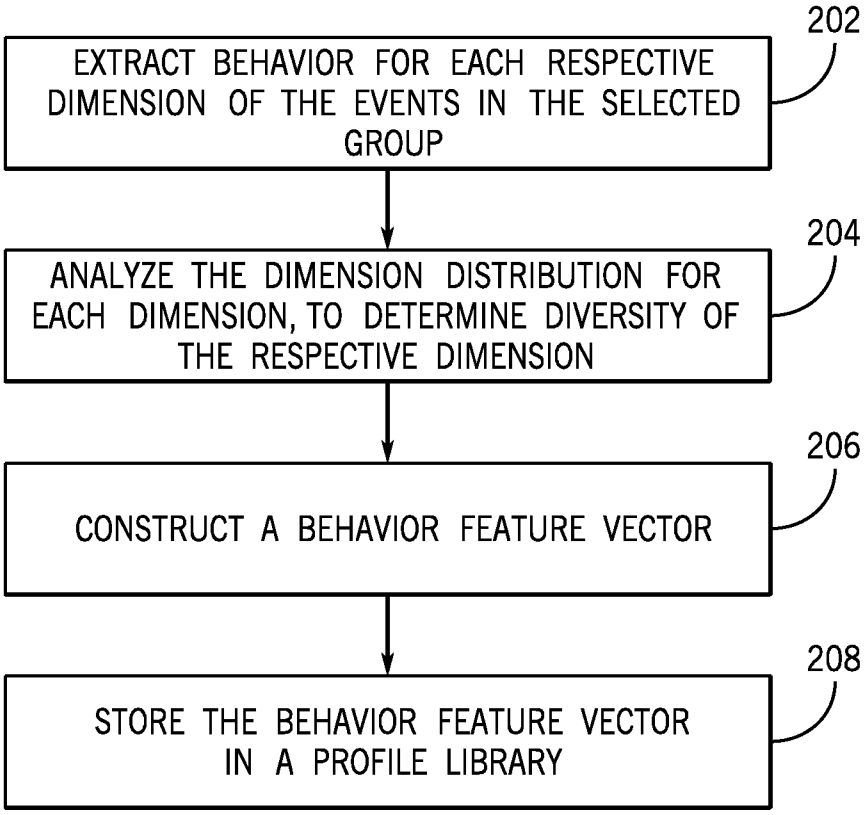
EXTRACT BEHAVIOR FOR EACH RESPECTIVE DIMENSION OF THE EVENTS IN THE SELECTED GROUP
202
ANALYZE THE DIMENSION DISTRIBUTION FOR EACH DIMENSION, TO DETERMINE DIVERSITY OF THE RESPECTIVE DIMENSION
204
CONSTRUCT A BEHAVIOR FEATURE VECTOR
206
STORE THE BEHAVIOR FEATURE VECTOR IN A PROFILE LIBRARY
208

FIG. 2

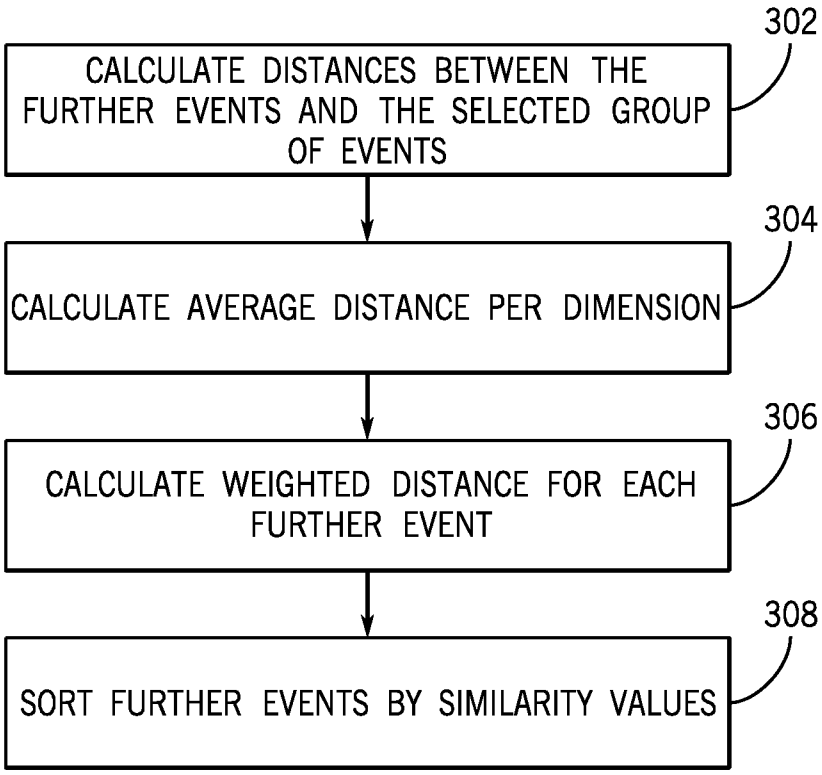
302
CALCULATE DISTANCES BETWEEN THE FURTHER EVENTS AND THE SELECTED GROUP OF EVENTS
304
CALCULATE AVERAGE DISTANCE PER DIMENSION
306
CALCULATE WEIGHTED DISTANCE FOR EACH FURTHER EVENT
308
SORT FURTHER EVENTS BY SIMILARITY VALUES

FIG. 3

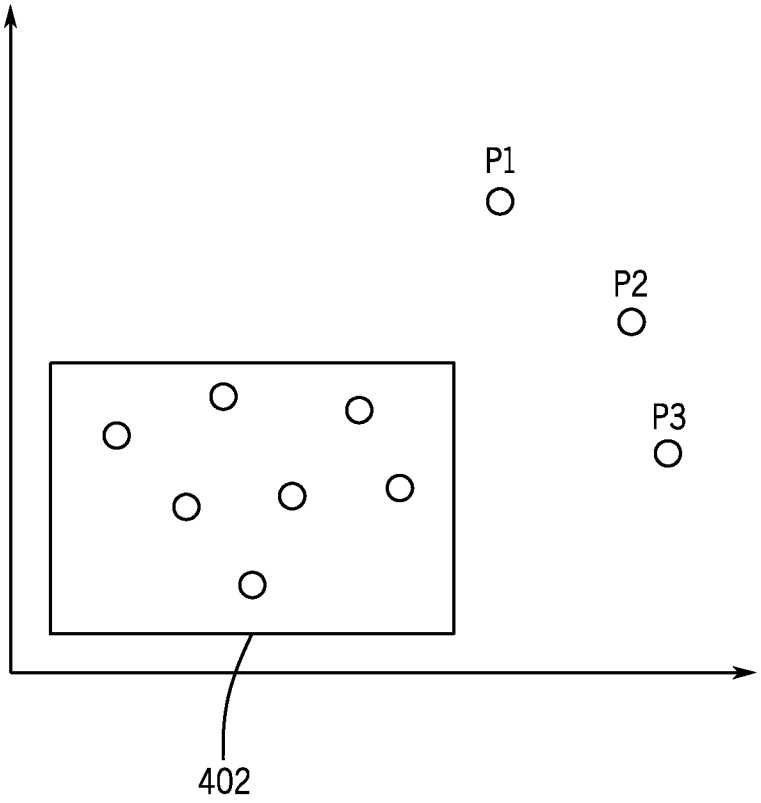

FIG. 4

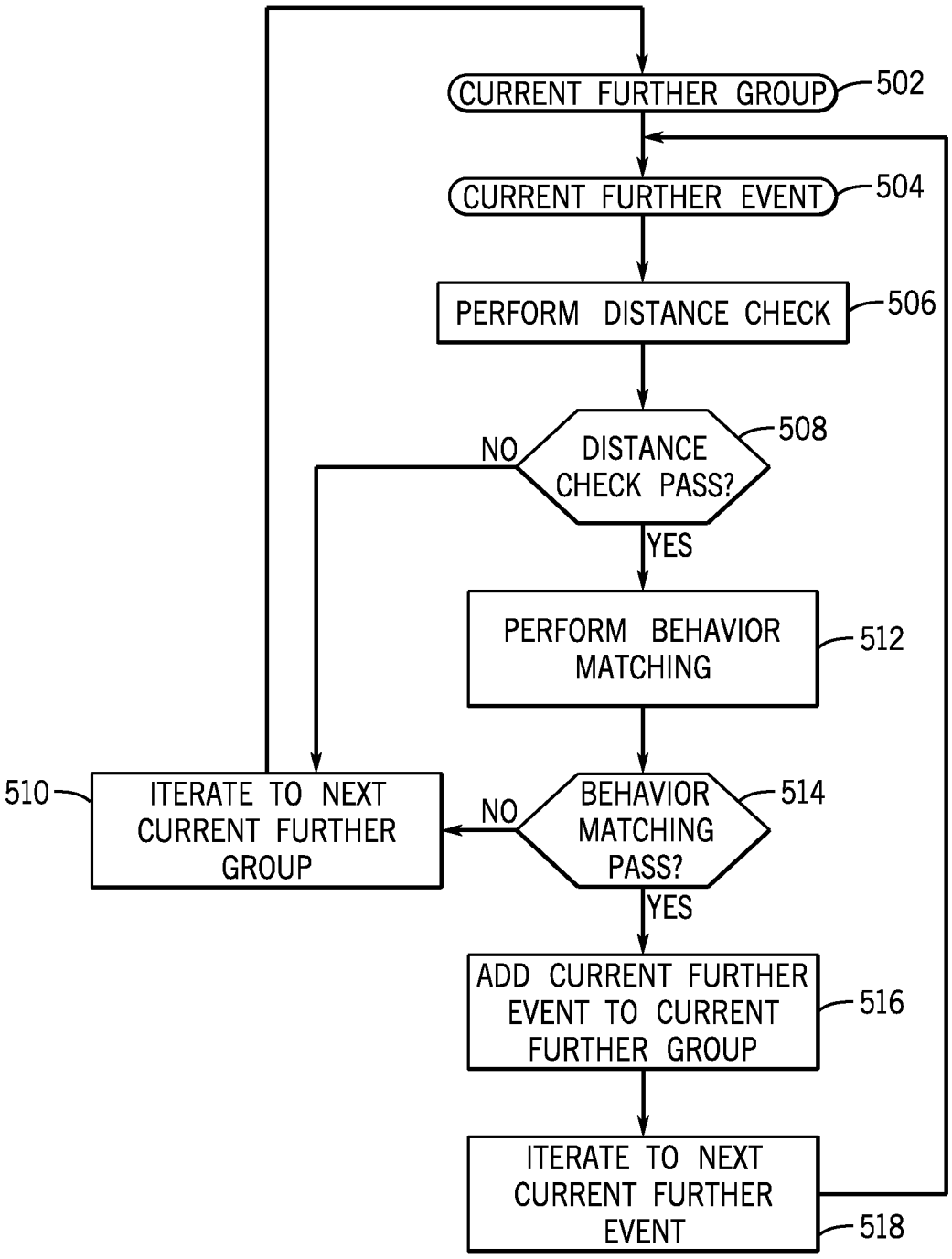
CURRENT FURTHER GROUP
502
CURRENT FURTHER EVENT
504
PERFORM DISTANCE CHECK
506
NO
DISTANCE CHECK PASS?
508
YES
PERFORM BEHAVIOR MATCHING
512
510
ITERATE TO NEXT CURRENT FURTHER GROUP
NO
BEHAVIOR MATCHING PASS?
514
YES
ADD CURRENT FURTHER EVENT TO CURRENT FURTHER GROUP
516
ITERATE TO NEXT CURRENT FURTHER EVENT
518

FIG. 5

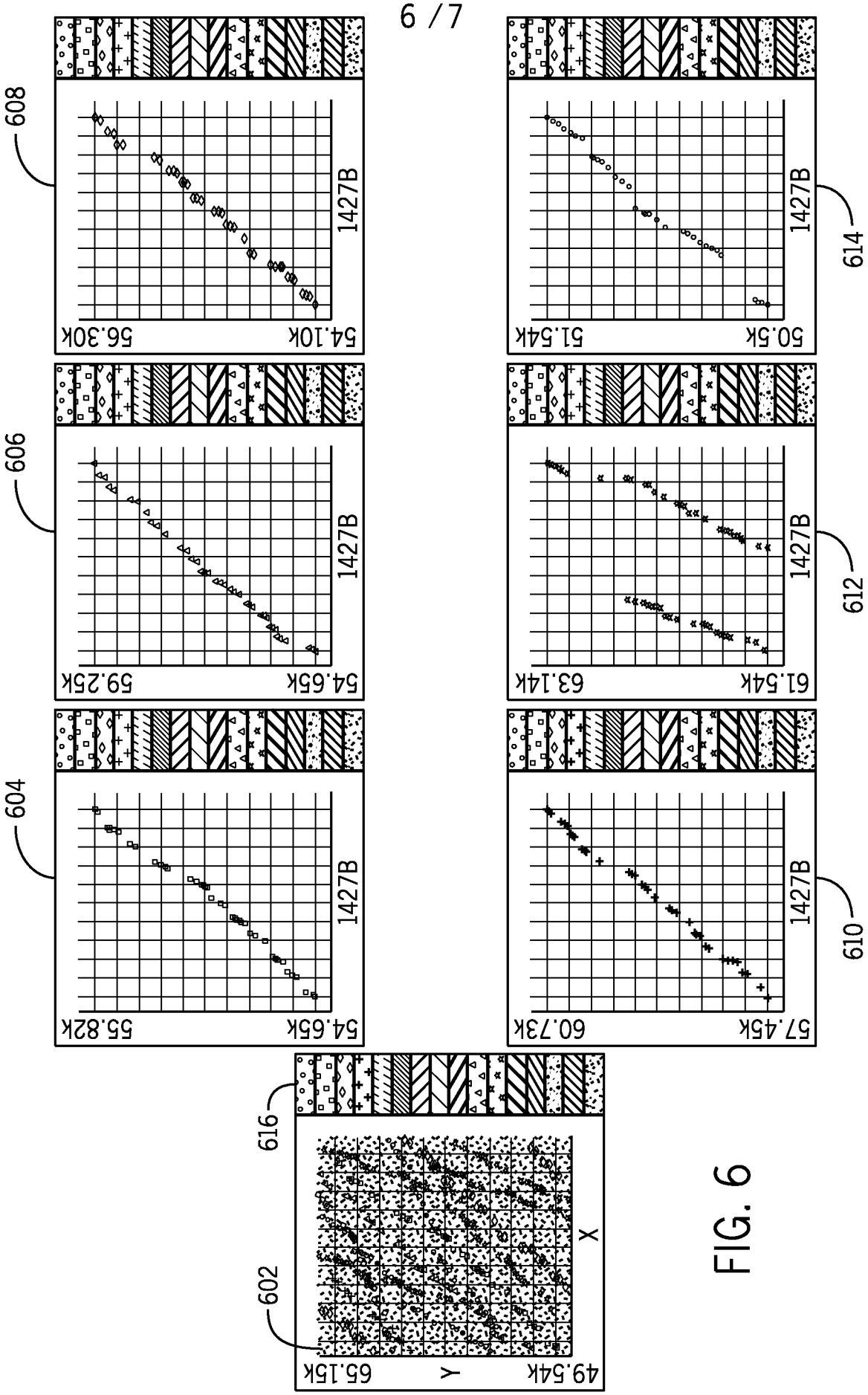
608
606
604
616
602
614
612
610
1427B
49.54K
65.15K
Y
X
54.10K
56.30K
54.65K
59.25K
54.65K
55.82K
50.5K
51.54K
61.54K
63.14K
57.45K
60.73K

FIG. 6

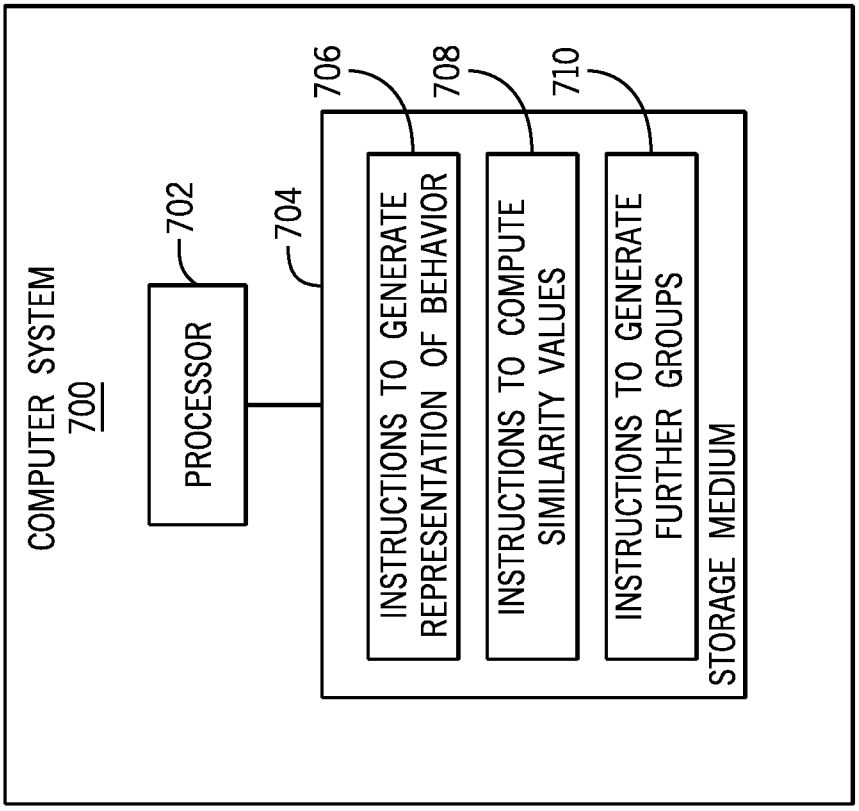
COMPUTER SYSTEM
700
PROCESSOR
702
704
706
INSTRUCTIONS TO GENERATE REPRESENTATION OF BEHAVIOR
708
INSTRUCTIONS TO COMPUTE SIMILARITY VALUES
710
INSTRUCTIONS TO GENERATE FURTHER GROUPS
STORAGE MEDIUM

FIG. 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2015/034906

A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/22(2006.01)i, H04L 12/26(2006.01)i, G06K 9/00(2006.01)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)
H04L 12/22; G06Q 99/00; G06F 15/173; G06F 17/00; G06Q 10/00; G06N 5/02; G06F 15/18; G06F 9/45; H04L 12/26; G06K 9/00

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: event, similarity, value, matching, behavior, distance, visualization

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2014-0046892 A1 (XURMO TECHNOLOGIES PVT. LTD.) 13 February 2014 See paragraphs [0011], [0059]; and figure 4.	1-15
A	US 8655800 B2 (RUTH BERGMAN et al.) 18 February 2014 See column 2, line 35 - column 3, line 30; and figures 1-2.	1-15
A	US 8005767 B1 (VINCENT A. CASSELLA) 23 August 2011 See column 16, line 4 - column 17, line 12; and figures 3-6.	1-15
A	US 2006-0074621 A1 (OPHIR RACHMAN) 06 April 2006 See paragraphs [0020H0026]; and figures 1-4.	1-15
A	US 2009-0248497 A1 (GEOFFREY J. HUETER) 01 October 2009 See paragraphs [0039]-[0042]; and figures 1-4.	1-15

☐ 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	Date of mailing of the international search report
03 March 2016 (03.03.2016)	**04 March 2016 (04.03.2016)**

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. +82-42-472-7140

Authorized officer
KIM, Seong Woo
Telephone No. +82-42-481-3348

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US20 15/034906

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2014-0046892 Al	13/02/2014	US 2014-·0046653 Al	13/02/2014
		us 2014-·0046877 Al	13/02/2014
		us 2014-·0046977 Al	13/02/2014
US 8655800 B2	18/02/2014	CN 102257487 A	23/11/2011
		CN 102257487 B	01/07/2015
		DE 112008004025 T5	01/03/2012
		GB 201106730 DO	01/06/2011
		GB 2476905 A	13/07/2011
		US 2011-·0202483 Al	18/08/2011
		₩O 2010-·042112 Al	15/04/2010
US 8005767 Bl	23/08/2011	None	
US 2006-0074621 Al	06/04/2006	None	
US 2009-0248497 Al	01/10/2009	US 08566256 B2	22/10/2013
		us 08788S445 B2	22/07/2014
		us 08903811 B2	02/12/2014
		us 2009-·0248494 Al	01/10/2009
		us 2009-·0248495 Al	01/10/2009
		us 2009-·0248496 Al	01/10/2009
		us 2009--0248523 Al	01/10/2009
		us 2009-·0248599 Al	01/10/2009
		us 2009-·0248682 Al	01/10/2009