



- (51) **International Patent Classification:**
G06F 17/30 (2006.01)
- (21) **International Application Number:**
PCT/US2010/030518
- (22) **International Filing Date:**
9 April 2010 (09.04.2010)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (71) **Applicants (for all designated States except US):** HEWLETT-PACKARD DEVELOPMENT COMPANY, L.P. [US/US]; 11445 Compaq Center Drive W., Houston, Texas 77070 (US). TRASTOUR, David [FR/FR]; SAP Labs France, 805 Avenue du Dr. Maurice Donat, B.P. 1216, F-06250 Mougins (FR).
- (72) **Inventors; and**
- (75) **Inventors/Applicants (for US only):** HICKEY, Marianne [GB/GB]; HP Ltd., Bristol Filton B1 (BUK01), M M131, Mail Stop M39, Fulton Rd. Stoke Gifford, Bristol Bristol BS34 8QZ (GB). BYDE, Andrew [GB/GB]; 9 Page Street, Cardiff, South Glamorgan, CF11 9LB (GB). RAHMOUNI, Maher [TN/GB]; HP Ltd., Building 3 (HP Labs), Long Down Ave., Bristol Bristol BS34 8QZ (GB). BARTOLINI, Claudio [IT/US]; 1501 Page Mill Road, Palo Alto, California 94304-1100 (US).
- (74) **Agents:** HART, Kevin M. et al.; Hewlett-Packard Company, Intellectual Property Administration, 3404 E. Harmony Road, Mail Stop 35, Fort Collins, Colorado 80527-2400 (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, 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, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, 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, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (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, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

- as to the identity of the inventor (Rule 4.17(i))
- as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))

Published:

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

(54) **Title:** METHOD AND SYSTEM FOR COMPARING AND LOCATING PROJECTS

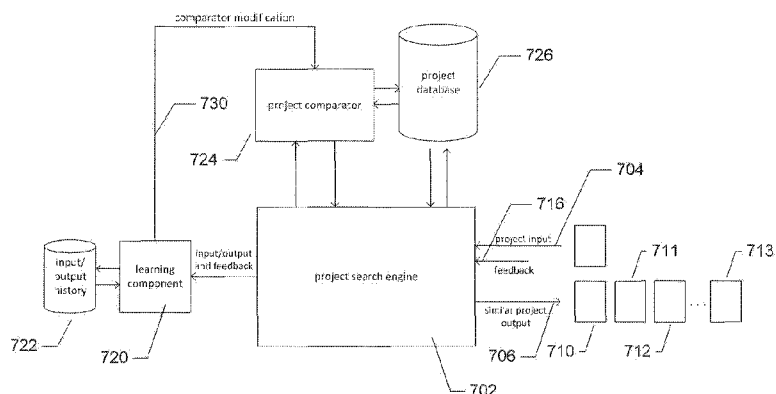


FIGURE 7

(57) **Abstract:** One embodiment of the present invention is directed to a project search system, implemented on one or more computer systems, comprising two or more stored projects represented by electronically stored, computer-accessible data that are stored in a project-database component; a project-comparator subsystem, executed on the one or more computer systems, that generates a similarity metric for two electronically-represented projects input to the project-comparator subsystem; and a project search engine, executed on the one or more computer systems, that receives data that represents an input project, identifies, by pairwise comparison of the stored projects with the input project using the project-comparator subsystem, one or more stored projects with greatest similarity to the input project, and outputs, to a graphical interface displayed on a computer display or to an application program, indications of the one or more stored projects with greatest similarity to the input project.

WO 2011/126489 A1

METHOD AND SYSTEM FOR COMPARING AND LOCATING PROJECTS

TECHNICAL FIELD

The present invention is related to project-management applications
5 and other project-based computational tools and applications.

BACKGROUND

Project management within various types of organizations and
institutions has become, in recent decades, a large and vital component of these
10 organizations and institutions. With the advent of cost-effective, networked
computing and software applications, a variety of different applications and
computing tools for facilitating project management have been developed. These
tools and applications may facilitate creation of electronic documents that describe
projects, tracking of project progress, project expenditures, project staffing, and other
15 components of projects. While these tools provide great utility and functionality to
project managers and others concerned with project management, project managers
and various other personnel and managers involved, at various levels, with projects
and project-management applications and computational tools continue to seek new
and improved project-management tools and applications that provide new and useful
20 functionality, increased efficiency, and other advantages.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates an electronic-data representation of a project.

Figure 2 illustrates a vector-based similarity metric.

25 Figures 3A-E illustrate one approach to generating a vector that
corresponds to a text document.

Figure 4 illustrates comparison of numeric values, extracted from an
electronic document, to produce a similarity metric that indicates the similarity of the
two documents based on the numeric values contained within them.

30 Figure 5 shows a method for computing a similarity metric between
two graphical images.

Figure 6 shows an overall architecture of a computational project comparator that represents one embodiment of the present invention.

Figure 7 illustrates a project search engine that represents a family of embodiments of the present invention.

5 Figures 8A-E provide control-flow diagrams that illustrate operation of a project-search engine that represents one embodiment of the present invention.

Figure 9 illustrates a typical electronic computer on which project search engines, project comparators, and other embodiments of the present invention may be implemented.

10

DETAILED DESCRIPTION

Currently, project-management applications and tools do not provide efficient, convenient, and accurate methods and systems for identifying projects, electronically encoded and stored in project databases and other computational facilities, related to a particular project. A current approach is to use keyword-based searching of text documents related to projects, but such keyword-based searching methods are tedious, error prone, and often produce poorly defined, large result sets that require a great detail of additional review and analysis. Another current approach involves executing relational-database queries, or other database-management-system queries, with respect to project-related data stored in database management systems. These approaches depend on a rational and efficient database schema design and on use of database management systems for storing significant portions of project-related data, and are often tedious, error-prone, and difficult for project managers and other personal to use.

25 One embodiment of the present invention is directed to a computational system for determining the similarity between two different projects by comparing the electronic-data representations the two projects using project-feature comparators and one or more feature-similarity aggregators. According to this project-comparator embodiment, information corresponding to each of a number of project features is extracted from both projects and the extracted information from
30 both projects for each feature is supplied to a corresponding feature comparator,

which outputs a feature-similarity value. The outputs from the feature comparators are weighted and aggregated by one or more feature-similarity aggregators to produce a final project-similarity value. The project comparator system can, in turn, be employed by a project search engine, representing an additional embodiment of the present invention, to carry out pairwise project comparisons in order to identify stored projects represented by electronic data that are similar to an input project.

Figure 1 illustrates an electronic-data representation of a project. The project can generally be considered to be a collection of electronic documents 102-106, each containing different types of information. Certain portions of the information may be expressed in formatted text, other portions of the information may be expressed numerically and in equations, yet other portions of the information may be expressed in graphics, graphical images, and photographs, other portions of the information may be expressed in unformatted text, and additional types of information may include application-specific application-data files.

In Figure 1, the electronic-data representation of the project includes a project-overview document 102 that contains formatted text encoding a project title 110, a brief project description 112, the name of a project manager 114, the names of various engineers assigned to the project 116, job titles, descriptions, or roles needed for the project, often needed prior to assignment of people to jobs, and other such information. The electronic-data representation of the project also contains additional documents, including a business proposal 103, an architecture document for the device to be developed as a result of successful undertaking of the project 104, a financial statement for the project 105, and a tasks document 106 including schedules, work flows, and other information. These additional documents may contain many different types of data, including the various types of data discussed above.

One embodiment of the present invention is a computational system for comparing the electronic-data representations of two projects to produce a similarity metric that expresses a computed similarity of the two projects. In one embodiment of the present invention, the computed similarity is a real value within the range [0, 1]. In one embodiment of the present invention, the project-similarity

value is computed from a number of feature-similarity values for various features of the electronic-data-representation of the project.

There are generally many ways to decompose or partition an electronic-data-representation of a project into a discrete number of representative features. For example, considering the electronic-data representation of a project shown in Figure 1, project features may include a character string containing the project title, a variable-length character string containing a description of the project, a date corresponding to the inception date of the project, graphical objects that represent architectural-diagram overviews of the project, text and numeric data describing the project plan, including sub-tasks, schedules, dependencies, and work breakdown structure, mixed text and numeric data containing financial parameters for the project, text and numeric data containing resource parameters for the project, including human resource roles and the number of personnel for each role, a list of character strings containing the names of project personnel, and other such features.

Identification of a set of characteristic features for a project, and extraction of data from the electronic-data representation of a project for each feature, is somewhat arbitrary. For example, a personnel feature used to characterize projects in one implementation of the project comparator may include a list of character strings that represents the names of all people involved with a project, while, in a different implementation of the project comparator, the names of those involved with the project may be distributed among roles such as project-manager, engineers, and support-staff features. In general, within any particular computational environment, electronic-data representations of projects may be somewhat standardized, and the data structures and data content for each of the various sections and entities within a standardized project description may be selected as features, for purposes of the project-comparison system that represents one embodiment of the present invention. The data associated with features may overlap, in certain implementations, and much of the data contained in an electronic-data representation of a project may be omitted from features and from consideration by the project comparator.

A variety of techniques can be used to compare the data content of a particular feature extracted from two projects being compared by a project

comparator system, according to various embodiments of the present invention. Comparison techniques that produce similarity-metric values often depend strongly on the underlying data type for the information corresponding to a feature. Next, a number of different comparison methods for different types of data are discussed, as
5 illustrations of various approaches used in feature-comparator components of a project comparator system, according to various embodiments of the present invention.

Figures 2-3E illustrate a text-comparison method that can be used to determine the similarity between two documents, pages, or paragraphs with formatted
10 or unformatted text. Figure 2 illustrates a vector-based similarity metric. In Figure 2, two three-dimensional vectors, vector A 202 and vector B 204 are compared by a well-known comparison technique. In general, the technique can be used for n -dimensional vectors. The two vectors can be expressed in terms of the coordinates of the points at the ends of the vectors, 206 and 208, using familiar coordinate-triplet
15 expressions 210. The dot product of two vectors is defined according to expression 212 in Figure 2. A magnitude, or length of the two vectors can be computed by computing the square root of the self dot product of the vectors, as shown in equation 214 in Figure 2. Using the definition of the dot product and the above-described computation of the magnitudes of the vectors, one can obtain an expression for the
20 cosine of the angle θ 216 between the two vectors in terms of the coordinates of the vectors, shown in expression 218 in Figure 2. The cosine of the angle between the two vectors can be used as a similarity metric for the two vectors. When the two vectors have the same direction, or, in other words, are parallel, the cosine of the angle between the two vectors is 1. When the two vectors have opposite directions,
25 and are anti-parallel, the cosine of the angle between the two vectors is -1. The value of $\cos \theta$ ranges from -1 to 1, and indicates the degree to which the two vectors have similar directions, or orientations. Using $\cos \theta$ as a similarity metric ignores differences in the magnitudes of the vectors which, for the purpose of comparing text documents of different lengths, is advantageous. The magnitudes of the vectors can
30 be included, as an additional term, in a more complex similarity metric computation, in alternative methods.

One general approach for computing a similarity metric for two text documents is to express each of the two text documents as a vector in an n -dimensional vector space and then to compute $\cos \theta$ for the angle defined by the two vectors in the n -dimensional vector space, as discussed above with reference to Figure 2. Figures 3A-E illustrate one approach to generating a vector that corresponds to a text document. As shown in Figure 3A, a text document 302 can be considered to be a sequence of words. In Figure 3A, each word, such as word 304, is represented by a rectangle. In the illustrated method, the vector corresponding to the text document 306 is a one-dimensional array with elements comprising numeric values. In Figure 3A, prior to construction of an n -dimensional vector that represents a text document, all the elements of the vector, or array, 306 are set to 0. Each element in the vector is indexed by a word, such as the word 308, which indexes the first element of the vector 310. In many text-comparison methods, the set of words, shown in column 312, that index the elements of a vector 306 representing a text document is an alphabetically ordered list of normalized terms encountered in the two documents that are being compared, or an alphabetically ordered list of terms obtained from a dictionary, analysis of a corpus of documents, or by some other means. Construction of the n -dimensional vector representing the text document also employs a normalization and filtering module 316 which normalizes and filters input words to generate either a normalized word within the index of terms 312 or an indication that input word does not correspond to a word in the index, and therefore does not contribute to construction of the n -dimensional vector. As one example, the normalization and filtering module may choose to filter out, or ignore, articles, pronouns, and other commonly encountered words of low information content. As another example, the normalization and filtering module may transform a variety of different derived forms of a particular word into the root word of the derived forms. For example, plural terms may be converted to singular terms, capitalized letters may be converted to corresponding lower-case letters, and various verb forms may be converted to the singular, present-tense form of the verb, synonyms may be converted to particular alternative terms, and other such normalization tasks may be carried out.

Instruction of the n -dimensional vector proceeds sequentially, by extracting each word, in order, from the text document, inputting each extracted word to the normalization and filtering module 316, and using the corresponding output from the normalization and filtering module to update at most one element in the vector for each word extracted from the document. In Figure 3A, a first word 320 is extracted from the document and input to the normalization and filter module 316. Output from the normalization and filtering module indicates that the first word 320 does not contribute to construction of the n -dimensional vector representing the document. The first word 320 may, for example, be an article. In Figure 3B, the second word 322 of the document is extracted and input to the normalization and filtering module 316, which outputs the root or stem of a word that occurs as index word 324 in the index. The corresponding element 326 of the n -dimensional vector is incremented, to reflect an occurrence of the root word in the document. In Figure 3C, the third word in the document 328 is processed similarly, resulting in incrementing a corresponding element 330 of the n -dimensional vector. Each next word in the document is similarly processed, resulting finally, as shown in Figure 3D, in an initial vector V_0 306 that includes, as elements, the number of times each of the indexing root words occurs in the text document. As shown in Figure 3E, this initial vector, V_0 , 306 is then arithmetically normalized to produce a normalized vector 340. Normalization may involve dividing each of the entries in the vector by the total sum of entries in the initial vector V_0 . Alternative normalization methods may be used. The normalized vector, V_n , 340 is then multiplied, element-by-element, by a normalized weight vector 342 to produce a final vector, V_f , 344 that represents the text document. The weights in the weight vector 342 are used to assign levels of significance, or importance, to each of the different indexing terms. Different normalized weight vectors may be used for text documents related to different subject areas, for example, to facilitate generation of meaningful similarity metrics on a document-domain basis.

Figure 4 illustrates comparison of numeric values, extracted from an electronic document, to produce a similarity metric that indicates the similarity of the two documents based on the numeric values contained within them. In Figure 4, two

electronic documents 402 and 404 are shown to contain a number of numeric values, each numeric value represented by a rectangle, such as rectangle 406. The numeric values can be extracted from each document and placed into an array in numeric order 408 and 410, with the first element in each array, such as element 412, containing the
5 numeric value with minimum value, and the final element, such as element 414, in each array containing the numeric value with maximum value, extracted from the corresponding document. A mean 416 and variance 418 can be easily computed for the numbers in each array. Moreover, a relative overlap 420 can be computed as the
10 length of a region of overlap 422 between the ranges represented by the two arrays divided by the total combined range 424 of the two arrays. A similarity value, generally ranging between [0, 1], can then be computed by a function 425 that receives, as input parameters, the means, variances, and relative overlap. Many different arithmetic combinations of these computed values may be used to compute
15 an overall similarity metric, in various alternative similarity-metric computational methods.

Figure 5 shows a method for computing a similarity metric between two graphical images. In Figure 5, electronic representations of a first graphical image 502 and a second graphical image 504 are compared. In a first step, the second graphical image is rotated 506 and scaled 508 to as closely approximate the first
20 graphical image 502 as possible. Then, the rotated and scaled second image is superimposed over the first image 510, and a degree of overlap of the two images is computed. In the example shown in Figure 5, the square 520, ellipses 522-523, and circle 524 of the first image and the second, rotated and scaled image overlap completely, while the outer ellipse 526 of the first image is not overlapped by a
25 corresponding element in the second image. The number of pixels in the overlapping regions divided by the total number of pixels for the two superimposed images may, for example, be used as the relative overlap of the two graphical images. A similarity metric can be computed as a function of the rotation, scale, and degree of overlap
528. Additional factors may consider colors in colored images, geometric primitives,
30 and other such image characteristics.

Figure 6 shows an overall architecture of a computational project comparator that represents one embodiment of the present invention. The computational project comparator that represents one embodiment of the present invention receives two electronic-data representations of projects 602 and 604 as input and generates a real-valued similarity metric output 606 within the range [0, 1] that represents the similarity of the two projects to one another. A feature demultiplexer 608 extracts information from the two projects for each of n different features on which the project comparison is based. For example, the data corresponding to a first feature is extracted from the two projects by the feature demultiplexer and output 610-611 to the inputs 612 and 613 of a feature comparator 616 for the first feature that compares the information extracted from the two projects for the first feature 612 and 613 to produce a similarity metric 618 that is input to a second-level comparator 620. The various feature comparators 616 and 624-629 for the n features are generally data-type-specific comparators, such as the various different types of data-type-specific comparators discussed with reference to Figures 3A-E, 4, and 5. Two different feature comparators may be two instances of a common comparator type. For example, two different features may both have text-document data types and may therefore use a text-document data-type comparator, such as that discussed with reference to Figures 3A-E. Many different hybrid feature comparators are also possible. For example, a feature comparator for comparing financial statements associated with a project may employ both a text-document-comparator function, such as that described with reference to Figures 3A-E, as well as a numeric-value-comparator function, such as that discussed with reference to Figure 4, producing a final similarity metric that is an arithmetic combination of similarity metrics produced by the text-document-comparison functionality and the numeric-value-comparison functionality.

The various different feature comparators generally include numerous tunable parameters. For example, the text feature comparator, discussed above with reference to Figures 3A-E, includes a weight vector, with the individual elements of the weight vector tunable in order to optimize the similarity metrics produced by the text feature comparator. The numeric feature comparator, discussed with reference to

Figure 4, may employ various constant multipliers of values computed from the mean, variance, and relative-overlap values. The second-level comparator 620 also generally applies multiplicative weights to the similarity metrics produced by the various feature comparators to produce a final similarity metric, and the weights applied by the second-level comparator to each of the similarity metrics produced by the feature comparators represent a set of tunable parameters for the project comparator, as a whole.

Figure 7 illustrates a project search engine that represents a family of embodiments of the present invention. The project search engine 702 receives, as input 704, a first project and returns 706, as output, a list of similar projects 710-713, in many embodiments sorted in descending or ascending project-comparator-produced similarity-metric-value order. A project search engine may additionally receive, as a final input 716, feedback from a user or from an application program related to the quality, or accuracy, of the output list of similar projects. This feedback, which, along with the original input 704 and output 706, is transmitted by the project search engine to a learning component 720 which stores the input/output history and feedback in a learning-component database 722 for subsequence analysis. The project search engine accesses a project comparator subsystem 724, discussed above with reference to Figure 6, and a database of projects 726. The learning component 720 produces feedback 730 to the project comparator 724 to tune and re-weight the project comparator, based on input/output and feedback history, in order to optimize the similarity metrics produced by the project comparator. The project search engine locates, in the project database, a number of projects similar to an input project by pairwise comparison of the input project to all or a subset of the projects contained in the project database using the project comparator subsystem.

Figures 8A-E provide control-flow diagrams that illustrate operation of a project-search engine that represents one embodiment of the present invention. Figure 8A provides a control-flow diagram for an event handling loop for a project-search engine. In step 802, the project-search engine waits for a next event to occur. In other words, the described project-search engine that represents one embodiment of the present invention is an event-driven, continuously available module that can be

accessed, through project-search-engine interfaces, by users and application programs. When an event occurs, and the event represents a request for a project search, as determined in step 804, then the project-search engine is invoked, in step 806, to search for projects related to an input project. Otherwise, when the event
5 corresponds to a request for comparison of two projects, as determined in step 808, the project comparator subsystem is invoked, in step 810, to provide a similarity metric value that represents the similarity of the two projects to be compared, in step 810. When a timer has expired for the learning module, as determined in step 812, then the learning module is invoked, in step 814, to analyze stored input/output and
10 feedback history in order to provide tuning modifications to the project comparator, as discussed with reference to Figure 7. Then, in step 816, the timer is reset. The types of events, detected in steps 804, 808, and 812, are representative of many different possible events that may be detected and handled by the event-handling loop of a project-search engine. Other events, which may include additional timers, events
15 associated with various maintenance and administrative functionality for the project-search engine, and other events are handled in a catch-all event handler routine 818. Thus, when awakened in step 802, the event handler may detect the occurrence of, and launch routines and other functionality to handle, multiple different events.

Figure 8B provides a control-flow diagram for a project comparator,
20 one implementation of which is discussed with reference to Figure 6, which represents one embodiment of the present invention. In step 820, the project comparator receives two project inputs. The projects may be input as files, electronic documents, or reference to files, electronic documents, or other computer-accessible electronic-data representations of the two projects. In the *for*-loop of steps 822-826,
25 each feature used by the project comparator in the comparison of the two projects is considered in a separate iteration of the *for*-loop. In step 823, information corresponding to the currently considered feature is extracted from the input projects and, in step 824, this extracted information is furnished to a feature comparator corresponding to the currently considered feature to produce a similarity-metric value,
30 as discussed above with reference to Figure 6. In step 825, the similarity-metric value produced by the feature comparator for the currently considered feature is then input

to the second-level comparator. As discussed above, with reference to Figure 6, the second-level comparator multiplies the similarity metrics input from the various feature comparators by corresponding weights, and computationally combines the weighted similarity metrics to produce an overall similarity-metric output, which is returned in step 828. Note that the project comparator may carry out iterations of the *for*-loop of steps 822-826 sequentially or in parallel fashion, depending on the operating system and/or hardware system above which the project comparator is implemented.

Figure 8C provides a control-flow diagram for the project-search engine, which represents one embodiment of the present invention, described above with reference to Figure 7. In step 830, the project-search engine receives a project as a file, set of electronic documents, or one or more references to files, electronic documents, or other electronic-data representation of the project. In step 832, a list of project/similarity-metric pairs is initialized to the empty list. Then, in the *for*-loop of step 834-842, the input project is compared to each of a number of projects stored in a project database to find similar projects which are entered into the list of project/similarity-metric pairs initialized in step 832, generally in similarity-metric-value order. Each iteration of this *for*-loop represents a pairwise comparison of the input project with a project extracted from the project database. In step 835, a filter is applied to the currently considered project extracted from the project database. This filter may determine, from the values of several key features or from other characteristics or parameters of the currently considered project, whether the project has sufficient indication of similarity to justify a project-comparator comparison, and is thus a candidate project. When the currently considered project is ascertained to be a candidate project, as determined in step 836, then the currently considered project and the input project are input to the project comparator, in step 837. If the list of project/similarity-metric pairs is full, as determined in step 838, and the output from the project comparator, called in step 837, is a similarity metric with a value greater than the minimum valued similarity metric contained in the list of project/similarity-metric pairs, as determined in step 839, then the project having the minimum similarity-metric value currently on the list is removed and the currently considered

project is added, along with the corresponding similarity metric, to the list so that the list remains in similarity-metric-value order, in step 840. When the list is not full, the currently considered project is added to the list of project/similarity-metric pairs, along with the similarity metric produced by the project comparator for the currently
5 considered project in step 837. When there are more projects in the project database to considered, as determined in step 842, control flows back to the beginning of the *for-loop* for a subsequent iteration of the *for-loop*. In step 844, the list of project/similarity-metric pairs, or a subset of the list of project/similarity-metric pairs with similarity metrics above a threshold value, are displayed to a user or returned to
10 a requesting application and, in step 846, when the user or application return feedback, then the list of project/similarity-metric pairs and input project are transmitted, along with that feedback, to the learning module, as discussed above with reference to Figure 7. Various applications and display interfaces may employ the similarity metrics associated with projects in the returned list to display projects in
15 ascending or descending similarity-metric-value order, color code projects according to similarity value, or otherwise provide indications of the projects most similar to the input project.

Figure 8D provides a control-flow diagram for the learning module, which represents one embodiment of the present invention, discussed above with
20 reference to Figure 7. The learning module is invoked, at intervals, to analyze stored learning-module data, provided to the learning module by the project-search engine, in order to provide feedback to the project comparator. In step 850, the learning module computes correlation coefficients for each feature used in project comparison with a list of projects associated with positive feedback. In step 852, the learning
25 module computes correlation coefficients of each feature with a list of projects associated with negative feedback. In step 854, features with correlation coefficients with respect to project lists associated with negative feedback greater than some threshold correlation-coefficient value are added to a first list. In step 856, the remaining features with correlation coefficients with respect to project lists associated
30 with negative feedback less than or equal to the threshold value are added to a second list. In step 858, features with computed correlation coefficients with respect to

project lists associated with positive results greater than a threshold value are removed from the second list. Thus, the first list includes features that appear to be strongly correlated with negative results and the second list includes features that are not strongly associated with negative results but also not strongly correlated to positive results. In step 860, the second portion of the learning-module functionality is invoked.

Figure 8E provides a control-flow diagram for the remaining functionality of the learning module, invoked in step 860 of Figure 8D. In the *for*-loop of steps 870-875, each feature on the first list is separately considered, in a single-pass traversal of the first list. In step 871, a change in feature-comparator parameters that would appear to ameliorate the correlation with negative results is searched for, by systematically altering feature-comparative parameters. When a change in the feature-comparator parameters is found to ameliorate the correlation with negative results, as determined in step 872, then those changes are output to a list a changes, in step 873, and the feature is removed from list 1. In a second *for*-loop of steps 880-885, a similar traversal of the second list is made, this time looking for changes to the feature-comparator parameters that would produce a correlation with positive results. If such changes are found, then they are output to the list of changes in step 883, and the currently considered features removed from list 2. In step 888, an optimization method is carried out on the weights applied by the second-level comparator, given changes to the feature comparators according to the list of changes, to produce second-level similarity metrics that best track feedback with respect to the project-list results. Finally, in step 890, the changes to the feature comparators and weights to be applied by the second-level comparator are transmitted to the project comparator, as discussed above with reference to Figure 7, to tune the project comparator based on the stored feedback. The learning module may be invoked at regular intervals, when sufficient feedback has accumulated to justify a next retuning of the project comparator, or according to other events and timings, such as when feedback from users or application programs indicates that the similarity metrics produced by the project comparator do not well correspond to expectations or external measures of similarity.

The project-search engine that represents one embodiment of the present invention, as discussed above with reference to Figure 7 and Figures 8A-E can, in turn, be used by a number of different project-management applications as well as directly by project managers and other users. In one application, as a project manager begins to develop an electronic-data representation of a project, a project-development tool searches for, and locates similar projects using the project-search engine, to assist the project manager in development of the current project. For example, the project manager may access the electronic-data representations of similar projects in order to use portions of those electronic-data representations in developing the new project. As another example, a project-manager-monitoring application may use the project-search engine to identify projects similar to those projects managed by a particular project manager in order to ascertain the relative success rate for the project manager as well as to identify positive and negative characteristics of the project manager.

Figure 9 illustrates a typical electronic computer on which project search engines, project comparators, and other embodiments of the present invention may be implemented. The computer system contains one or multiple central processing units ("CPUs") 902-905, one or more electronic memories 908 interconnected with the CPUs by a CPU/memory-subsystem bus 910 or multiple busses, a first bridge 912 that interconnects the CPU/memory-subsystem bus 910 with additional busses 914 and 916, or other types of high-speed interconnection media, including multiple, high-speed serial interconnects. These busses or serial interconnections, in turn, connect the CPUs and memory with specialized processors, such as a graphics processor 918, and with one or more additional bridges 920, which are interconnected with high-speed serial links or with multiple controllers 922-927, such as controller 927, that provide access to various different types of mass-storage devices 928, electronic displays, input devices, and other such components, subcomponents, and computational resources. Embodiments of the present invention may also be implemented on distributed computer systems and can also be implemented partially in hardware logic circuitry.

Although the present invention has been described in terms of particular embodiments, it is not intended that the invention be limited to these embodiments. Modifications will be apparent to those skilled in the art. For example, a project comparator subsystem and project-search engine that represent
5 embodiments of the present invention may be implemented in a variety of different ways by varying any of the many well-known implementation parameters, including programming language, operating system, modular organization, data structures, control structures, and other such parameters. As discussed above, any of various different feature comparators may be employed within the project comparator, based
10 on the data types of features and other considerations. The second-level comparator within the project comparator may simply sum weighted similarity metrics produced by the various feature comparators, or may carry out any of various other computational combinations of the results produced by the feature comparators to produce a final project-comparator similarity metric. The project-search engine that
15 represents one embodiment of the present invention may be directly used by human users through a project-search-engine interface or may alternatively be invoked by any of a variety of different project-management applications and tools

The foregoing description, for purposes of explanation, used specific nomenclature to provide a thorough understanding of the invention. However, it will
20 be apparent to one skilled in the art that the specific details are not required in order to practice the invention. The foregoing descriptions of specific embodiments of the present invention are presented for purpose of illustration and description. They are not intended to be exhaustive or to limit the invention to the precise forms disclosed. Many modifications and variations are possible in view of the above teachings. The
25 embodiments are shown and described in order to best explain the principles of the invention and its practical applications, to thereby enable others skilled in the art to best utilize the invention and various embodiments with various modifications as are suited to the particular use contemplated. It is intended that the scope of the invention be defined by the following claims and their equivalents:

CLAIMS

1. A project search system, implemented on one or more computer systems, comprising:
 - two or more stored projects represented by electronically stored, computer-accessible
5 data that are stored in a project-database component;
 - a project-comparator subsystem, executed on the one or more computer systems, that generates a similarity metric for two electronically-represented projects input to the project-comparator subsystem; and
 - a project search engine, executed on the one or more computer systems, that
10 receives data that represents an input project,
identifies, by pairwise comparison of the stored projects with the input project using the project-comparator subsystem, one or more stored projects with greatest similarity to the input project, and
outputs, to a graphical interface displayed on a computer display or to an
15 application program, indications of the one or more stored projects with greatest similarity to the input project.
2. The project search system of claim 1 wherein the project-comparator subsystem comprises:
 - 20 a feature demultiplexer that extracts information from two input projects related to one or more project features;
 - two or more feature comparators that each receives information extracted by the feature demultiplexer from each of the two input projects related to a corresponding project feature, compares the received information, and outputs a feature-similarity metric; and
 - 25 a second-level comparator that receives the feature-similarity metrics output by the two or more feature comparators, applies a weighting factor to each received feature-similarity metric to produce a corresponding weighted feature-similarity metric, and combines the weighted feature-similarity metrics to produce an output project-similarity metric.
- 30 3. The project search system of claim 1 wherein the project-comparator subsystem includes feature comparators that compare extracted text from the two input projects by

creating a term-occurrence vector from the extracted text for each input project and computing a cosine for the angle between the two term-occurrence vectors.

4. The project search system of claim 1 wherein the project-comparator subsystem
5 includes feature comparators that compare extracted numerical values from the two input projects by computing a numerical-value-range overlap between the numerical values extracted from a first of the two input projects and the numerical values extracted from a second of the two input projects, by computing one or more statistical values from the extracted numerical values from the two input projects, including the mean, variance, and
10 standard deviation, and by comparing the one or more statistical values computed from the first of the two input projects with the statistical values computed from the second of the two input projects.

5. The project search system of claim 1 wherein the project-comparator subsystem
15 includes feature comparators that compare a figure or graphical data extracted from the two input projects by rotating and scaling the figure or graphical data extracted from a first of the two input projects with respect to the figure or graphical data extracted from a second of the two input projects in order to compute a metric corresponding to a maximum overlap of features within the figure or graphical data extracted from the first of the two input projects
20 and the figure or graphical data extracted from the second of the two input projects.

6. The project search system of claim 1 wherein the project-comparator subsystem includes feature comparators that compare a combination of two or more types of data extracted from the two input projects including:

25 a figure or graphical data;
numerical values;
text; and
application-program data files.

30 7. The project search system of claim 1 further including a learning module that

receives feedback regarding output of the project search system along with input projects and output of the project search system;

electronically stores the received feedback, input, and output; and

at defined points in time, computes adjustments to the project-comparator subsystem that optimize similarity metrics produced by the project-comparator subsystem with respect to received and stored feedback.

8. The project search system of claim 3 wherein the defined points of time represent one or more of:

10 fixed intervals;

times at which sufficient new feedback has been received to carry out a next project-comparator-subsystem adjustment; and

times at which feedback indicates that similarity metrics produced by the project-comparator subsystem do not correspond to external measures of project similarity.

15

9 A method for identifying projects related to a particular project, the method comprising:

inputting a data representation of the particular project into a project search system, implemented on one or more computer systems, that includes

20 two or more stored projects represented by electronically stored, computer-accessible data that are stored in a project-database component;

a project-comparator subsystem that generates a similarity metric for two electronically-represented projects input to the project-comparator subsystem; and

a project search engine, executed on one or more computer systems, that

25 receives data that represents an input project,

identifies, by pairwise comparison of the stored projects with the input project using the project-comparator subsystem, one or more stored projects with greatest similarity to the input project, and

30 outputs, to a graphical interface displayed on a computer display or to an application program, indications of the one or more stored projects with greatest similarity to the input project; and

analyzing the indications of the one or more stored projects to identify projects most related to a particular project.

10. The method of claim 9 wherein data stored to represent the identified projects are used
5 to facilitate further development and management of the particular project.

11. The method of claim 9 wherein data stored to represent the identified projects are used to compare progress of the particular project with progress of the identified projects, in order evaluate performance of a manager of the particular project..

10

12. The method of claim 9 wherein the project-comparator subsystem comprises:
a feature demultiplexer that extracts information from two input projects related to one or more project features;

15 two or more feature comparators that each receives information extracted by the feature demultiplexer from each of the two input projects related to a corresponding project feature, compares the received information, and outputs a feature-similarity metric; and

20 a second-level comparator that receives the feature-similarity metrics output by the two or more feature comparators, applies a weighting factor to each received feature-similarity metric to produce a corresponding weighted feature-similarity metric, and combines the weighted feature-similarity metrics to produce an output project-similarity metric.

13. The method of claim 9 wherein the project search system further includes a learning module that

25 receives feedback regarding output of the project search system along with input projects and output of the project search system;

electronically stores the received feedback, input, and output; and

at defined points in time, computes adjustments to the project-comparator subsystem that optimize similarity metrics produced by the project-comparator subsystem with respect to received and stored feedback.

30

14. A project-comparator system, implemented within a computer system, that comprises:

a feature demultiplexer, implemented by routines executed by the computer system, that extracts information from two input projects related to one or more project features;

two or more feature comparators, implemented by routines executed by the computer system, that each receives information extracted by the feature demultiplexer from each of the two input projects related to a corresponding project feature, compares the received information, and outputs a feature-similarity metric; and

a second-level comparator, implemented by routines executed by the computer system, that receives the feature-similarity metrics output by the two or more feature comparators, applies a weighting factor to each received feature-similarity metric to produce a corresponding weighted feature-similarity metric, and combines the weighted feature-similarity metrics to produce an output project-similarity metric.

15. The project-comparator system of claim 14 including one or more of:

feature comparators that compare extracted text from the two input projects by creating a term-occurrence vector from the extracted text for each input project and computing a cosine for the angle between the two term-occurrence vectors;

feature comparators that compare extracted numerical values from the two input projects by computing a numerical-value-range overlap between the numerical values extracted from a first of the two input projects and the numerical values extracted from a second of the two input projects, by computing one or more statistical values from the extracted numerical values from the two input projects, including the mean, variance, and standard deviation, and by comparing the one or more statistical values computed from the first of the two input projects with the statistical values computed from the second of the two input projects;

feature comparators that compare a figure or graphical data extracted from the two input projects by rotating and scaling the figure or graphical data extracted from a first of the two input projects with respect to the figure or graphical data extracted from a second of the two input projects in order to compute a metric corresponding to a maximum overlap of features within the figure or graphical data extracted from the first of the two input projects and the figure or graphical data extracted from the second of the two input projects; and

feature comparators that compare a combination of two or more types of data extracted from the two input projects including:

a figure or graphical data;

numerical values;

5 text; and

application-program data files.

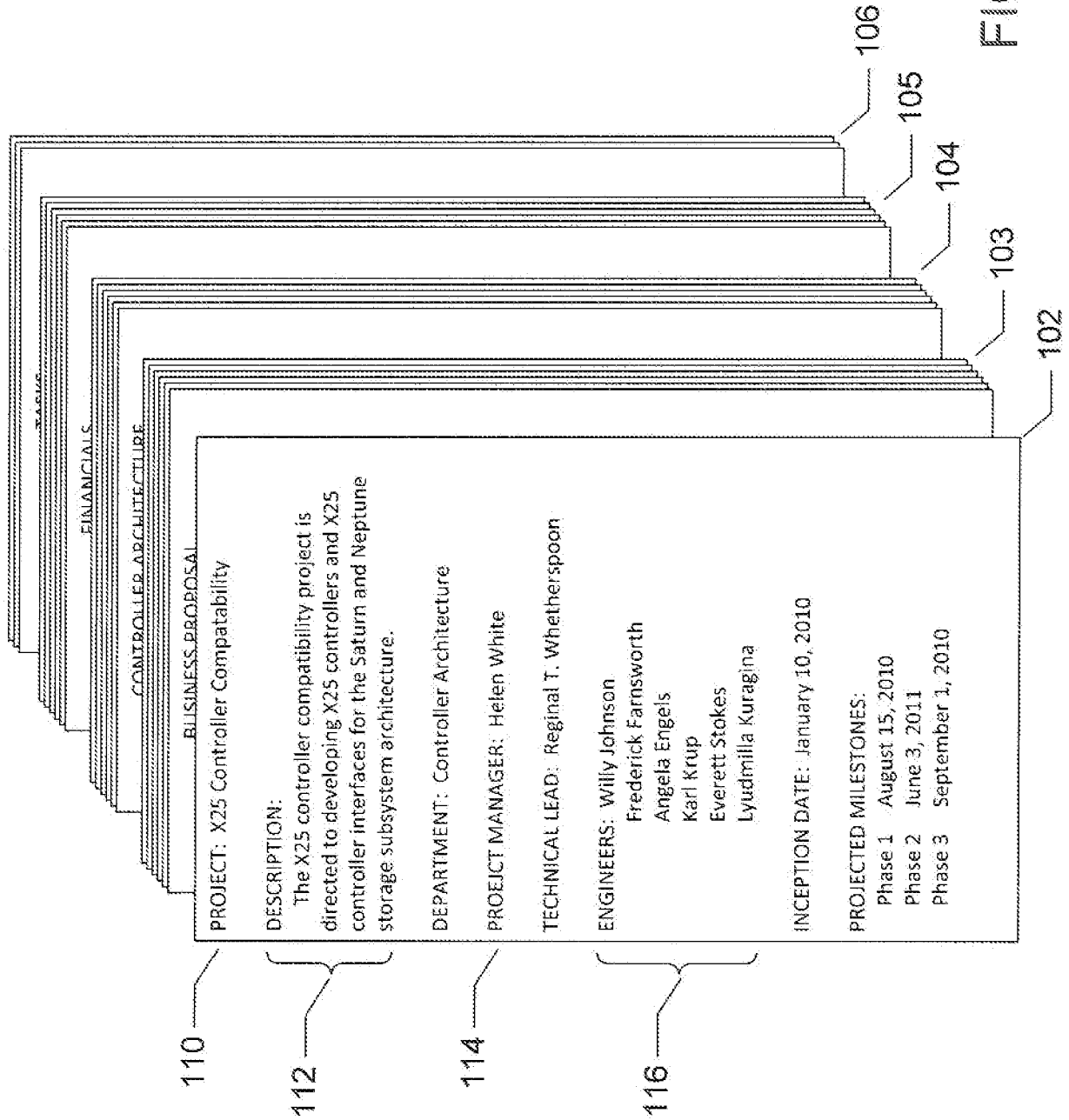
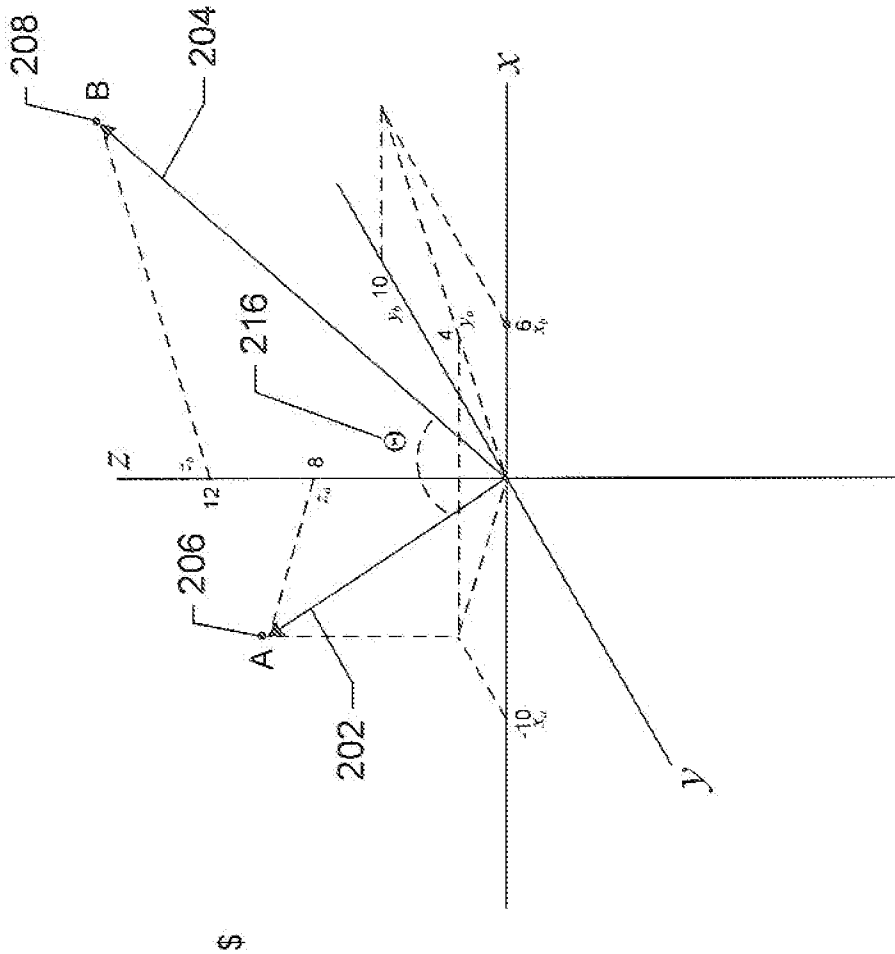


FIGURE 1



$$A = (x_a, y_a, z_a); B = (x_b, y_b, z_b)$$

$$A \cdot B = |A||B| \cos \Theta = x_a x_b + y_a y_b + z_a z_b$$

$$|A| = (A \cdot A)^{1/2} = \sqrt{x_a^2 + y_a^2 + z_a^2}$$

$$|B| = (B \cdot B)^{1/2} = \sqrt{x_b^2 + y_b^2 + z_b^2}$$

$$\cos \Theta = \frac{A \cdot B}{|A||B|} = \frac{x_a x_b + y_a y_b + z_a z_b}{\sqrt{x_a^2 + y_a^2 + z_a^2} \sqrt{x_b^2 + y_b^2 + z_b^2}}$$

$x_a = -10$
 $y_a = 8$
 $z_a = 12$
 $x_b = 6$
 $y_b = 4$
 $z_b = 10$
 $A \cdot B = 76$
 $|A| = 13.416$
 $|B| = 16.733$
 $\cos \Theta = 0.3385$
 $\Theta = 70.2^\circ$

FIGURE 2

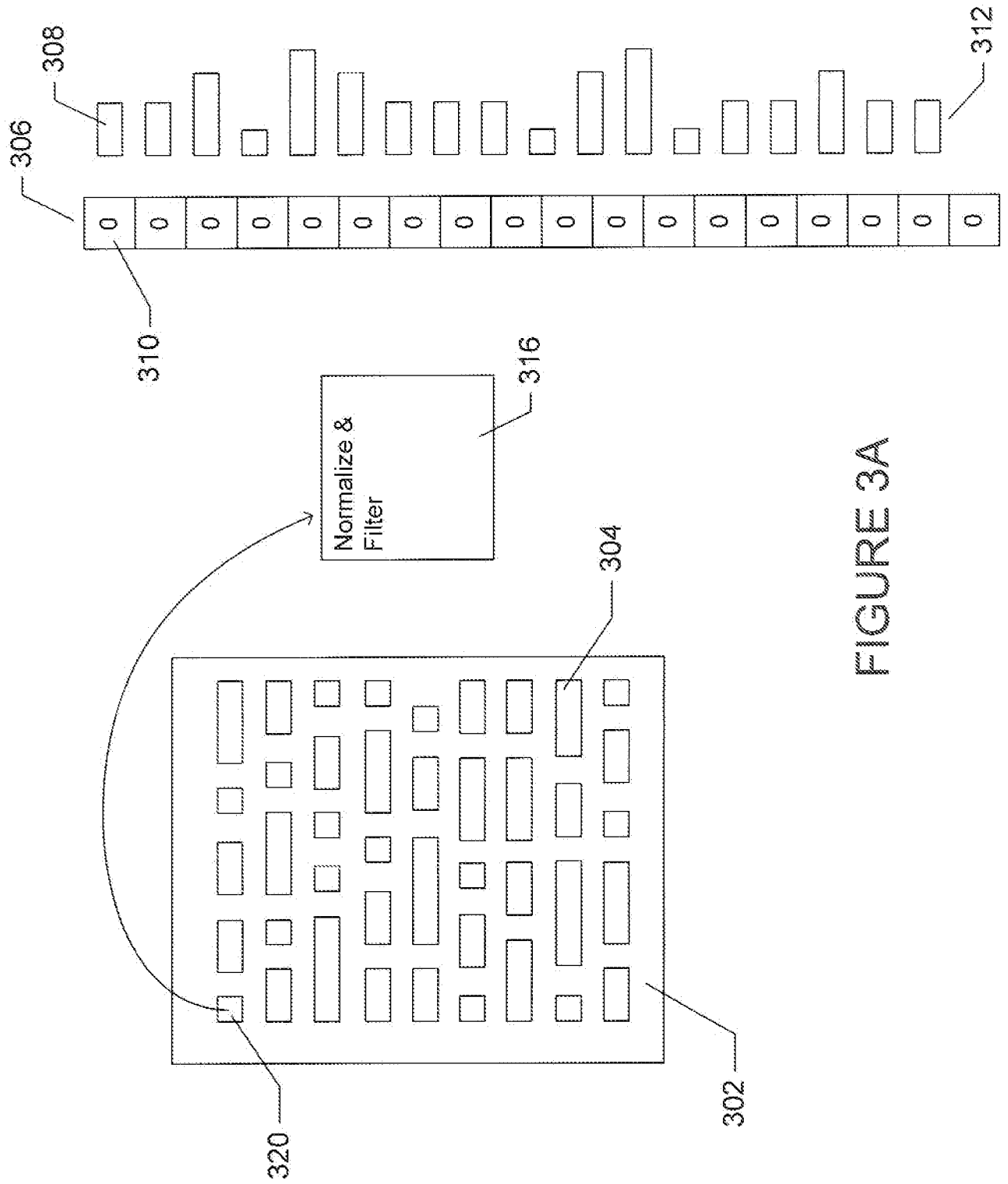


FIGURE 3A

4/17

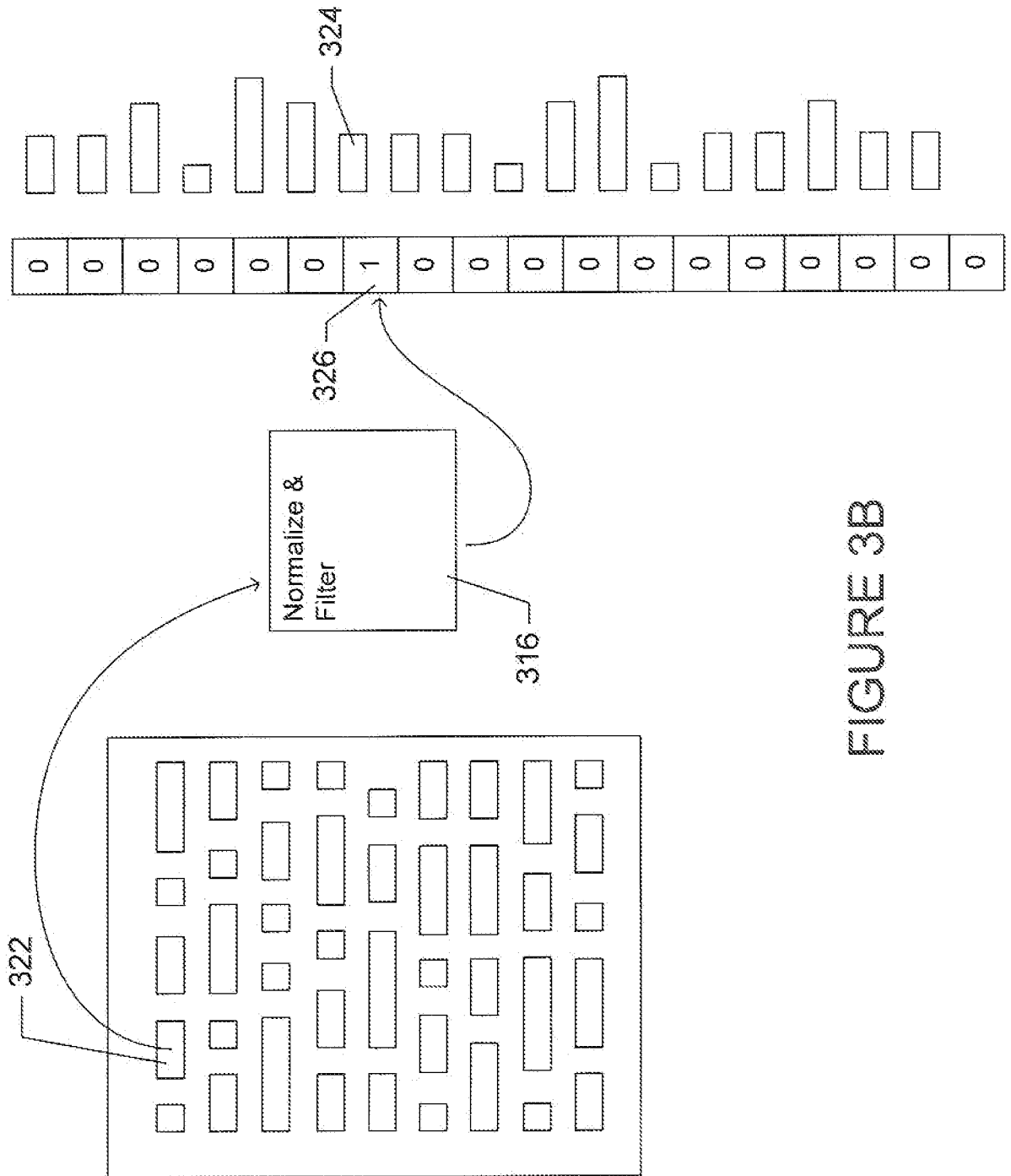


FIGURE 3B

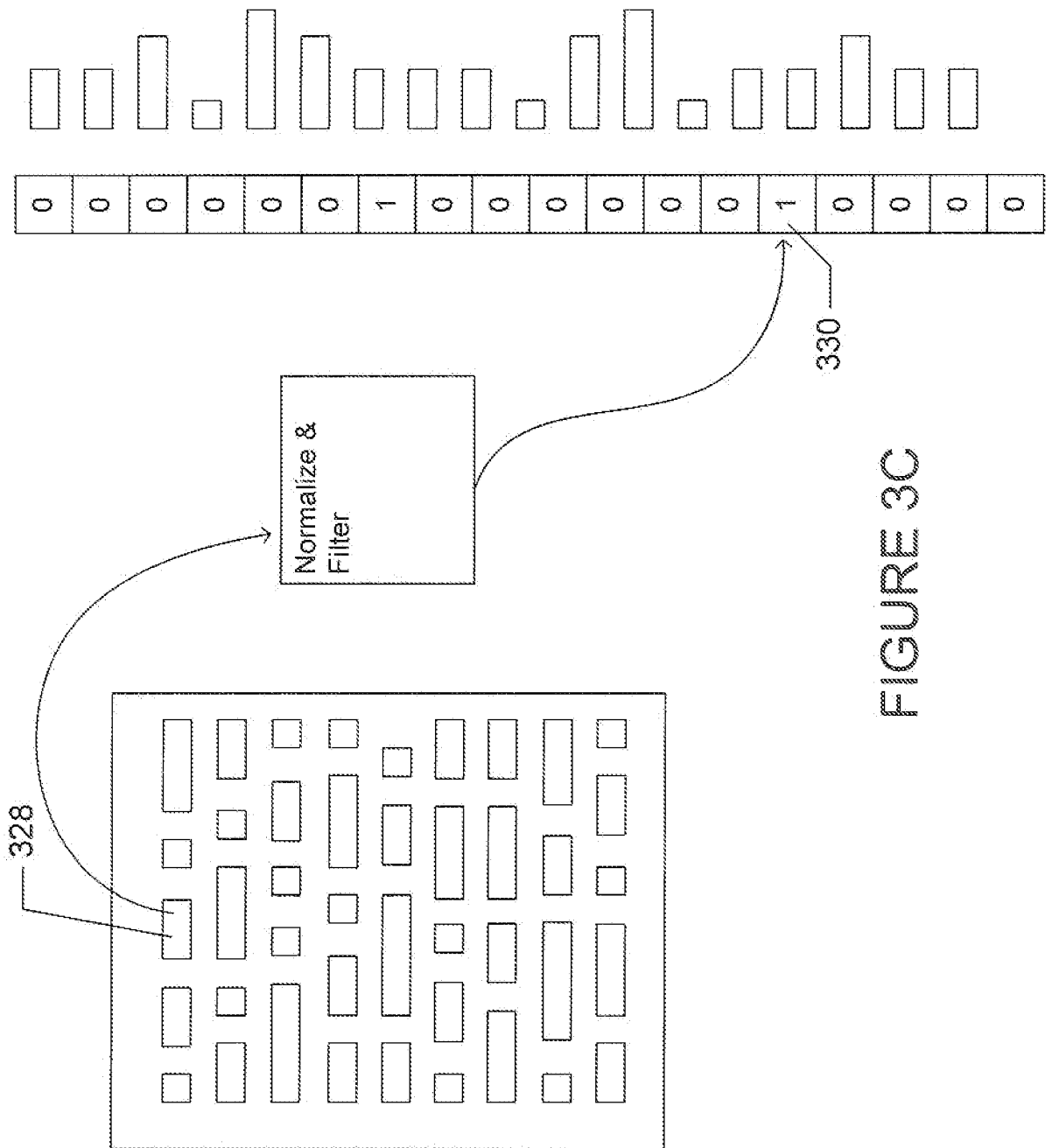


FIGURE 3C

6/17

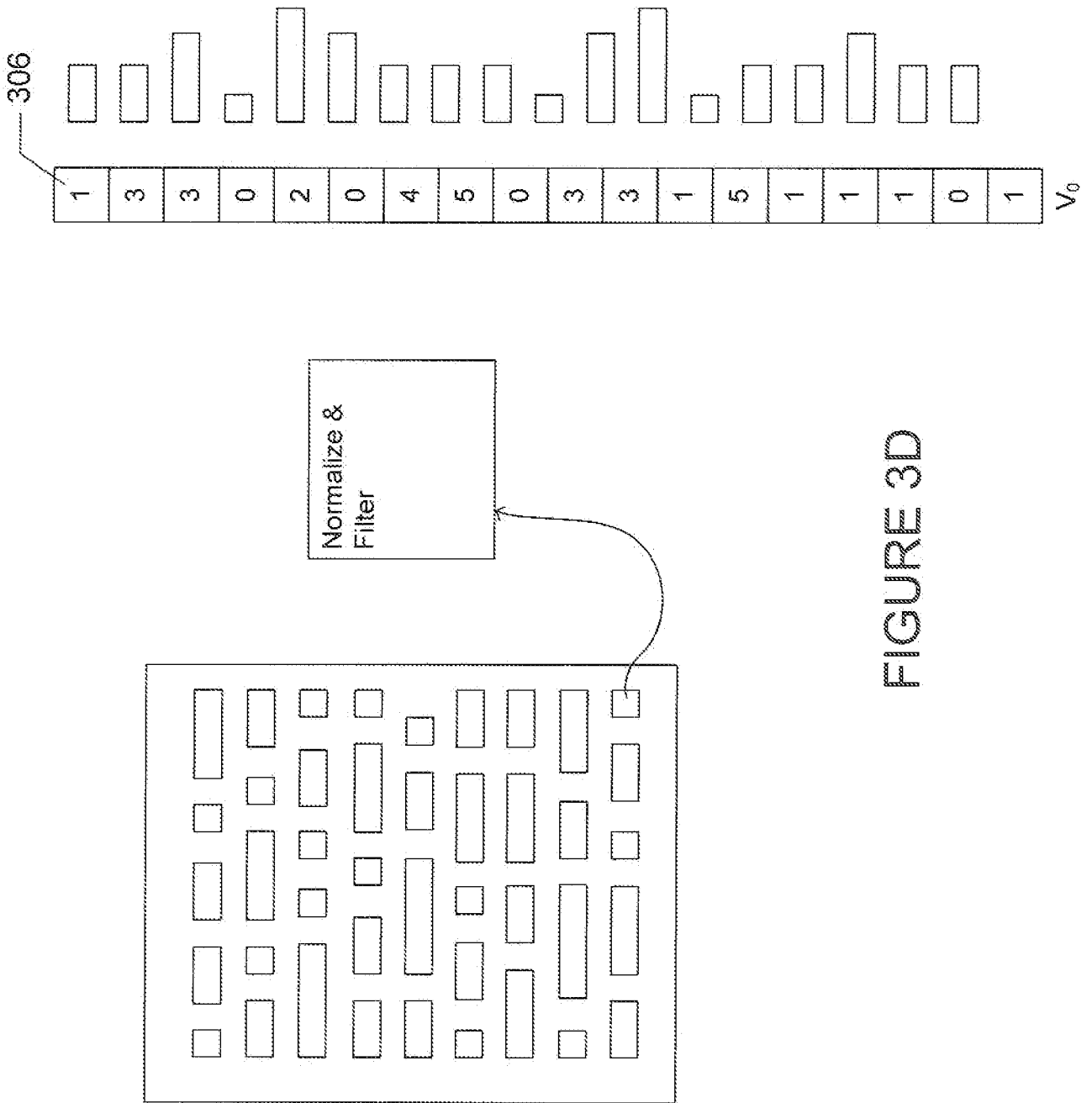


FIGURE 3D

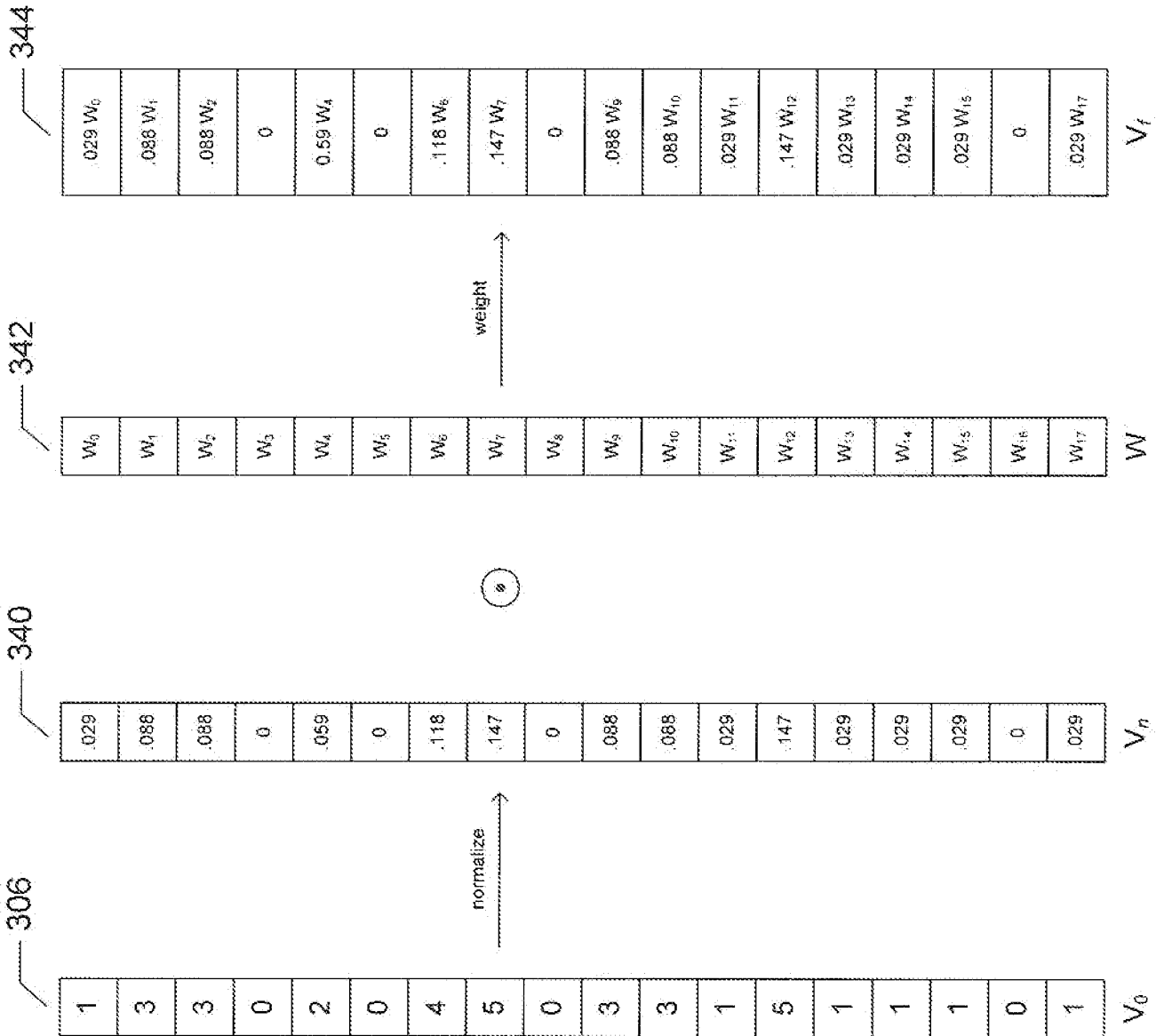


FIGURE 3E

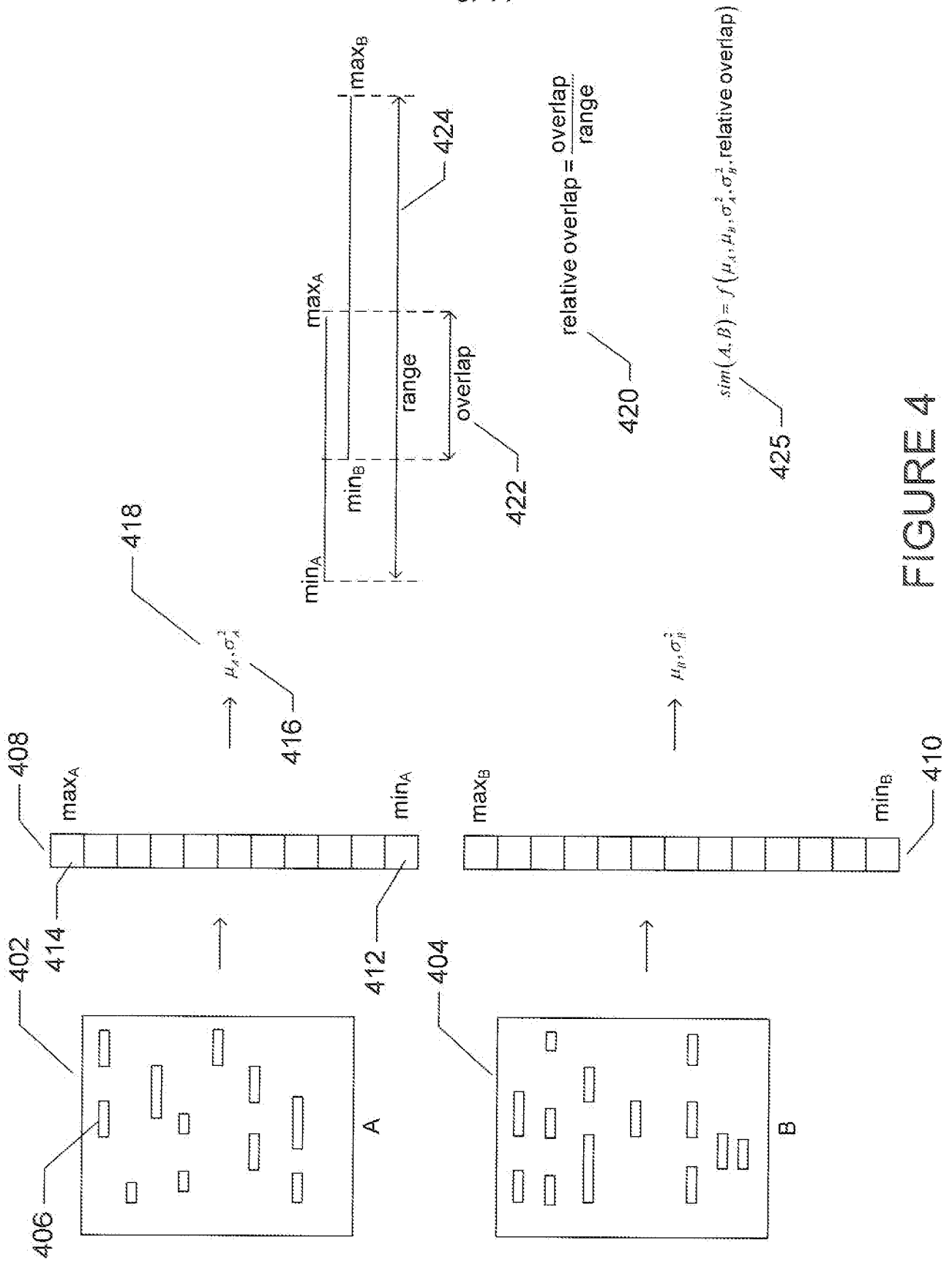


FIGURE 4

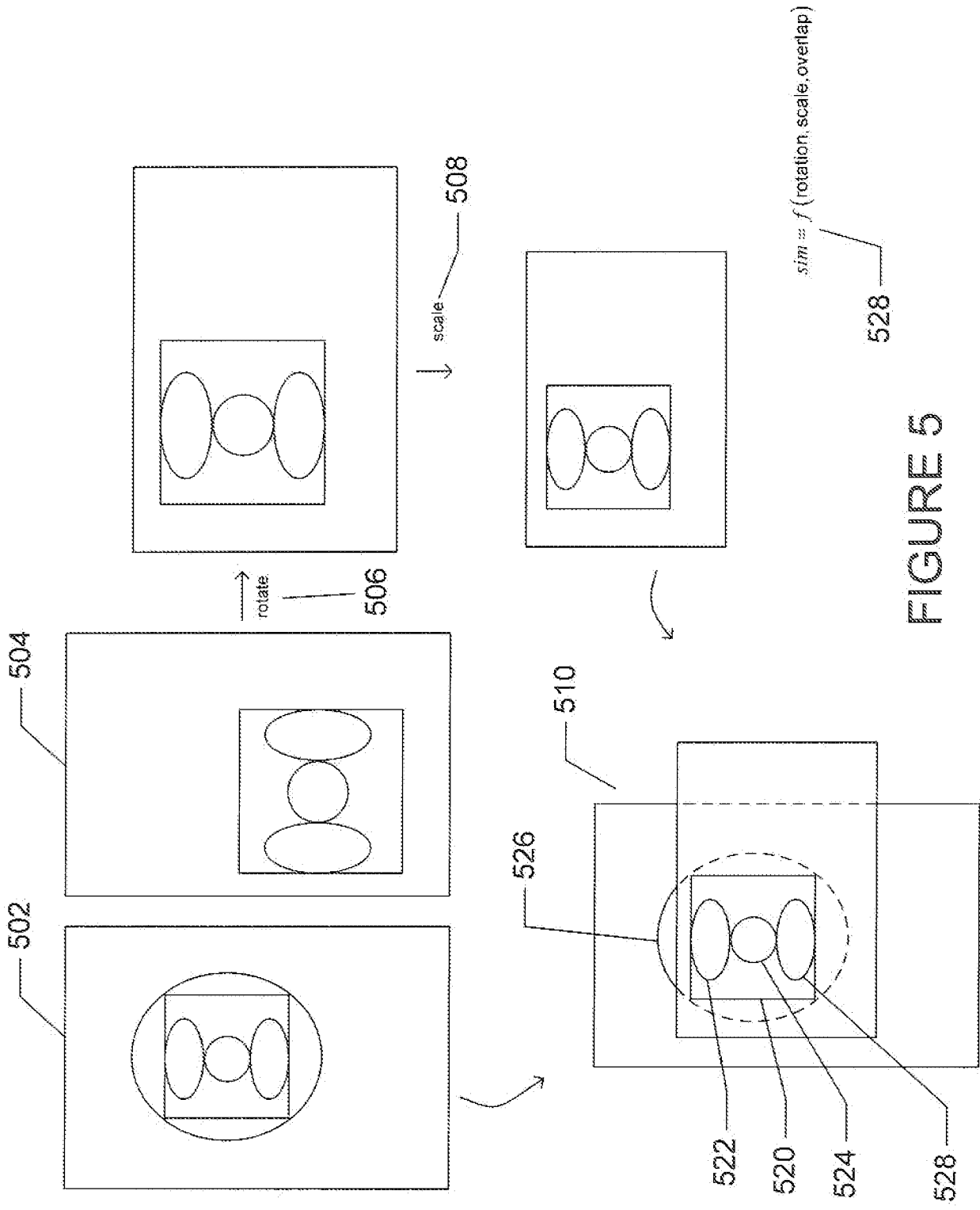


FIGURE 5

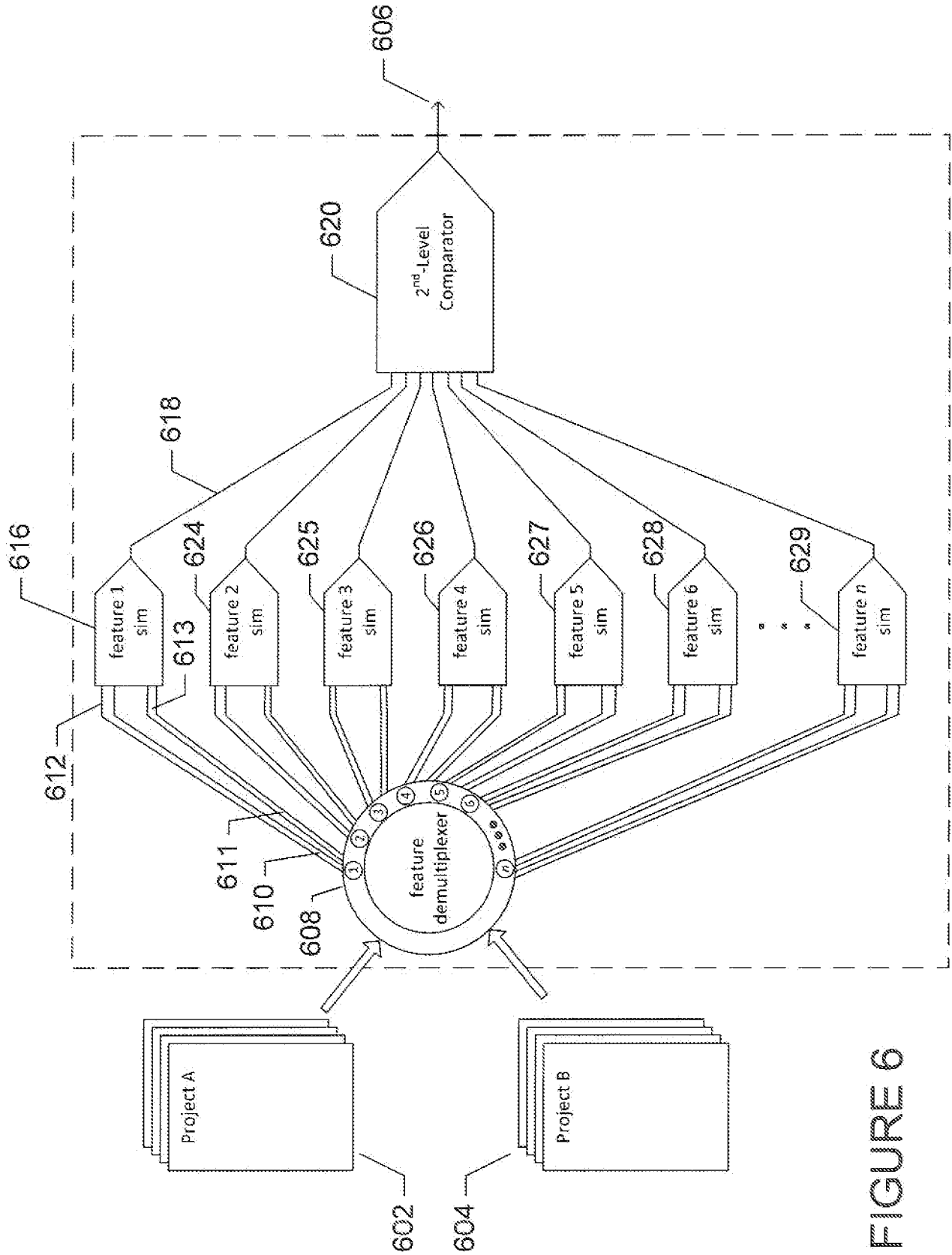


FIGURE 6

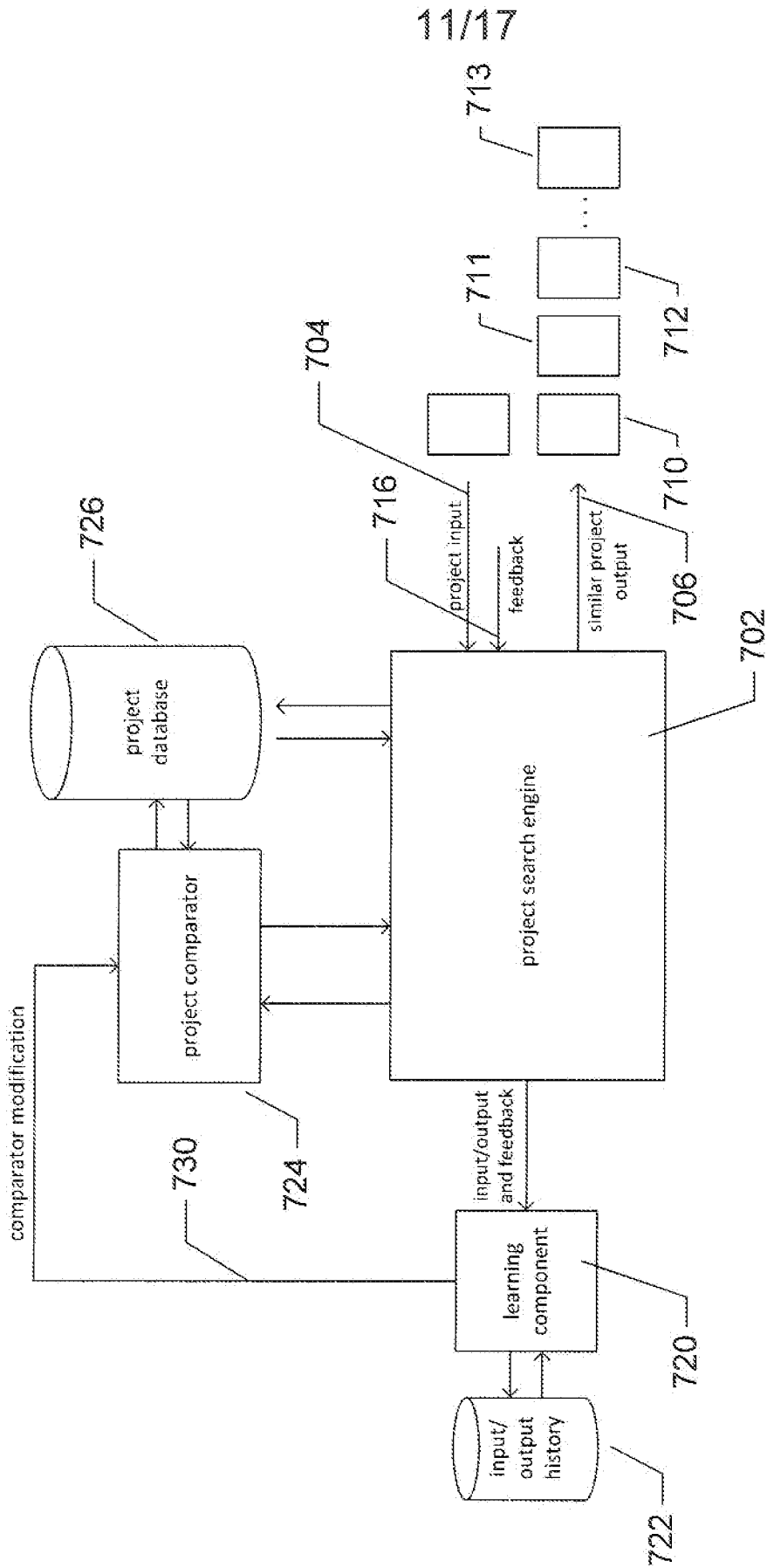


FIGURE 7

12/17

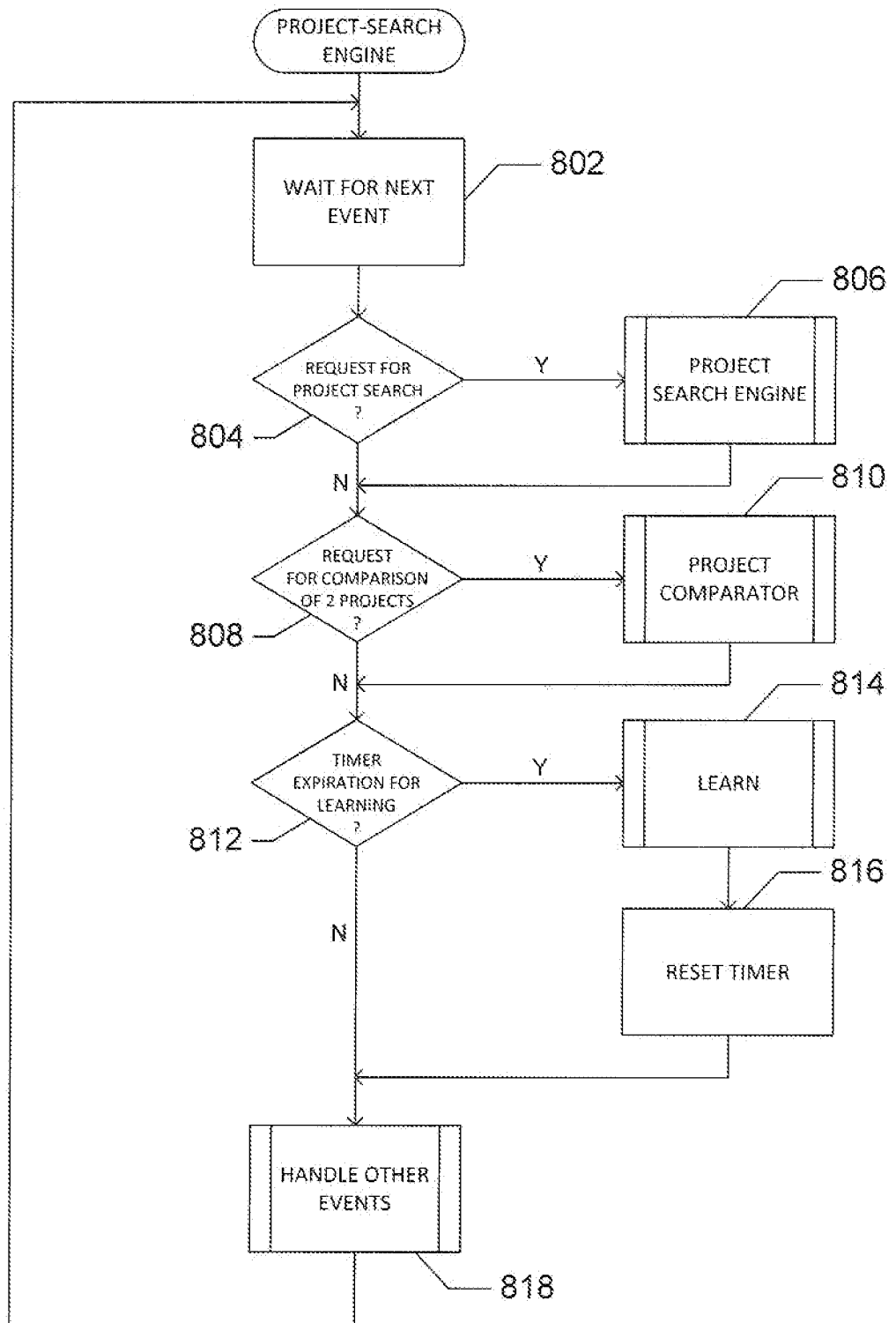


FIGURE 8A

13/17

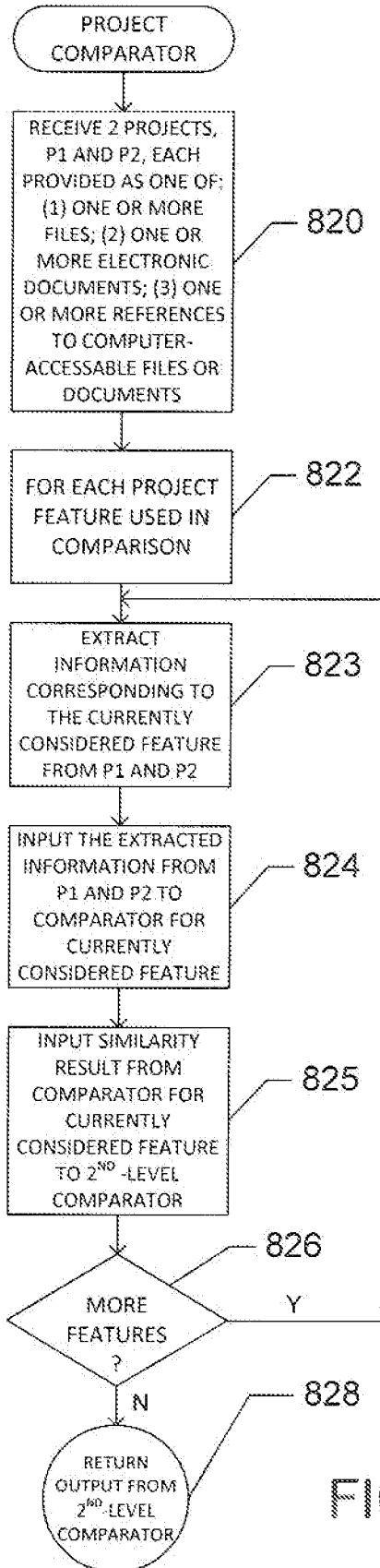


FIGURE 8B

14/17

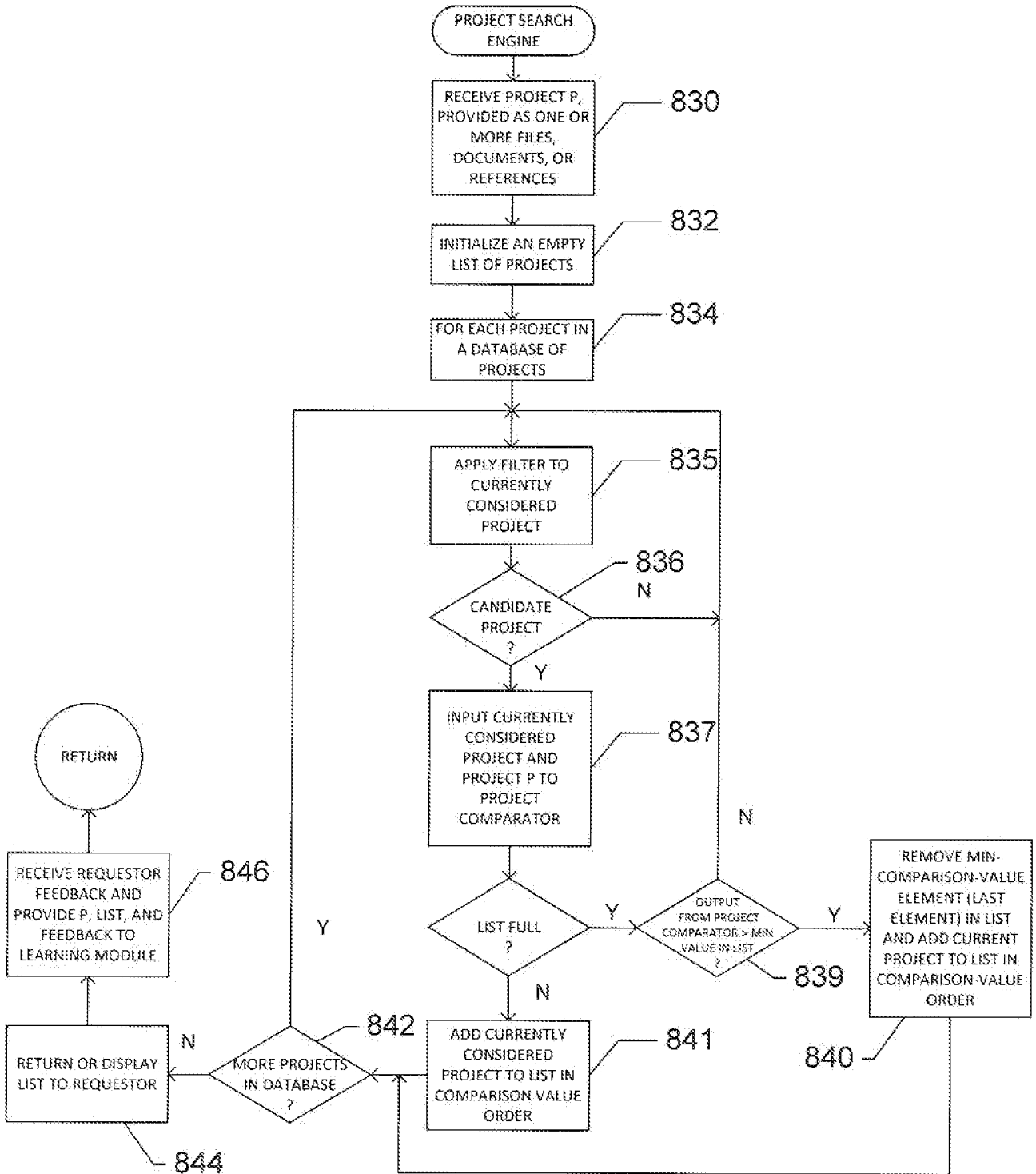


FIGURE 8C

15/17

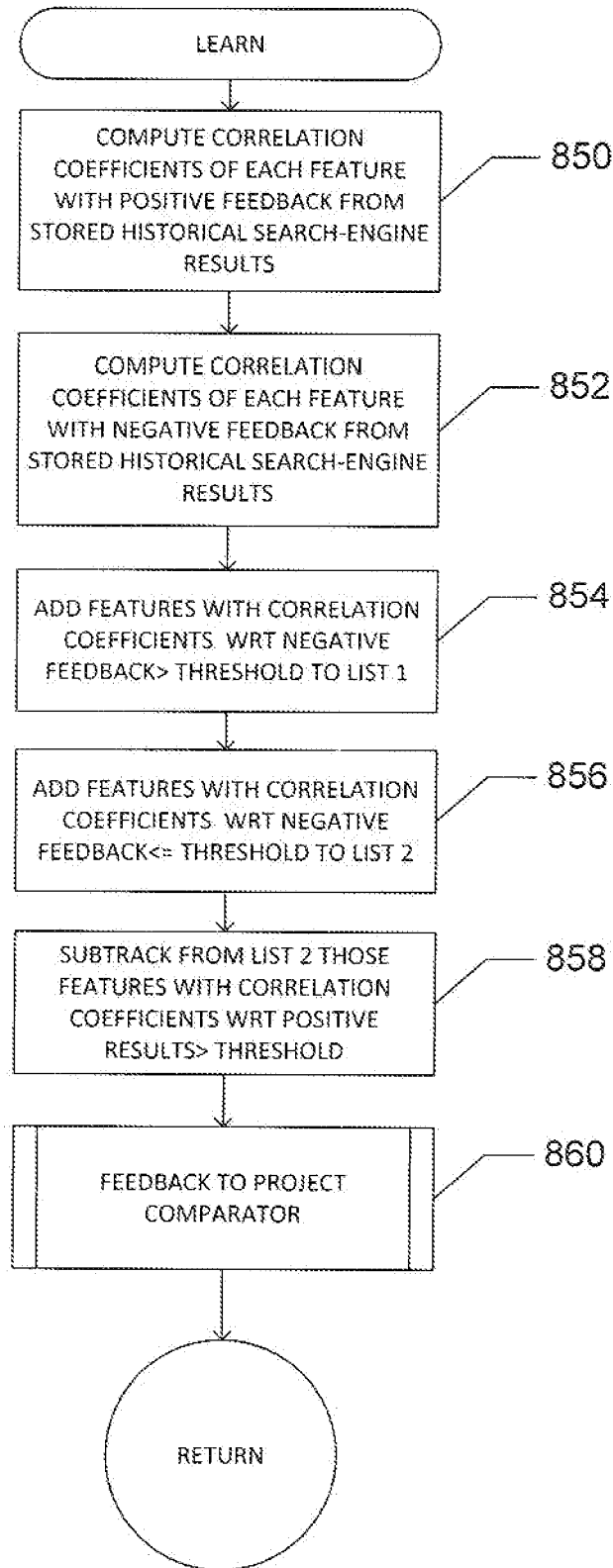


FIGURE 8D

16/17

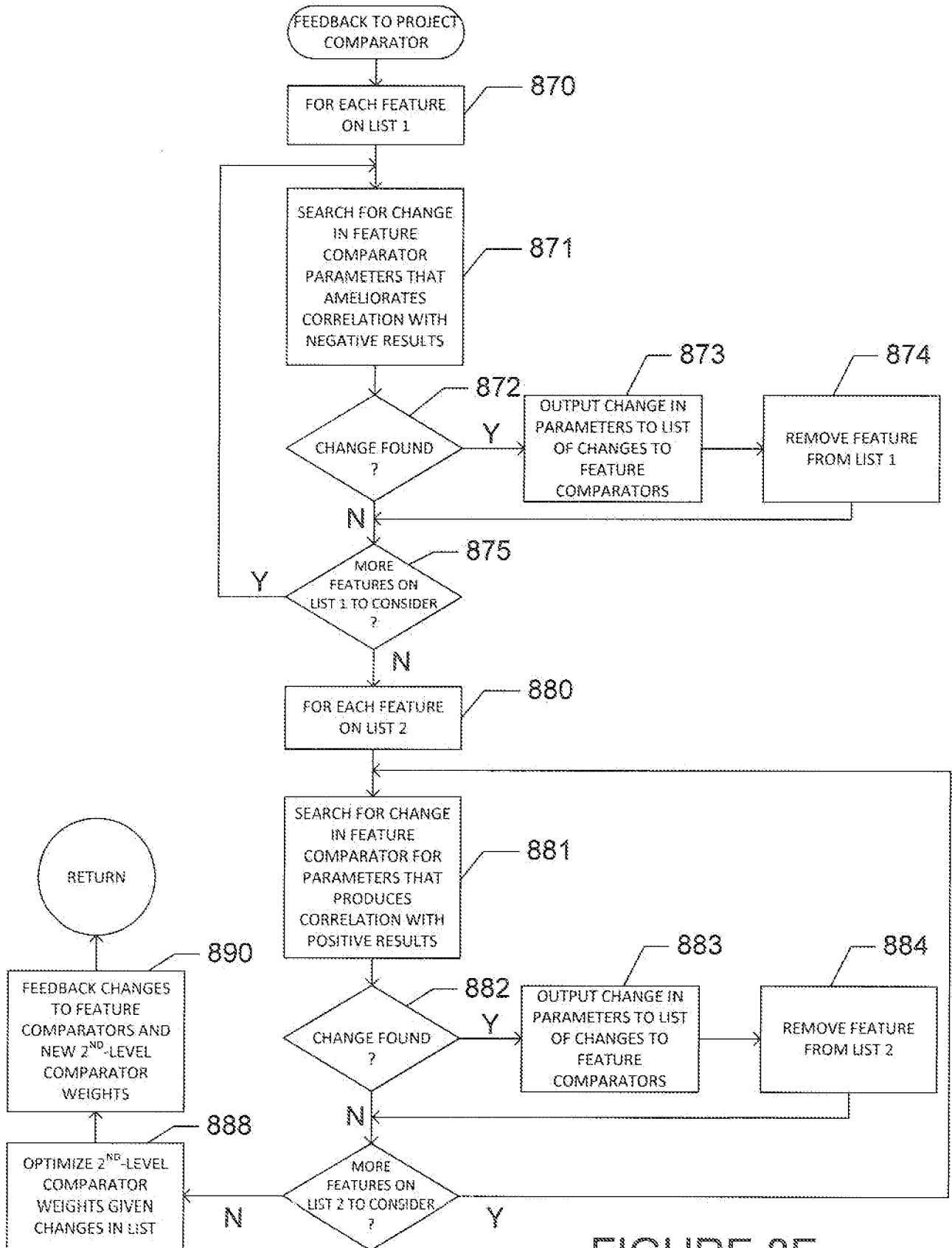
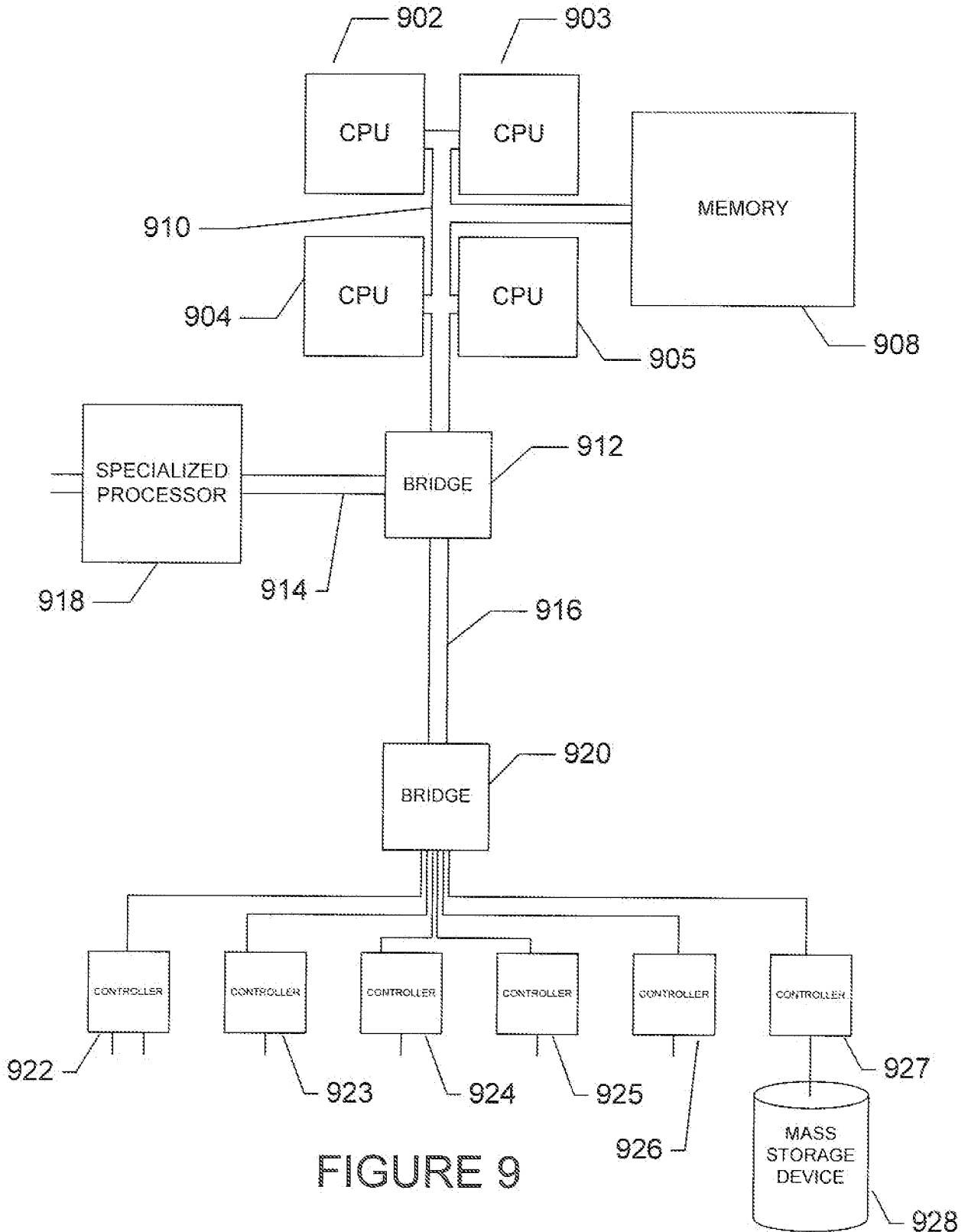


FIGURE 8E

17/17



A. CLASSIFICATION OF SUBJECT MATTER**G06F 17/30(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)

G06F 17/30; G06F 7/08; G06F 17/00; G06F 17/21; G06F 7/04

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: "project, search, compare, comparator, feature, relation"

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Towards a unified approach to document similarity search using manifold-ranking of blocks (Xiaojun Wan et al.) In: Information Processing & Management, 2008, Vol.44, pp.1032-1048. See the whole document.	1-15
A	Finding Similar Documents in Document Collections (Thorsten Brants and Reinhard Stolle) In: Third International Conference on Language Resources and Evaluation (LREC-2002), June 2002 (2002). See the whole document.	1-15
A	US 2008-0319941 A1 (GOLLAPUDI SREENIVAS) 25 December 2008 See abstract, paragraphs 20-42.	1-15
A	US 7158961 B1 (Moses Samson Charikar) 02 January 2007 See abstract, column 2, line 52-column 7, line 35.	1-15
A	KR 10-2009-0039092 A (CONTENTS COMPANY J CO., LTD.) 22 April 2009 See abstract, paragraphs 9-76.	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 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

24 JANUARY 2011 (24.01.2011)

Date of mailing of the international search report

24 JANUARY 2011 (24.01.2011)

Name and mailing address of the ISA/KR

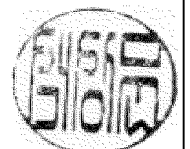
Korean Intellectual Property Office
Government Complex-Daejeon, 139 Seonsa-ro, Seo-gu, Daejeon 302-701, Republic of Korea

Facsimile No. 82-42-472-7140

Authorized officer

Soak, Sang Moon

Telephone No. 82-42-481-8470



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2010/030518

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2008-0319941 A1	25.12.2008	None	
US 7158961 B1	02.01.2007	None	
KR 10-2009-0039092 A	22.04.2009	None	