



(51) International Patent Classification:
G06Q 50/00 (2006.01)

(21) International Application Number:
PCT/US2009/040881

(22) International Filing Date:
16 April 2009 (16.04.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
12/103,712 16 April 2008 (16.04.2008) US

(71) Applicant (for all designated States except US): **MICROSOFT CORPORATION** [US/US]; One Microsoft Way, Redmond, WA 98052-6399 (US).

(72) Inventor: **LAI, Wei**; C/o Microsoft Corporation, Lca, International Patents, One Microsoft Way, Redmond, WA 98052-6399 (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, 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, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, 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, 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, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— without international search report and to be republished upon receipt of that report (Rule 48.2(g))

(54) Title: FORUM WEB PAGE CLUSTERING BASED ON REPETITIVE REGIONS

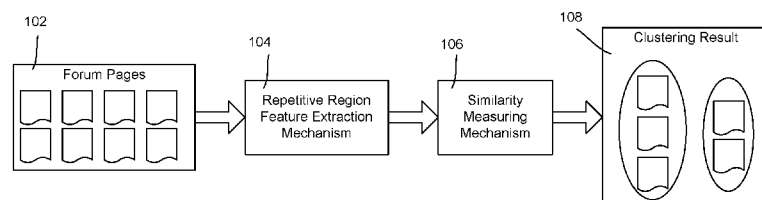


FIG. 1

(57) Abstract: Described is a technology by which forum web pages are processed into clusters for classification purposes, including by determining repetitive regions between pages and associating pages that have similar repetitive regions into a common cluster. Patterns corresponding to the regions are determined, and a feature set based at least in part on those patterns (e.g., pattern frequency) is extracted from the page. The feature set of a page is compared against the feature set of another page to determine similarity therewith, e.g., via a feature space distance computation that is evaluated against a threshold distance.



FORUM WEB PAGE CLUSTERING BASED ON REPETITIVE REGIONS

BACKGROUND

[0001] User created content (UCC) is becoming an important data resource on the Internet. One popular type of user created content, directed towards user discussions, is referred to as a web forum (also named a bulletin board or discussion board). The data of a web forum are becoming very valuable for various web applications. For example, commercial search engines have begun to integrate forum data into their searches to improve the quality of search results. As another example, recent research efforts have tried to mine forum data to obtain useful information, such as business intelligence and expertise. In any such application, a general goal is to fetch data pages from various forum sites distributed over the Internet.

[0002] To download forum data effectively and efficiently, the characteristics of forums need to be understood, which involves understanding the forum pages and relationships between pages. Forum pages tend to be semi-structured, and are typically generated based upon pre-defined templates.

[0003] As a result of the structuring, the pages of a given forum site may be classified into several categories, in which each category represents a specific function. For example, generic forums usually have list-of-board pages, post-of-thread pages, user profile pages, and so forth; to extract post-of-thread content, identification of the post-of-thread pages is required.

[0004] Once classified, page classification may be used in forum page understanding, and for further analysis of forum data. Page classification is also valuable in forum crawling, e.g., page classification is a component used in recovering the structure of the forum site, and determine an optimized route for a crawler. Further, page classification can help filter out invalid pages and reduce duplicate pages; for example, the same pages (or other content) having different Uniform Resource Locators (URLs) are often generated for different requests, such as "view by date" or "view by title" requests.

[0005] To categorize forum pages, URL pattern analysis may be used, particularly with respect to sites hosted by commercial forum service providers. For example, `"*/forumdisplay.php?fid=*"` refers to list-of-post pages,

while `"*/viewthread.php?tid=*`" refers to post-of-thread pages. However, in many cases, a URL is ambiguous and does not reveal a page's function. As one example, professional forums and communities of large enterprises usually define their own forms, whereby, for example, a URL such as `"http://www.wxyz-forums.net/"` provides no readily apparent URL patterns indicative of different types of pages.

[0006] Another technique used in categorizing forum pages utilizes Document Object Model (DOM) tree-based structure criteria to describe target pages. However, using DOM trees in forum sites for categorization does not provide a sufficient and robust solution, as similar pages may have different numbers of advertisements, images, and even complex sub-structures from user posts.

SUMMARY

[0007] This Summary is provided to introduce a selection of representative concepts in a simplified form that are further described below in the Detailed Description. This Summary is not intended to identify key features or essential features of the claimed subject matter, nor is it

intended to be used in any way that would limit the scope of the claimed subject matter.

[0008] Briefly, various aspects of the subject matter described herein are directed towards a technology by which web pages, particularly forum pages for example, are processed into clusters for classification purposes, including by determining repetitive regions in pages and selecting pages with similar repetitive regions for a common cluster. In one aspect, regions on forum web pages are determined, and those regions used to determine which forum web pages have similar regions for clustering purposes. In one implementation, patterns corresponding to the regions are determined, and a feature set based at least in part on those patterns (e.g., pattern frequency) is extracted from the page. The feature set of a page is compared against the feature set of another page to determine similarity therewith, e.g., via a feature space distance computation that is evaluated against a threshold distance.

[0009] In one aspect, a feature set may be based upon frequency of occurrence of repetitive patterns. Region area data, location data and/or size data are other

possible features. Still other features that may be extracted and compared include an ordering of repetitive regions, a number of rows in repetitive regions, a number of columns in the repetitive regions, a URL pattern of links in the repetitive regions, and/or a pattern of anchor texts in the repetitive regions.

[0010] In one example aspect, once the clusters are obtained, a new page may be classified based on evaluation against the clusters. To this end, a feature set extracted from a new page is compared against sets of clustered pages to classify the new page based on feature set similarity to one of the sets of clustered pages. For example, the feature space distance to a cluster center is computed to determine whether a distance threshold is met, and if so, the new page is associated with that cluster.

[0011] Other advantages may become apparent from the following detailed description when taken in conjunction with the drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] The present invention is illustrated by way of example and not limited in the accompanying figures in which like reference numerals indicate similar elements and in which:

[0013] FIGURE 1 is a block diagram representing example components directed towards forum web page clustering based on repetitive regions.

[0014] FIGS. 2A-2D are representations of rendered web pages showing detection of repetitive regions.

[0015] FIG. 3 is a representation of which sets of repetitive regions are alike and not alike based on patterns.

[0016] FIG. 4 is a flow diagram showing example steps of a process to discover repetitive patterns and generate feature descriptions for sampled pages.

[0017] FIG. 5 is a flow diagram showing example steps for categorizing a new page based on existing sets of clustered pages.

[0018] FIG. 6 shows an illustrative example of a computing environment into which various aspects of the present invention may be incorporated.

DETAILED DESCRIPTION

[0019] Various aspects of the technology described herein are generally directed towards using repetitive regions that are present among forum pages to identify different types of forum pages. This takes advantage of the observation that for most forum sites, there is a common characteristic of the pages; forum pages, especially pages containing rich information, have strong patterns with respect to content layout.

[0020] In general, a repetitive region on a web page comprises a block area containing multiple data records in a uniform formation. For example, a region may include a list of thread records, with each record containing fields such as title and author of each thread. As described herein, repetitive regions are typical on forum pages, and significant information such as posts, navigation bars, advertisements, and so forth is shown in repetitive regions.

[0021] A repetitive pattern is generally an abstract representation of the records in a repetitive region; a repetitive region may be treated as an instance of a repetitive pattern on a related page. In one example implementation, a repetitive pattern may be described via a tree structure basically comprising an extended DOM tree with regular expression-like signs for each node. In one example implementation, each different repetitive pattern discovered in a forum site may be indexed with a unique identifier.

[0022] As described herein, the repetitive patterns and regions are processed to identify the category and function of a forum page. A forum page may be characterized by its layout, e.g., by what kinds of repetitive regions it contains, and possibly by where these regions are located on the page. Based on such layout information, the forum pages may be effectively clustered into categories.

[0023] While the examples described herein are generally directed towards using repetitive regions for categorization of forum pages, it is readily understood that this is only one such usage. For example, repetitive region-based categorization may be used with other types of

documents. Further, repetitive region-based categorization may be combined with other categorization techniques, e.g., it may be used in conjunction with URL-based categorization (e.g., for sub-clustering) and/or DOM-based categorization.

[0024] As such, the present invention is not limited to any particular embodiments, aspects, concepts, structures, functionalities or examples described herein. Rather, any of the embodiments, aspects, concepts, structures, functionalities or examples described herein are non-limiting, and the present invention may be used in various ways that provide benefits and advantages in computing and classification in general.

[0025] Turning to FIG. 1, a clustering process automatically groups pages with similar content layout (e.g., generated by the same templates) by comparing their distances in feature space as constructed by repetitive patterns. FIG. 1 shows forum pages 102 that are to be clustered being input into a feature extraction mechanism 104 by which the features of the pages are extracted. Then, the distance between each pair of feature samples are computed by a feature similarity measuring mechanism 106. A clustering result 108 is provided, comprising the pages

clustered into one or more groups based upon the feature distance between the pages.

[0026] FIGS. 2A-2D are representations of four rendered pages, in which components such as post titles, authors, post dates, advertisement links, and navigation bars are displayed in a repetitive manner. The repetitive regions are marked by ellipses, in which like different types of lines (e.g., continuous, dashed, dotted) of the ellipses indicate which regions are similar.

[0027] Although the examples show gray stripes rather than specific text, it is readily apparent from these examples that different types of pages have dissimilar repetitive patterns in their layouts. More particularly, FIGS. 2A and 2B correspond to list-of-post pages 202 and 203, respectively, while FIGS. 2C and 2D correspond to post-of-thread pages 204 and 205, respectively.

[0028] FIG. 3 is a representation of how the regions are used in determining whether pages are similar. In FIG. 3, the ellipse patterns overlaid on pages 202 and 203 (of FIGS. 2A and 2B) are each represented as large-dashed, dotted, solid; those of pages 204 and 205 (of FIGS. 2C and 2D) are each represented as small-dashed, dashed/dotted mix, solid.

By comparing the repetitive regions as indicated by the line-patterned ellipses, pages 202 and 203 are deemed similar to each other, while being distinguished from pages 204 and 205, (which themselves are deemed similar to each other).

[0029] Note that in FIGS. 2A and 2B, the page contains two distinct banner images of advertisements, however because each page has one such banner image, the advertisements are not marked as a repetitive region; in one example implementation, repetitive regions contain at least two elements.

[0030] In general, page classification based upon repetitive regions is often more robust than using DOM tree-based classification. For example, the DOM tree structure of a post-of-thread page having only two posts is quite different from that of another page having twenty posts, and thus such pages may not be considered alike. However, two such two pages do have the same kind of a repetitive region, namely a region comprising posts, whereby repetitive region processing may properly identify such otherwise different kinds of pages.

[0031] Turning to feature extraction, a descriptor may be defined to represent the repetitive region information in a feature space, which facilitates quantitative computing. Based on this, a similarity measure between two pages may be defined. In the similarity measure, various factors of repetitive regions, such as the location of the regions, size of the regions, and the occurrence frequency of the repetitive patterns may be considered in making an effective comparison of pages. Note that in one example implementation, the order of the repetitive regions is not used as a feature. However, in alternative implementations, order information may be among the features of a page, and used as part of the distance measure computation.

[0032] Additional features may be used in the distance measurement, such as location of the repetitive regions, e.g., including left, right, top and bottom positions. Others feature may be based on the number of rows and/or columns in the repetitive regions, a URL pattern of links, and/or a pattern of anchor texts in the repetitive regions.

[0033] An example process of repetitive pattern-based feature extraction is represented in the flow diagram of

FIG. 4, which also generally corresponds to the pseudo-code below:

```

input: a set of sampled Web pages  $\mathbb{S}$  and a threshold  $\epsilon$ 
output: 1) a list of discovered repetitive patterns
 $\mathbb{P}$ ; and
           2) a set of feature descriptions  $\mathbb{F}$  for pages
in  $\mathbb{S}$ 

begin
   $\mathbb{P} = \phi$ ;  $\mathbb{F} = \phi$ ;
  foreach  $s \in \mathbb{S}$  do
     $\mathbb{R} = \text{RepetitiveRegionDetection}(s)$ ;
    foreach  $r \in \mathbb{R}$  do
      create an empty  $\mathbf{p}^* = \{p_{tree}^*, p_{area}^*, p_{support}^*\}$ 
       $p_{tree}^* = \text{AlignRecordsInRegion}(r)$ ;
       $p_{area}^* = \text{area of } r / \text{screen area}$ ;  $p_{support}^* = 1$ ;
      foreach  $\mathbf{p}^i \in \mathbb{P}$  do
        if  $\text{TreeAlignmentCost}(p_{tree}^*, p_{tree}^i) < \epsilon$ 
          then  $\mathbf{p}^i = \text{Update}(\mathbf{p}^i, \mathbf{p}^*)$  and break;
        end // foreach  $\mathbf{p}^i \in \mathbb{P}$  do loop
        if cannot align  $\mathbf{p}^*$  with any  $\mathbf{p}^i \in \mathbb{P}$  then  $\mathbb{P} \leftarrow \mathbf{p}^*$ ;
      end // foreach  $r \in \mathbb{R}$  do loop
       $\mathbf{f} = [n^1, n^2, \dots, n^{|\mathbb{P}|}]$ ,  $n^i = \text{number of } \mathbf{p}^i \text{ occurs in } s$ ;
       $\mathbb{F} \leftarrow \mathbf{f}$ ;
    end // foreach  $s \in \mathbb{S}$  do loop
  end // for begin

```

Note that the *RepetitiveRegionDetection*,

AlignRecordsInRegion, and *TreeAlignmentCost* algorithms are described by Y. Zhai and B. Liu in the reference

"Structured data extraction from the Web based on partial tree alignment," IEEE Trans. Knowl. Data Eng.,

18(12):1614-1628, Dec. 2006; and by S. Zheng, R. Song, J.-R.

Wen, and D. Wu in the reference "Joint optimization of

wrapper generation and template detection" In Proc. 13th KDD, pages 894–902, San Jose, CA, USA, Aug. 2007.

[0034] In general, the output of this process generates a pattern for every repetitive region r in every page s , and investigates whether the pattern being evaluated (p^*) can be matched with any existing pattern (p^i) in the list $\mathbb{P}$. Step 402 represents initializing the sets to null sets and selecting a page, with step 404 detecting the regions and step 406 selecting a region. Step 408 creates an empty pattern structure for filling based on tree alignment, area and number of page data $\{p_{tree}^*, p_{area}^*, p_{support}^*\}$. Note that data records in a repetitive region may be extracted by aligning the DOM tree with one or more corresponding repetitive patterns. The records may be stored, e.g., with link-related fields retained to construct a link table for each repetitive region.

[0035] The features only partially represent a page's characteristics. More particularly, different patterns have different capabilities in distinguishing forum pages; patterns tend to be important to users if they have large rendering sizes on the screen, as users tend to pay attention to salient blocks on a page and ignore small ones.

Also popular patterns are unimportant; note that patterns appearing on all of the pages are removed before clustering.

[0036] Because of differences in patterns, in one implementation two parameters, namely p_{area}^i and $p_{support}^i$, are used to describe a pattern besides the tree structure p_{tree}^i . In one example implementation, the rendering information is based upon known Vision-based Page Segmentation (VIPS). Then, p_{area}^i is the average area ratio of the repetitive regions generated by p^i , and $p_{support}^i$ is the number of pages having such repetitive regions. These p_{area}^i are normalized to sum to one, and are taken as weights of patterns in distance measure. Moreover, similar to the well-known "term frequency-inverse document frequency" (or TF×IDF) in text retrieval, the page feature f is further revised by integrating $p_{support}^i$, to lessen the influence of those frequent (popular) patterns, as:

$$f_i = (n^i / \sum_{j=1}^{\|\mathbb{P}\|} n^j) \times \log_{10}(\|\mathbb{S}\| / p_{support}^i) \quad 1)$$

The distance between two pages s_a and s_b is defined as:

$$dist(s_a, s_b) = \sqrt{\sum_{i=1}^{\|\mathbb{P}\|} p_{area}^i \times (f_a^i - f_b^i)^2} \quad 2)$$

As there is no prior knowledge about how many kinds of pages may exist in the target forum, in clustering, a single linkage algorithm is utilized to agglomerate these pages in a hierarchical way. The agglomeration is stopped when the minimum distance between pages of each cluster is larger than a pre-defined threshold.

[0037] Returning to FIG. 4, step 410 selects a pattern p^* and tries to match it to an existing pattern p^i based on alignment cost (step 412). If matched with an existing pattern (as repeatedly attempted via steps 416 and 418), the existing pattern p^i is updated with p^* at step 414; otherwise p^* is added to $\mathbb{P}$ as a new pattern at step 422.

[0038] When the repetitive regions processed following steps 424 and 426, at step 428 the feature description f is created for this page s by recording n^i , comprising the number of existing patterns p^i that occur in s , considering that in a page there may be several regions that are generated by the same pattern.

[0039] Once the clusters have been decided as in the example steps of FIG. 4, the clusters may be used to classify a new page, such as in online crawling. FIG. 5 is a flow diagram showing example steps that may be taken to

do so, beginning at step 502 where a new page is located, followed by step 504 where the feature set f is extracted for the new page. A cluster is selected at step 506, and at step 508 the distance in feature space is compared to the cluster center. If at step 510 the minimum distance is less than a clustering threshold distance, the new page is classified into the corresponding cluster at step 512. If not, the process repeats (steps 514 and 516) with each other cluster until the page is classified or no more clusters remain, in which event the classification attempt fails and this page is discarded at step 518.

EXEMPLARY OPERATING ENVIRONMENT

[0040] FIG. 6 illustrates an example of a suitable computing and networking environment 600 in which the examples of FIGS. 1-5 may be implemented. The computing system environment 600 is only one example of a suitable computing environment and is not intended to suggest any limitation as to the scope of use or functionality of the invention. Neither should the computing environment 600 be interpreted as having any dependency or requirement

relating to any one or combination of components illustrated in the exemplary operating environment 600.

[0041] The invention is operational with numerous other general purpose or special purpose computing system environments or configurations. Examples of well known computing systems, environments, and/or configurations that may be suitable for use with the invention include, but are not limited to: personal computers, server computers, handheld or laptop devices, tablet devices, multiprocessor systems, microprocessor-based systems, set top boxes, programmable consumer electronics, network PCs, minicomputers, mainframe computers, distributed computing environments that include any of the above systems or devices, and the like.

[0042] The invention may be described in the general context of computer-executable instructions, such as program modules, being executed by a computer. Generally, program modules include routines, programs, objects, components, data structures, and so forth, which perform particular tasks or implement particular abstract data types. The invention may also be practiced in distributed computing environments where tasks are performed by remote

processing devices that are linked through a communications network. In a distributed computing environment, program modules may be located in local and/or remote computer storage media including memory storage devices.

[0043] With reference to FIG. 6, an exemplary system for implementing various aspects of the invention may include a general purpose computing device in the form of a computer 610. Components of the computer 610 may include, but are not limited to, a processing unit 620, a system memory 630, and a system bus 621 that couples various system components including the system memory to the processing unit 620. The system bus 621 may be any of several types of bus structures including a memory bus or memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures. By way of example, and not limitation, such architectures include Industry Standard Architecture (ISA) bus, Micro Channel Architecture (MCA) bus, Enhanced ISA (EISA) bus, Video Electronics Standards Association (VESA) local bus, and Peripheral Component Interconnect (PCI) bus also known as Mezzanine bus.

[0044] The computer 610 typically includes a variety of computer-readable media. Computer-readable media can be

any available media that can be accessed by the computer 610 and includes both volatile and nonvolatile media, and removable and non-removable media. By way of example, and not limitation, computer-readable media may comprise computer storage media and communication media. Computer storage media includes volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information such as computer-readable instructions, data structures, program modules or other data. Computer storage media includes, but is not limited to, RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical disk storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to store the desired information and which can be accessed by the computer 610. Communication media typically embodies computer-readable instructions, data structures, program modules or other data in a modulated data signal such as a carrier wave or other transport mechanism and includes any information delivery media. The term "modulated data signal" means a signal that has one or more of its characteristics set or changed in such a manner as to encode information in the

signal. By way of example, and not limitation, communication media includes wired media such as a wired network or direct-wired connection, and wireless media such as acoustic, RF, infrared and other wireless media. Combinations of the any of the above may also be included within the scope of computer-readable media.

[0045] The system memory 630 includes computer storage media in the form of volatile and/or nonvolatile memory such as read only memory (ROM) 631 and random access memory (RAM) 632. A basic input/output system 633 (BIOS), containing the basic routines that help to transfer information between elements within computer 610, such as during start-up, is typically stored in ROM 631. RAM 632 typically contains data and/or program modules that are immediately accessible to and/or presently being operated on by processing unit 620. By way of example, and not limitation, FIG. 6 illustrates operating system 634, application programs 635, other program modules 636 and program data 637.

[0046] The computer 610 may also include other removable/non-removable, volatile/nonvolatile computer storage media. By way of example only, FIG. 6 illustrates

a hard disk drive 641 that reads from or writes to non-removable, nonvolatile magnetic media, a magnetic disk drive 651 that reads from or writes to a removable, nonvolatile magnetic disk 652, and an optical disk drive 655 that reads from or writes to a removable, nonvolatile optical disk 656 such as a CD ROM or other optical media. Other removable/non-removable, volatile/nonvolatile computer storage media that can be used in the exemplary operating environment include, but are not limited to, magnetic tape cassettes, flash memory cards, digital versatile disks, digital video tape, solid state RAM, solid state ROM, and the like. The hard disk drive 641 is typically connected to the system bus 621 through a non-removable memory interface such as interface 640, and magnetic disk drive 651 and optical disk drive 655 are typically connected to the system bus 621 by a removable memory interface, such as interface 650.

[0047] The drives and their associated computer storage media, described above and illustrated in FIG. 6, provide storage of computer-readable instructions, data structures, program modules and other data for the computer 610. In FIG. 6, for example, hard disk drive 641 is illustrated as storing operating system 644, application programs 645,

other program modules 646 and program data 647. Note that these components can either be the same as or different from operating system 634, application programs 635, other program modules 636, and program data 637. Operating system 644, application programs 645, other program modules 646, and program data 647 are given different numbers herein to illustrate that, at a minimum, they are different copies. A user may enter commands and information into the computer 610 through input devices such as a tablet, or electronic digitizer, 664, a microphone 663, a keyboard 662 and pointing device 661, commonly referred to as mouse, trackball or touch pad. Other input devices not shown in FIG. 6 may include a joystick, game pad, satellite dish, scanner, or the like. These and other input devices are often connected to the processing unit 620 through a user input interface 660 that is coupled to the system bus, but may be connected by other interface and bus structures, such as a parallel port, game port or a universal serial bus (USB). A monitor 691 or other type of display device is also connected to the system bus 621 via an interface, such as a video interface 690. The monitor 691 may also be integrated with a touch-screen panel or the like. Note that the monitor and/or touch screen panel can be

physically coupled to a housing in which the computing device 610 is incorporated, such as in a tablet-type personal computer. In addition, computers such as the computing device 610 may also include other peripheral output devices such as speakers 695 and printer 696, which may be connected through an output peripheral interface 694 or the like.

[0048] The computer 610 may operate in a networked environment using logical connections to one or more remote computers, such as a remote computer 680. The remote computer 680 may be a personal computer, a server, a router, a network PC, a peer device or other common network node, and typically includes many or all of the elements described above relative to the computer 610, although only a memory storage device 681 has been illustrated in FIG. 6. The logical connections depicted in FIG. 6 include one or more local area networks (LAN) 671 and one or more wide area networks (WAN) 673, but may also include other networks. Such networking environments are commonplace in offices, enterprise-wide computer networks, intranets and the Internet.

[0049] When used in a LAN networking environment, the computer 610 is connected to the LAN 671 through a network interface or adapter 670. When used in a WAN networking environment, the computer 610 typically includes a modem 672 or other means for establishing communications over the WAN 673, such as the Internet. The modem 672, which may be internal or external, may be connected to the system bus 621 via the user input interface 660 or other appropriate mechanism. A wireless networking component 674 such as comprising an interface and antenna may be coupled through a suitable device such as an access point or peer computer to a WAN or LAN. In a networked environment, program modules depicted relative to the computer 610, or portions thereof, may be stored in the remote memory storage device. By way of example, and not limitation, FIG. 6 illustrates remote application programs 685 as residing on memory device 681. It may be appreciated that the network connections shown are exemplary and other means of establishing a communications link between the computers may be used.

[0050] An auxiliary subsystem 699 (e.g., for auxiliary display of content) may be connected via the user interface 660 to allow data such as program content, system status

and event notifications to be provided to the user, even if the main portions of the computer system are in a low power state. The auxiliary subsystem 699 may be connected to the modem 672 and/or network interface 670 to allow communication between these systems while the main processing unit 620 is in a low power state.

CONCLUSION

[0051] While the invention is susceptible to various modifications and alternative constructions, certain illustrated embodiments thereof are shown in the drawings and have been described above in detail. It should be understood, however, that there is no intention to limit the invention to the specific forms disclosed, but on the contrary, the intention is to cover all modifications, alternative constructions, and equivalents falling within the spirit and scope of the invention.

WHAT IS CLAIMED IS:

1. In a computing environment, a method comprising:
detecting (404) regions on forum web pages;
determining (412) which forum web pages have similar regions with which other web pages; and
clustering (414) forum web pages into categories based on their similar regions.
2. The method of claim 1 wherein determining which forum web pages have similar regions comprises determining (408) patterns corresponding to the regions.
3. The method of claim 1 wherein determining which forum web pages have similar regions comprises extracting (428) a feature set for each page.
4. The method of claim 3 wherein extracting the feature set comprises determining (414) a number of patterns on each page.
5. The method of claim 3 wherein extracting (428) the feature set comprises determining locations of the regions, size of the regions, or frequency of occurrence of

repetitive patterns, or any combination of the locations of the regions, the size of the regions, or the occurrence frequency of repetitive patterns.

6. The method of claim 3 wherein extracting the feature set (428) comprises determining an ordering of repetitive regions, a number of rows in repetitive regions, a number of columns in the repetitive regions, a URL pattern of links in the repetitive regions, or a pattern of anchor texts in the repetitive regions, or any combination of a number of rows, a number of columns, a URL pattern of links, or a pattern of anchor texts.

7. The method of claim 1 wherein determining which forum web pages have similar regions comprises extracting a feature set for each page and computing a distance (508) in feature space between pages.

8. The method of claim 1 further comprising, comparing (508, 510) a feature set of a new page against sets of clustered pages to classify the new page based on feature set similarity to one of the sets of clustered pages.

9. The method of claim 7 wherein comparing the feature set of the new page comprises evaluating (510) a feature space distance to a cluster center to determine whether a distance threshold is met.

10. In a computing environment, a system comprising:
a feature extraction mechanism (104) that extracts features of web pages including region-based features; and
a similarity measurement mechanism (106) coupled to the feature extraction mechanism, the similarity measurement mechanism configured to determine similarities between pages based on their features, and to cluster (108) similar pages together.

11. The system of claim 10 wherein the feature extraction mechanism (104) extracts a feature set from each page including a number of patterns on each page.

12. The system of claim 10 wherein the feature extraction mechanism (104) extracts a feature set from each page including locations of the regions, size of the regions, frequency of occurrence of repetitive patterns,

an ordering of repetitive regions, a number of rows in repetitive regions, a number of columns in the repetitive regions, a URL pattern of links in the repetitive regions, or a pattern of anchor texts in the repetitive regions, or any combination of the locations of the regions, the size of the regions, the occurrence frequency of repetitive patterns, a number of rows, a number of columns, a URL pattern of links, or a pattern of anchor texts.

13. The system of claim 10 wherein the similarity measurement mechanism (104) computes a distance in feature space between pages to determine similarity, including performing a computation based on area data and pattern frequency data.

14. The system of claim 10 wherein the similarity measurement mechanism (106) includes means for lessening an influence due to frequent patterns.

15. The system of claim 10 further comprising, means for comparing (508, 510) a feature set of a new page against sets of clustered pages to classify the new page.

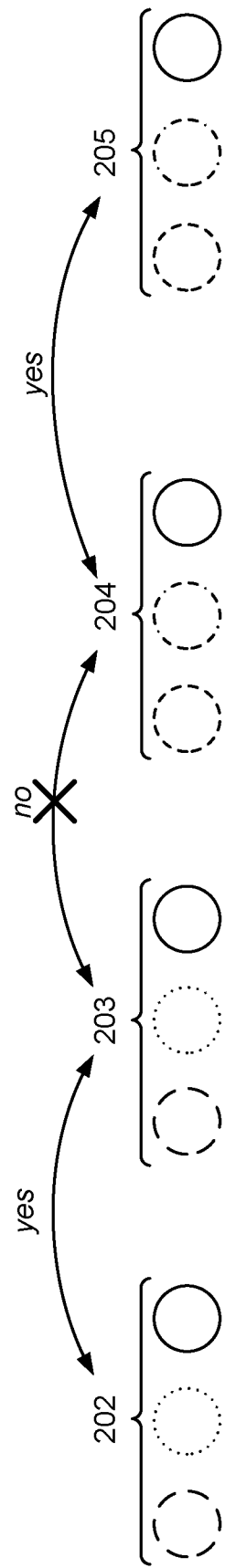
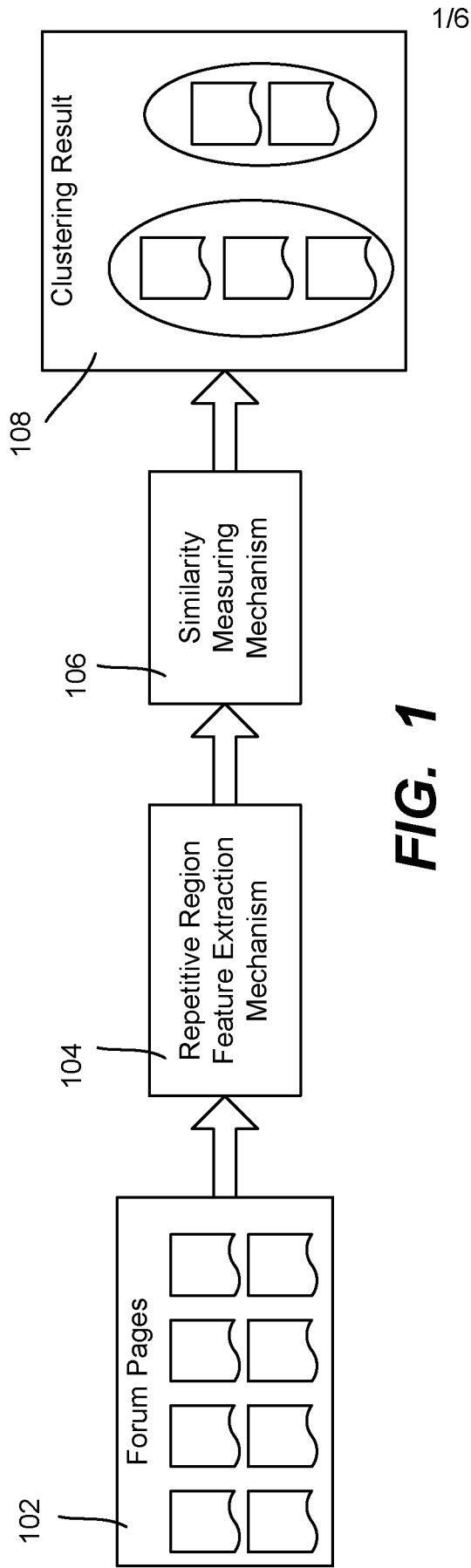
16. The system of claim 10 wherein the means for comparing includes means for evaluating (510) a feature space distance to a cluster center to determine whether a distance threshold is met.

17. A computer-readable medium having computer-executable instructions, which when executed perform steps, comprising, processing a plurality of pages into clusters (108), including by determining (404) repetitive regions between pages and selecting (412, 414) pages with similar repetitive regions for a common cluster.

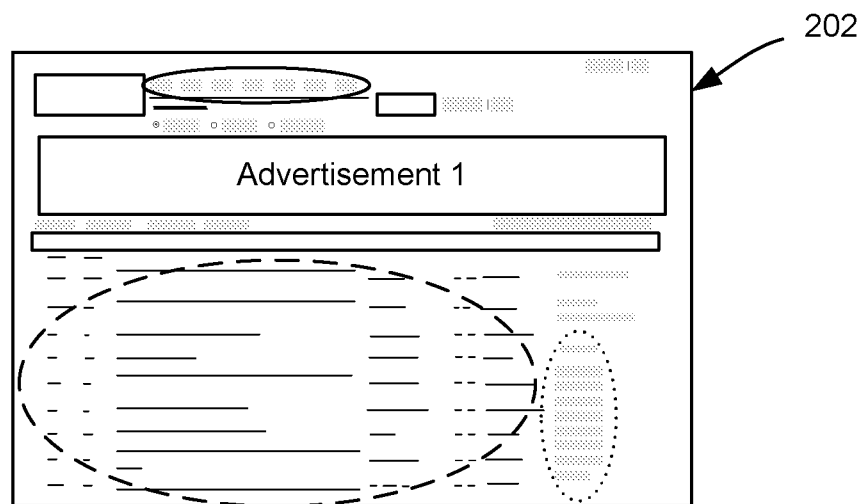
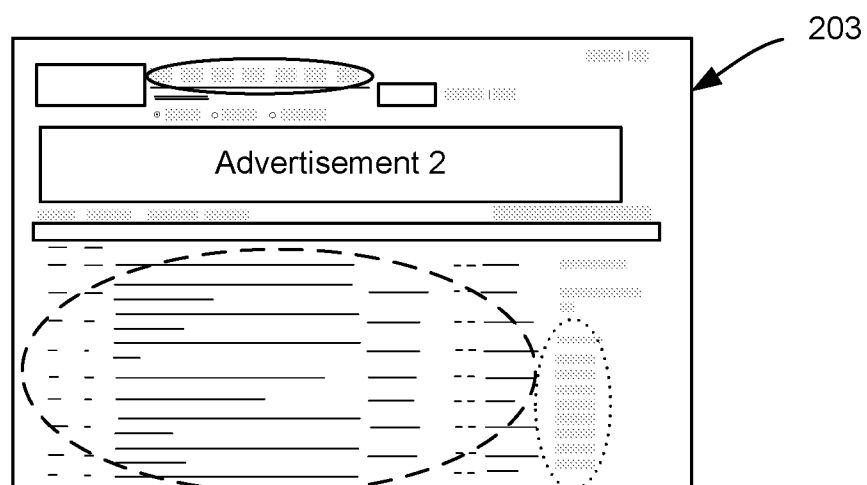
18. The computer-readable medium of claim 17 wherein determining the repetitive regions between pages comprises determining whether patterns corresponding to one page's regions align (412) with any existing patterns of other pages.

19. The computer-readable medium of claim 17 wherein selecting pages with similar repetitive regions comprises extracting (428) a feature set for each page, and comparing (508) the feature sets of two pages.

20. The computer-readable medium of claim 17 having further computer-executable instructions comprising, locating (502) a new page, extracting (504) a feature set from the new page, and selecting (512) a selected cluster for the new page based on similarity of the feature set extracted from the new page to a feature set of at least one of the clusters.



2/6

**FIG. 2A****FIG. 2B**

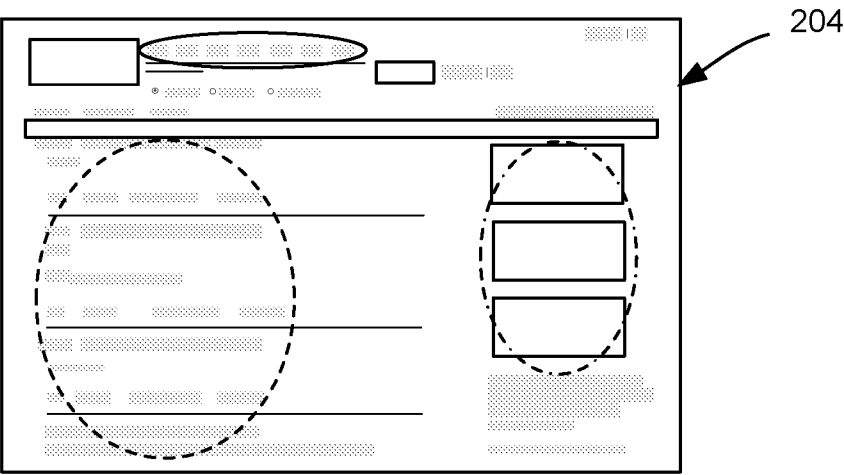


FIG. 2C

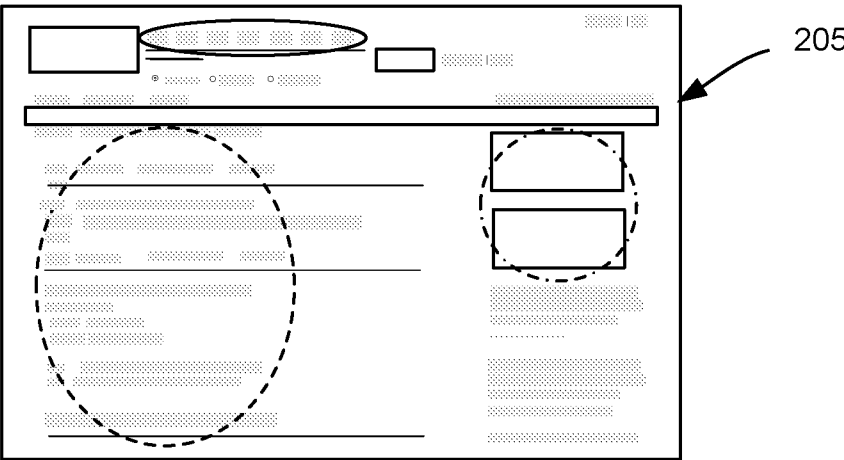
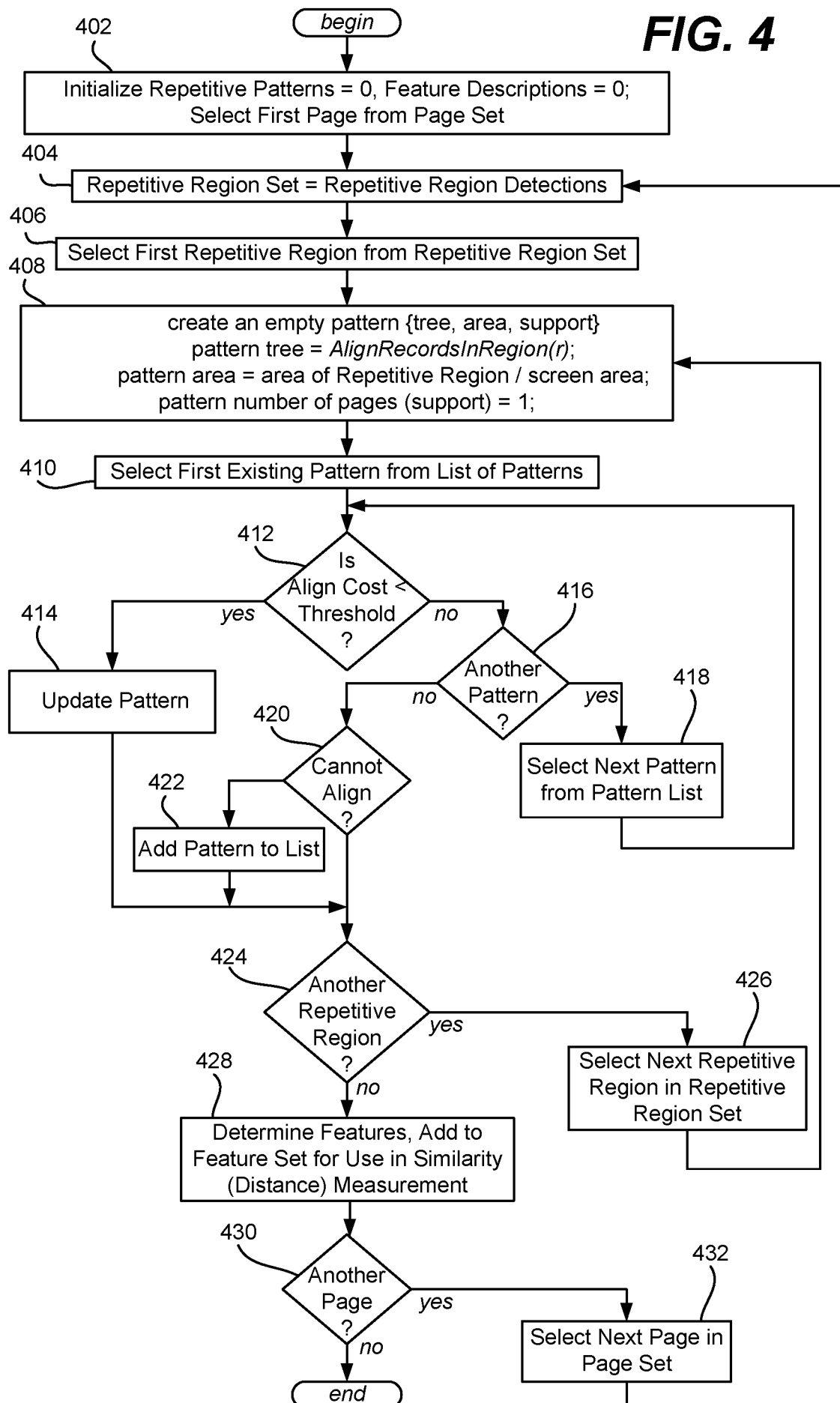
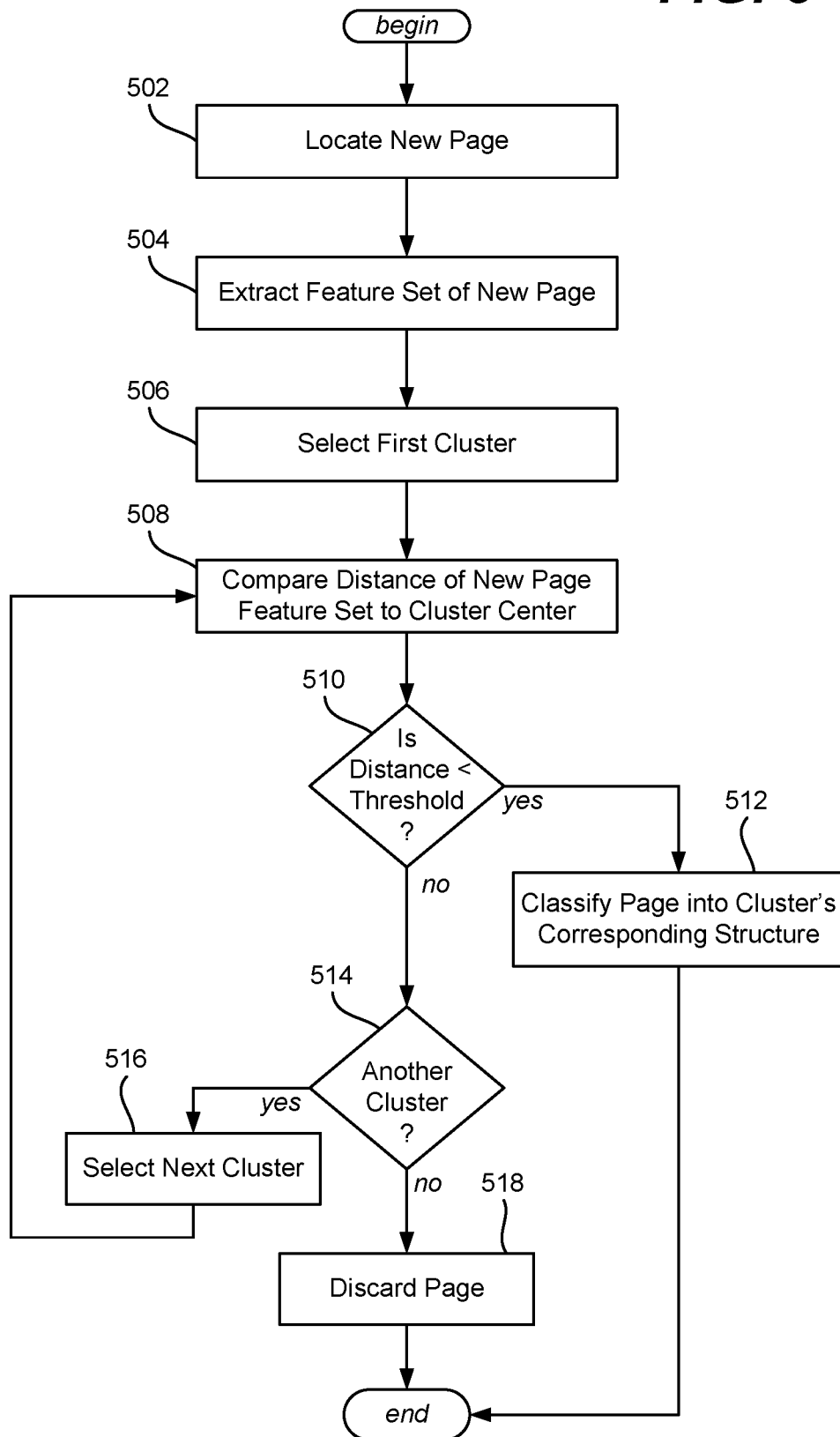


FIG. 2D

4/6

FIG. 4

5/6

FIG. 5

6/6

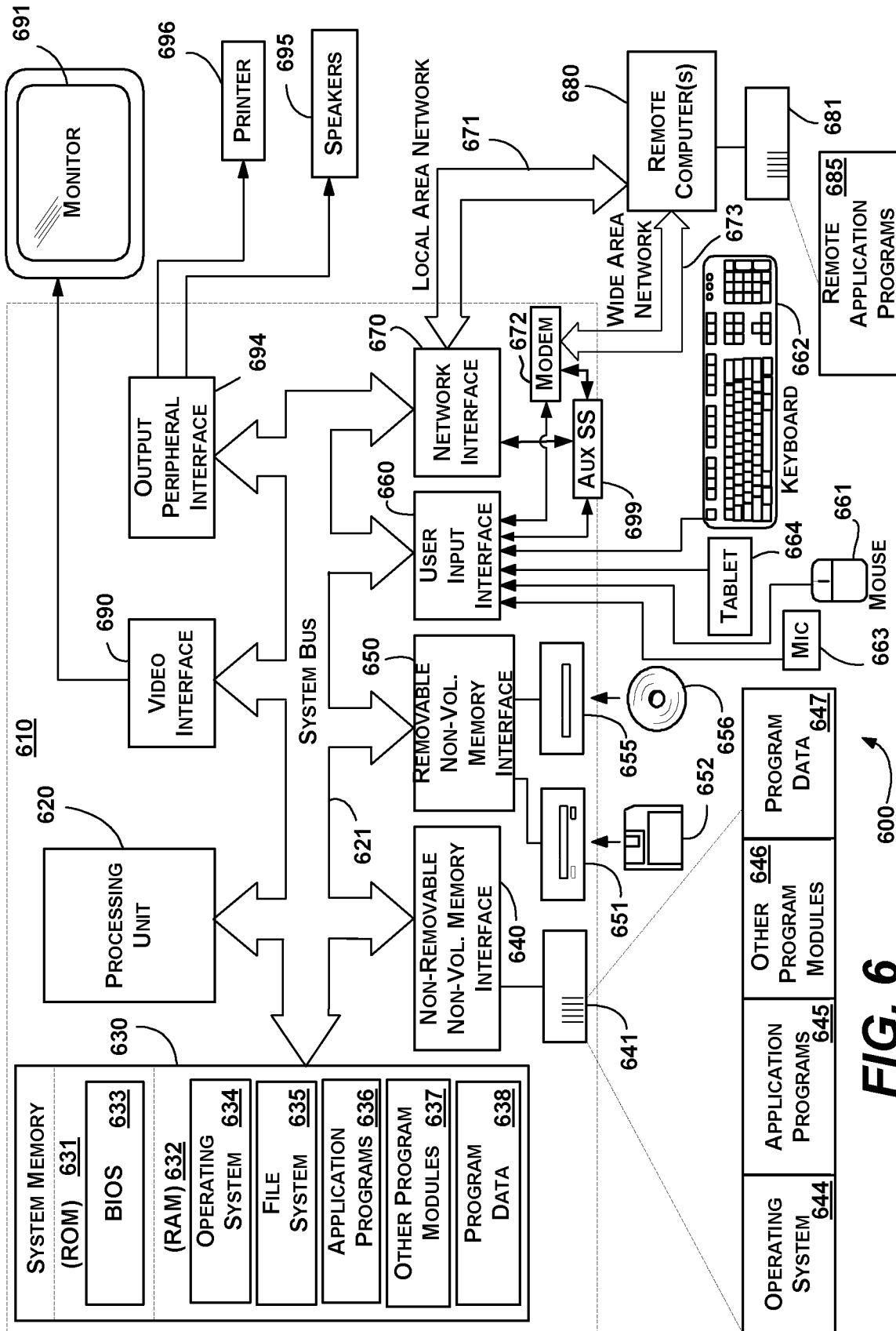


FIG. 6