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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
2 August 2001 (02.08.2001)

PCT

(10) International Publication Number
WO 01/55905 A1

(51) International Patent Classification[7]: G06F 17/30, 11/00

(21) International Application Number: PCT/US00/02280

(22) International Filing Date: 28 January 2000 (28.01.2000)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **WEBSENSE, INC.** [US/US]; 10240 Sorrento Valley Road, San Diego, CA 92121 (US).

(72) Inventors: **KESSINGER, Joshua, A.**; Apartment #308, 2701 Second Avenue, San Diego, CA 92103 (US). **ROBINSON, Nathanial, C.**; 2527 San Joaquin Court, San Diego, CA 92109 (US).

(74) Agents: **HUNT, Dale, C.** et al.; 16th floor, 620 Newport Center Drive, Newport Beach, CA 92660 (US).

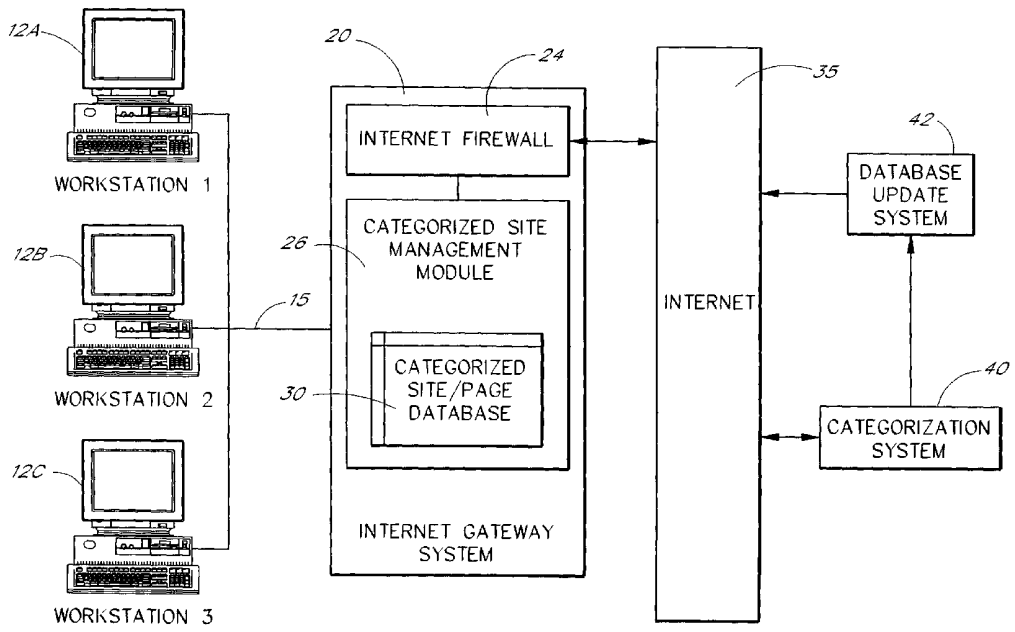
(81) Designated States *(national)*: AE, AL, AM, AT, AT (utility model), AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, CZ (utility model), DE, DE (utility model), DK, DK (utility model), DM, EE, EE (utility model), ES, FI, FI (utility model), GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KR (utility model), KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SK (utility model), SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: AUTOMATED CATEGORIZATION OF INTERNET DATA



(57) **Abstract:** A method and system for automatically categorizing Internet sites is described. The system (20) first assigns relevances of word pairs and word adjacencies to particular categories (30). New Internet pages to be categorized are then parsed so that each word pair and adjacencies on the New Internet page (100A, 100B, 100C), a determination can be made whether the new site (100A, 100B, 100C) should be associated with one or more category (42). The system can then allow users (12A, 12B, 12C) selective access to only sites that are within (or outside of) specific categories (30).

WO 01/55905 A1

AUTOMATED CATEGORIZATION OF INTERNET DATA

Background of the Invention

Field of the Invention

This invention relates to systems and methods for selectively blocking access to particular Internet websites and pages. More specifically, embodiments of this invention relate to systems and methods for automatically categorizing Internet sites and pages so that users can be blocked from accessing specific categories of information.

Description of the Related Art

The Internet is a global system of computers that are linked together so that the various computers can communicate seamlessly with one another. Internet users access server computers in order to download and display informational pages. Once a server has been connected to the Internet, its informational pages can be displayed by virtually anyone having access to the Internet.

The easy access and inexpensive cost of retrieving Internet pages has led to several problems for restricting access to inappropriate information, such as pornography. Several solutions to this problem have been proposed, including rating systems similar to that used for rating movies so that a parent or employer could restrict access to Internet servers, or pages, that have a particular rating. Unfortunately, this mechanism requires each person running an Internet server to voluntarily rate their site. Because of the free-wheeling nature of the Internet, this type of voluntary rating scheme is unlikely to be very efficient for preventing access to sites, such as those containing pornography, that most parents or businesses desire to block.

In addition to a rating scheme, others have developed databases that contain the uniform resource locater (URL) address of sites to be blocked. These databases are integrated into network computer systems and Internet firewalls so that a person wishing access to the Internet first has their URL request matched against the database of blocked sites. Any URL found in the database cannot be accessed by the user. One such system is described in U.S. Patent No. 5,678,041 to Baker et al. Unfortunately, such systems rely on the database of accessed sites to be complete. Because new servers are being added to the Internet on a daily basis, as well as current servers being updated with new information, these databases do not provide a complete list of sites that should be blocked.

Thus, what is needed in the art is a system for restricting access to Internet sites in view of the constantly changing content appearing on Internet servers. The present invention provides such a system.

Summary of the Invention

One embodiment of the invention is a computerized system for controlling access to Internet sites. This embodiment includes: a first module that categorizes an Internet site into predefined subject-matter categories; a second module that stores the address of the Internet site and its associated category to a database; and a third

module that controls access to the Internet site, the third module comprising instructions that block access to the Internet site if the Internet site is within a particular category.

Another embodiment of the invention includes a method for creating a database of categorized Internet addresses. This embodiment provides a method of: retrieving a first Internet page; parsing the Internet page to determine the lexical elements on the page; comparing the lexical elements to a table of category relevancies to determine the relevance of each lexical element to a subject matter category; determining the subject matter category of the Internet page based on the relevance of each lexical element to a category; and storing the address and subject matter category of the Internet page to a database.

Yet another embodiment is a method of controlling access to an Internet site, that includes comprising: categorizing a first Internet site into a predefined subject-matter category; storing the address of the Internet site and its associated category to a database; capturing a user request to view the site; and determining whether the user has permission to view the category of sites and, responsive to the determination, controlling access to the site.

Still another embodiment is a method of controlling access to Internet sites. This embodiment includes a method of: providing a training database, the training database comprising lexical elements and their relevance to subject matter categories; determining the relevance of a plurality of Internet pages to subject matter categories; storing the address of the plurality of Internet pages and their relevance to subject matter categories to a categorized database; comparing the Internet address of pages requested by users with the categorized database to determine if said users have permission to view the pages.

Brief Description of the Drawings

Figure 1 is a block diagram providing an overview of one embodiment of a system for blocking access to Internet sites.

Figure 2 is a block diagram illustrating the categorization system found in Figure 1.

Figure 3 is a block diagram of the tables within the training database described in Figure 2.

Figure 4 is a flow diagram illustrating the process of a user requesting access to an Internet page.

Figure 5 is a flow diagram illustrating the "Analyze Word Content of Page" process found in Figure 4.

Figure 6 is a flow diagram illustrating the process of training data that is performed within the training module of Figure 2.

Figure 7 is a flow diagram illustrating one embodiment of a process for retrieving pages performed by the site/page retrieval module of Figure 2.

Detailed Description

Embodiments of the invention include systems and methods for automatically categorizing Internet pages to create and update a database of categorized sites. This categorized database is then used within an Internet access control system to control user's access to Internet sites within certain categories. For example, if the system

described herein assigns a particular Internet page to a "Sports" category, users that are restricted from viewing sports pages on the Internet will not be granted access to the requested site. In one embodiment, the system is installed within an Internet Gateway computer that controls traffic from the user to the Internet. Because the system described herein becomes more accurate with each page that is scored, minimal user intervention is required to assign pages to categories.

As will be described in detail below, embodiments of this system include a training database that is created by analysis of lexical elements appearing on Internet sites that are strongly associated with a particular category. In this context, a lexical element is a word or plurality of words that appear on the site under analysis. Examples of lexical elements include individual words, word pairs, adjacent words, and triplets of words. Thus, in order to train a "Sports" category, for example, a site for a football team would be fed into the system.

As a first step, each category, such as Sports, is trained to recognize words, words pairs and word adjacencies that are particularly relevant to their category. As discussed herein, a word pair means any two words that appear anywhere on a page. In contrast, a word adjacency is any two words that appear next to one another. Thus, the word adjacency "football team" would be given a strong relevance score to the Sports category. However, this same word adjacency would be given a low relevance score to the Internet Commerce category.

Once a training database has been created of word pairs and word adjacencies, along with their relevance score for each predefined category, any new pages appearing on the Internet can then be analyzed based on the relevance of word pairs/adjacencies appearing in the new pages. For example, a new Internet page having the word adjacency "football team" would be scored highly for the Sports category, but have a low relevance to the Internet Commerce category.

Moreover, by continuing to train each category with pages that have been confirmed to be within a particular category, the system can become increasingly accurate. With each training session, the relevance scores of lexical elements within each page are either increased to indicate a higher relevance to the category, or decreased to indicate a lower relevance to the category.

By using an automated Internet site retrieval program, embodiments of the system provide a database of categorized Internet sites and pages that is constantly updated with new Internet pages as they appear on the World Wide Web. Thus, embodiments of the system provide an efficient system for scoring and categorizing Internet pages.

Overview of the Process

An embodiment of the automated categorization system, as described below, includes computer instructions that, when run, evaluate the source page of an Internet site and categorize the given URL into one of several categories. The system includes three equations that score for:

1. Single Word Relevance — Example: In Category 2, "sex" = 4040.
2. Word Pair Relevance — Example: In Category 2, "sex" and "porn" = 6005
3. Word Adjacency Relevance — Example: In Category 2, "hardcore sex" = 8050

In addition, in other embodiments, equations which score for multiple word associations, such as word pairs, word adjacencies and combinations of higher degrees (triplet, quadruplets, etc.) can be implemented.

The categorization system is first trained by collecting a representative number of Internet sites that best represent the various facets of a given category. These sites are run through a training algorithm that assigns a relevance score to the words, word pairs and word adjacencies found in the Internet sites to the selected category. The result of the training process is a composite of the Internet sites called a "category prototype." The category prototype is a collection of the single word, word pair, and word adjacency relevance scores.

Once a category prototype has been generated for each category, the words, word pairs and word adjacencies from new Internet sites are tested against the category prototypes to determine if the new page should be categorized within any particular category. For example, if the word "sex" occurs on a source page, the computer checks the category prototype and retrieves a relevance score of 4040 for this word within Category 2 (Sex). If the word pair, "sex, porn" occurs on a source page, the computer checks the category prototype and retrieves the score of 6005 within Category 2 (Sex) for the word pair "sex, porn". This process is repeated for every word pair and word adjacency on the retrieved page. These scores are then used to calculate a category rating for the retrieved page.

The category rating is used to evaluate the probability that a page should be placed in a given category. For instance, if a URL has a category rating of 5000 within category two, then its associated probability of being within that category might be .99. This means that if there were 100 sites, each with a category two rating of 5000, then 99 of those sites belong in category two. In general, as the category rating increases, the probability that the corresponding site belongs to that category also increases. Consequently, it is possible to use this feature to establish a cut-off point that maintains 99% accuracy (or any other accuracy).

One goal of the process is to obtain two cut-off points within each category: the alpha point and the beta point. These two points create benchmarks against which decisions concerning a site's categorization can be made. The alpha point is chosen to maintain a sorting accuracy of, for example, 99%. As is known, the sorting accuracy is simply the computer's ability to correctly sort sites into a specific category. The Alpha point can be calculated for any category by using the following equation:

$$Ap = M7 + 4\,(SD7),$$

where, **Ap** = alpha point, **M7** = the average category rating of the incorrectly sorted sited within the specific category, and **SD7** = the standard deviation of the category rating for the incorrectly sorted sites within the specific category. This ensures 99 percent sorting accuracy because we are calculating four standard deviations away from the mean score, and should generalize to the Internet at large for the given category.

The beta point's sorting accuracy will undoubtedly vary between categories. However, it may generally maintain a sorting accuracy between the ranges of 75 to 85 percent. The beta point can be found using the equation:

$$Bp = M7 + 1\,(SD7),$$

where, **Bp** = beta point, **M7** = the average category rating of incorrectly sorted sites within the specific category and **SD7** = the standard deviation of the category rating for the incorrectly sorted sites within the specific category. Sites that fall between the beta point and the alpha point will be placed into a Suggest Database to be viewed by Web Analysts or technicians. It should be noted that each category will be assigned its own unique alpha and beta points.

As discussed below, embodiments of the system include the one or more modules. These modules include software instructions that are run on processors within the computer system. The modules can also include storages, such as Random Access Memory (RAM), Read Only Memory (ROM), Electrically Erasable Programmable Read Only Memory (EEPROM), hard disks or other computer storage devices.

Figure 1 provides an overview of a system 10 for controlling access to particular sites on the Internet. As shown, a plurality of workstations 12A-C are connected through a local area network 15 to an Internet gateway system 20. The workstations 12A-C are preferably Intel Pentium class personal computers operating under the Microsoft Windows Operating System. Of course, it should be realized that any conventional personal computer, such as those manufactured by Apple, IBM, Compaq, Dell, Digital Equipment Corp. (DEC) or other system can be used in a similar manner.

The local area network 15 is preferably an Ethernet 10baseT topology, but can be based on any well-known networking protocol, including wireless networks, token ring networks and the like. The local area network 15 communicates with the Internet Gateway system 20 in order to provide the workstations 12 A-C with TCP/IP communication to sites on the Internet 35. Such gateways are well known in the art and normally communicate through routers or other data packet switching technology for translating Internet TCP/IP protocols into the proper protocols for communicating across the local area network 15.

Within the Internet gateway system 20 is an Internet firewall module 24 that monitors data packets flowing to and from the Internet 35. The firewall module 24 controls access between the workstations 12A-C and the Internet so that unauthorized users cannot gain access to computer resources on the local area network 15. Thus, all communication between the Internet and the network server 15 first passes through the firewall 24. Many firewall software programs are available, such as Firewall-1 (Check Point software, Redwood City, California). However, it should be realized that while the embodiment described in Figure 1 relies on a firewall to control access of data packets between the Internet and the workstations 12A-C, other similar access control systems are available. For example, the Microsoft proxy server (Microsoft Corp., Redwood City, WA), Netscape proxy server (Netscape Corp) and the Open Server implementation of Cisco's Pix Firewall (Cisco Corp.) are currently available and can be implemented in place of the firewall 24.

Within the Internet gateway system 20, and communicating with the firewall 24 is a categorized site management module 26 that includes instructions for analyzing Internet site requests from the workstations 12A-C and then comparing those Internet site requests with a categorized site/page database 30. If the requested page is found within the database 30, it will either be blocked or allowed depending on the access rights granted to the user

within the management module 26. As illustrated, the categorized site management module 26 communicates with the firewall 24 to control access to the Internet 35.

Also connected to the Internet 35 is a categorization system 40 that, as described below, categorizes websites and pages in order to create the categorized site database 30. Once sites on the Internet have been categorized by the categorization system 40, a database update system 42 thereafter routinely copies the updated database from the categorization system 40 to the Internet gateway system 20. As can be imagined, the system can include hundreds of gateway systems, each of which is updated regularly by the database update system 42 to provide an updated database of blocked Internet sites. Moreover, the database update system 42 can preferably only transfer portions of the database to the gateway system 20 so that the entire database does not need to be transmitted.

Overall, Figure 1 illustrates one embodiment of a system for providing controlled access of workstation computers to the Internet. Each request from a workstation for an Internet address (e.g.: page or site) is first compared to a categorized database of Internet addresses. If the requested address is found within the categorized database, a management module accesses a user permissions table to determine if the requesting user has rights to view sites within the category that is associated with the requested page. If the user has access rights to view pages within the category, the page request is sent to the Internet. However, if the user does not have any access rights, the user is blocked from receiving the requested page from the Internet.

Referring to Figure 2, the categorization system 40 (Figure 1) is explained in more detail. As illustrated, Internet pages 100A, B and Internet site 100C are retrieved by a site/page retrieval module 110. Within the site/page retrieval module 110 are instructions for searching and retrieving Internet pages and sites from the Internet. One exemplary method for retrieving such sites is illustrated below in Figure 7.

Once an Internet site or page has been retrieved by the retrieval module 110, it is forwarded to an analysis module 120 in order to determine which category (or categories) is most strongly related to the retrieved site. The process for analyzing an Internet page for its relevance to one or more categories is explained in more detail below in Figure 5.

As illustrated, the analysis module 120 is linked to a copy of the categorized database 30' and a training database 125. The analysis module 120 calculates the relevance of the retrieved Internet page to each of the predefined categories by analyzing the word pairs and word adjacencies within the page. In order to provide this analysis, the training database 125, as explained below, includes category relevance scores for each word pair and word adjacency that might be found on the page. Thus, by comparing the word pairs and word adjacencies within the retrieved page to the scores for those word pairs and adjacencies within the training database, a total relevance score for the page within each category can be determined. Once a page relevance score has been calculated for the page in each category, a determination is made whether the relevance score for each category is high enough to warrant assigning the retrieved score to any category.

As discussed below, the determination of whether to assign a retrieved page to a particular category is made by comparing the page's relevance score for a particular category with a predetermined alpha value. If the page relevance score is higher than the alpha value for the category, the page is assigned to that category. If the score is lower than the alpha value, but greater than a beta value, the page is forwarded to a manual scoring system wherein technicians view the retrieved page and determine whether or not to include the page within the category. If the relevance of the page for a category is below the beta value, the page address is stored to a database of analyzed sites, and the system continues to score additional addresses.

The data within the training database 125 is created by providing training data 130 to a training module 135, as illustrated. The training data 130 includes Internet pages strongly associated with each category to be trained. For example, in order to train a Sports category, the training data might include the Internet address of a sports franchise or other sports website. The training module 135 then parses the word pairs and word adjacencies for each page within the given sports site. Any unique word pairs and word adjacencies, as described below, are then assigned high relevance scores in the Sports category within the training database. Thus, similar words and word pairs appearing on new pages will be given high relevance scores to the Sports category.

Referring to Figure 3, one embodiment of a training database 125 is illustrated. Within the training database 125 is a word identification table 200 that includes lists of words and a corresponding ID number for each word. This table allows every word pair or word adjacency referenced in the database to be represented by two numbers instead of two words. Since, in general, the number of characters in the ID number is less than the number of characters in the word itself, much less data storage space is required within the training database to store numerical representations of each word instead of the word itself. In addition, well-known words, such as "the" and "and" can be represented by single-digit numbers so that only one byte of data is taken to represent these common words. However, as discussed below, such common words are normally discarded prior to scoring an Internet page so that the lexical elements on each page will be more readily differentiated from every other Internet page. This provides a more advantageous page scoring system.

In addition to the word identification table 200 is a category identification table 205 that provides a category ID number for each category within the system. The category identification table 205 also includes an alpha and beta score that provide the cut-off values for assigning a particular page to the selected category. For example, as illustrated in Figure 3, the Sports category includes an alpha score of 920 and beta score of 810. If an Internet page is found to have a page relevance score of greater than 920 for the Sports category, it will be assigned to the Sports category. However, if the Internet page is found to have a page relevance score of between 810 and 920, it will be flagged for manual follow-up by a technician to determine whether or not it belongs within the Sports category. If the Internet page is found to have a page relevance score of below 810 for the Sports category, then it will not be flagged as being related to the Sports category. By using these values, the system determines whether or not to assign a particular page to one of the predefined categories.

Also within the training database 125 is a word relevance table 210 that provides the relevance scores of word pairs and word adjacencies with particular categories in the system. For example, the word "Cleveland" (ID No. 234) and the word "Browns" (ID No. 198) are illustrated with a word adjacency relevance score of 900 to category 1 (Sports). Because, in this illustration, the maximum relevance score is 1,000, the word adjacency "Cleveland Browns" is very strongly associated with the Sports category. Thus, any Internet page having the words "Cleveland Browns" adjacent one another will have their total page score raised in the Sports category due to the strong relevance of these words to sports.

Note that the words "diamond" (ID No. 755) and "jewelry" (ID No. 1345) only have a relevance score of 290 within the Sports category. However, the word pair "diamond" and "jewelry" is illustrated with a relevance score of 940 in category 3 (Shopping). Thus, as illustrated, any page having both of these words will be more strongly associated with the shopping category, and more weakly associated with the Sports category.

Referring to Figure 4, an overall process 300 of requesting access to an Internet page or site is illustrated. The process 300 begins at a start state 302 and then moves to a state 306 wherein an Internet browser on a workstation computer 12A-C requests an address on the Internet. Well-known browsers include Microsoft Explorer and Netscape Navigator. The browser request is normally made after a user has entered a desired URL into their browser software.

The user's request is then sent across the local area network 15 to the Internet Gateway system 20. The process 300 then moves to a state 308 wherein the requested Internet address is matched against the categorized database 30. It should be noted that the address can be a single page within an Internet site, or the default address of the site (e.g.: www.company.com).

A determination is then made at a decision state 310 whether an address match has been made with any address stored in the categorized database. If no match was found within the categorized database 30, the requested page is retrieved from the Internet at a state 312 and the process terminates at an end state 314.

However, if an address match between the requested address and the categorized database is found, the process 300 moves to a decision state 315 wherein a determination is made whether the current user has restricted access rights to specific categories of Internet pages. This determination can be made by reference to a list of network users, and an associated permissions table for each category found within the categorized database. Thus, a particular user may be restricted from access to all Sports and Pornography categories but not restricted from Internet Commerce or Travel categories. An exemplary list of Internet categories is provided below in Table 1.

Table 1

Listing of Categories

Category	Description
Abortion Advocacy	Abortion advocacy, pro or con.
Activist Groups	Organizations with a cause. This is a broad category that can include environmental groups and any other activist group not covered under other categories. **Note:** No special exceptions are made for Freedom of Speech activist sites.
Adult Entertainment	Full or partial nudity of individuals. This might include strip clubs, lingerie, adult-oriented chat rooms, erotica, sex toys, light adult humor and literature, escort services, password-verification sites, prostitution, and so forth. Sexually explicit language describing acts that would fit into this category are also categorized here.
Alcohol/Tobacco	Any site promoting, containing, or selling liquor or tobacco products, or their accessories.
Alternative Journals	Online equivalents to supermarket tabloids, or non-mainstream periodicals. **Note:** This category may contain materials that are sexual in nature.
Cult/New Age	Promoting or containing information on witchcraft, black arts, voodoo, spirituality, horoscopes, alternative religions, cult, UFOs. All religions not covered under the Religion category.
Drugs	Promotion of illegal drugs and/or drug culture information, or drug-related contraband. **Note:** As legality of drugs varies by country, the drug laws of the United States are used.
Entertainment	Sites promoting/containing information on movies, radio, television, books, theater, sedentary hobbies, magazines (non-business related), music, pets, humor/jokes, and sites containing downloadable software of an entertaining nature. **Note:** Computer magazines containing technical information are not included in this category.
Gambling	Any site that promotes gambling or allows online gambling.
Games	Information about or advocacy of board games, electronic games, video games, computer games, or on-line games. Includes both hardware and software.
Gay/Lesbian Lifestyles	Information about gay and lesbian lifestyles that does not contain sexually explicit images or text. Dating services and shopping sites that cater to gay or lesbian customers.
Hacking	Any site promoting questionable or illegal use of equipment and/or software to hack passwords, create viruses, gain access to other computers, and so on. Does not include security information sites.
Illegal	Promotion or information describing how to commit non-violent, illegal activity such as drunk driving, mail fraud, picking locks, white or blue collar crime of a non-technical nature. **Note:** U.S. laws are used as a guide.
Job Search	Personal job/career search sites.
Militancy	Any site promoting or containing information on militia operations, terrorist activity, war, riots, rebellion groups. Advocates of violence to overthrow governments.
Personals/Dating	People meeting other people, personal ads, mail order brides. Sites combining heterosexual and gay personals on the same site are included here. Dating and personals sites that accommodate only gay and lesbian lifestyles are categorized

Category	Description
•	under Gay/Lesbian Lifestyles.
Politics	Political advocacy of any type. Any site promoting or containing information on any political party, pro or con. This includes all registered and otherwise officially recognized political parties. Excludes all official government sites.
Racism/Hate	Ethnic impropriety, hate speech, anti-Semitism, racial clubs/conflict.
Religion	Religious advocacy, pro and con. Limited to: Atheism, Buddhism, Christianity, Hinduism, Island, Judaism and Shintoism.
Sex 1	Heterosexual activity involving one or two persons, hard-core adult humor and literature. Sexually explicit language describing acts that would fit into this category are also categorized here.
Sex 2	Heterosexual acts involving more than two people, homosexual and bisexual acts, orgies, swinging, bestiality, sadism/masochism, child pornography, fetishes and related hardcore adult humor and literature. Sexually explicit language describing acts that would fit into this category are also categorized here.
Shopping	Consumer-oriented online shopping. Includes real estate shopping. Excludes sites that sell sex toys, weapons, alcohol, tobacco, vehicles and vehicle parts or travel services. **Note:** The entire site is screened if the intent of the site is selling.
Sports	Sports and sports-related recreation. Team or individual activities, indoor or outdoor, with a physical component. For example, body building, hiking, camping, and football.
Tasteless	Offensive or useless sites, grotesque depictions caused by "acts of God."
Travel	Sites promoting or containing information on travel, leisure, vacation spots, transportation to vacation destinations.
Vehicles	Any site promoting vehicles, including: cars, vans, trucks, boats/water craft, ATV's, trains, planes and any other personal vehicles and vehicle parts. Vehicles within this category do not carry weapons.
Violence	Any site promoting or containing information on violent acts, murder, rape, violent criminal activity, gangs, gross depictions caused by acts of man, excess profanity.
Weapons	Any site promoting/containing information on guns, knives, missiles, bombs, or other weapons.
Web Chat	Chat sites via http protocol, chat rooms (non-IRC), forums and discussion groups. Home pages devoted to IRC.

Once a determination has been made at the decision state 315 that the user has restricted categories, the process 300 moves to a state 316 to determine which categories have been blocked for this particular user. This determination is made by reference to permissions list associated with the user.

The process 300 then moves to a decision state 320 to determine whether the requested page is within any of the restricted categories for this particular user. This determination is made by first determining the category of the requested address from the categorized database, and then comparing that result with the restricted categories for the user. If a determination is made that the requested page is not within one of the user's restricted categories, the revised page is retrieved at a state 324 and the process terminates at the end state 314.

If a determination is made at the decision state 320 that the requested page is within one of the user's restricted categories, the process 300 moves to a state 340 wherein access to the page is blocked. This blocking can occur by discarding the packet request from the user to the Internet, or simply closing the connection that was requested by the Internet browser to the requested page. The process 300 then returns an appropriate page notifying the user that their request has been denied. The process 300 then terminates at the end state 314.

Thus, Figure 4 provides an overview of one process for requesting and blocking access to particular Internet addresses based on whether the requested page appears within the categorized database 30. Figure 5 provides a method for creating the categorized database 30 by analyzing the content of word pairs and word adjacencies within Internet pages.

Referring to Figure 5, a process 328 of analyzing the word content of pages to determine their relevance to particular categories is illustrated. The process 328 begins at a start state 400 and then moves to a state 402 wherein the first word in an Internet page is retrieved. As used herein, the term "word adjacency" includes words that are directly adjacent one another. The term "word pair" includes any two words that are located on the same Internet page.

Once a first word from the page has been retrieved at the decision state 402, the process 328 moves to a state 404 wherein the relevance of every word pair that contains the first word in the page is determined for each of the defined categories. Thus, the first word and the third word in the page are determined, and that word pair is compared against the word relevancy table 210 in the training database to determine its relevance score in each of the listed categories. This relevance score is determined by reference to the word relevance table 210 (Figure 3) which lists each word pair and its associated relevance to every category. In one embodiment, the relevance score of a word pair within a particular category varies from 0 to 1,000, with 1,000 being a word pair that is perfectly associated with a category. Of course, various scoring systems can be developed that reflect the relevance of a particular word pair to a category. It should also be understood that a maximum distance between any two words within a word pair can be set. For example, the system may only analyze word pairs that are 10, 20, 30, 40 or more words apart, and then move to begin analyzing the next word in on the page.

The determined word pair relevance scores are then stored to a memory for later manipulation. The first word is then paired with the fifth word in the page to determine the new word pair's relevance to each category. This process is repeated for every possible two-word pair in the page that includes the first word.

The process 328 then moves to a state 405 wherein the relevance of the word adjacency of the first word and the second word is calculated by matching these words to the word relevance table 210 in the training database to determine their relevance to each category.

Once the relevance score for the retrieved word adjacency has been determined for every category, the process 328 moves to a state 408 wherein the relevance scores determined at the state 404 for each of the word pairs is added to the total page score for each category.

Thus, if the word pair "Cleveland" and "Browns" returned a relevancy score of 900 from the word relevancy table in the Sports category, the numerical value 900 would be added to the total page score for category 1 (Sports). Thus, word pairs having higher relevance scores in a category will result in a higher overall page relevance score in the current category for that page. Similarly, word pairs having lower relevance scores in a particular category will reduce the overall page relevance score to that category.

Once the word pair relevancy scores for the page have been added to the total page relevance score, the process 328 moves to a state 409 wherein the word adjacency relevancies that were determined at state 405 for each category are added to the page relevance category scores for the current Internet page.

Now that the page scores for each category have been calculated, a determination is made at a decision state 416 whether more words exist on the page to be analyzed. If a determination is made that no more words are available for analysis on the retrieved Internet page, the process 328 moves to a state 420 wherein the total page relevance score for each category is normalized to take into account the fact that pages with more words will have higher scores. For example, since page scores are determined by adding the relevancies of word pairs and word adjacencies, a page with 500 words will have a substantially higher score in each category than a page with 100 words. Thus, for example, dividing the page relevance score within each category by the total number of words on the page will normalize the page score so that pages of differing lengths will have approximately the same page score in each category. It should be noted that categories having higher average relevance scores for each word pair and word adjacency will have a higher page score than those categories having word pairs with lower relevance scores.

Once a normalized page score has been determined in each category for the retrieved page, the process 328 moves to a decision state 422 to determine whether the page relevance score for the category is greater than the alpha relevance score for that category. This determination is made by reference to the category ID table 205 in the training database 125. If the page relevance score is not greater than the alpha score, the process 328 moves to a decision state 424 to determine if the page relevance score is greater than the beta score for the category. If a determination is made that the page relevance score is not greater than the beta score, the process 328 moves to a state 426 wherein the retrieved site is stored to a table and flagged as having been analyzed, but not within any category. The process 328 then terminates at an end state 430.

If a determination is made at the decision state 422 that the page relevance score is above the alpha score for the category, the process 328 moves to a state 432 wherein the retrieved address is added to the categorized database 30. It should be noted that the categorized database 30 includes not only the address of the Internet addresses to block, but also the category that the Internet site is associated with so that a determination can be made whether a user having particular permissions should be provided access to the site, even though it is categorized within the database.

In an alternative embodiment, if a determination is made that the page score is greater than the alpha score for the category, the system may run instructions that access the current page on the Internet. The instructions then begin to score the hierarchical pages of the site while moving towards the main domain address

(e.g.:www.company.com). If a determination is made that any of higher nodes on the site are also above the alpha score for the same category, those sites are also added to the database. This provides the system with a mechanism for not only rating an individual page, but also the plurality of pages that appear below a specific node on an Internet site.

In one embodiment, the number of words that are considered on any page is limited to a predetermined number. For example, the system might be limited to only considering the first 100, 250, 500 or 1000 words on any page. Any words that follow the predetermined number would not be considered.

If a determination is made at the decision state 424 that the page relevance score is greater than the beta score, but lower than the alpha score, the process 328 moves to a state 434 wherein this address is flagged for further analysis by a technician. The process then terminates at the end state 430.

If a determination is made at the decision state 416 that more words are left to be analyzed in the retrieved page, the process 328 moves to a state 436 wherein the next word in the page is selected as the first word for each word pair and word adjacency. In this manner, the system "walks" across the page by analyzing each word in the page in conjunction with every other word. This provides a complete analysis of every possible word pair and word adjacency in the page.

Through the process 328 illustrated in Figure 5, a newly retrieved Internet page is scored and associated with one or more categories within the system. Each page that is found to have relevancy score within any category that is greater than the alpha score for that category is added to the categorized database 30 for the categories that it is associated with. In addition, any page that is found to have a relevancy score that is greater than the less stringent beta score is flagged for analysis by a technician so that it can be manually added to the categorized database, if necessary. Through this mechanism, new Internet pages are added to the system on a regular basis.

Referring to Figure 6, a process 500 for creating the word relevance table 210 within the training database 125 is described. The process 500 begins at a start state 502 and then moves to a state 504 wherein a first category to train is selected. The category might be, for example, the Sports category. The process 500 then moves to a state 508 wherein web pages that have been predetermined to be within the chosen category (e.g., sports) are retrieved. Thus, because these pages are known to be within the category selected at state 504, the relevance of each word pair and word adjacency within the chosen page can be assigned a high relevance to the current category.

Once web pages within the chosen category are retrieved, the process 500 moves to a state 510 wherein a target page score is determined for the currently selected page. Normally, a page that is highly relevant to a particular category is given a score of, for example, 1,000. However, it should be realized that any similar type of scoring scale that is used to relate words to a category can similarly be implemented. Once the target page score is determined at the state 510, the process 500 moves to a state 516 wherein the first page of the retrieved pages is selected for analysis.

The number of words on the selected page is then counted at the state 520 and the process thereafter moves to a state 526 wherein the number of unique word pairs are divided by the target page score (1000) so that if

the word pairs were re-scored, the total page relevance score would be 1000. Similarly, the target page score (1000) is divided by the number of unique word adjacencies to result in a word adjacency score that, if added together, would result in a page relevancy score of 1000 (extremely high relevance to the trained category). It should be noted that common words such as "a", "the" and "and" are ignored to minimize processing time and increase the accuracy of the scoring process. Moreover, computer language instructions and hypertext headers are also ignored in order to increase the accuracy of scoring the pages.

The process then moves to a state 530 wherein the current score for each word pair and word adjacency (1000) is averaged with the same word pair and word adjacency scores already stored in the word relevance table. Thus, if we are training the Sports category, and the word adjacency "Cleveland Browns" is found within the current page, it might be assigned a word adjacency value of 105 in the Sports category. However, if the term "Cleveland Browns" is already scored within the Sports category at a value of 89, the 105 value and the 85 value would be averaged to normalize the word adjacency score to the Sports category. This system therefore allows words that are used over and over within certain categories to be "up-trained" so that their relevance score with the chosen category will go up as they appear on more pages that are scored. In addition, it should be understood that the system is capable of parallel processing of a plurality of sites simultaneously.

The process 500 then moves to a state 534 wherein the alpha and beta scores for the category being trained are determined. The alpha score is the numerical score that, when exceeded, indicates that the selected page is clearly within a category. The beta score is the numerical score that, when exceeded, indicates that the selected page may be within a category. As discussed above, the alpha score is normally chosen so that 99% of the pages having that score are within the chosen category. The beta score is normally chosen so that 75-85% of the pages having that score are within the chosen category. These scores are determined by analyzing the average score of the trained pages in the category to determine cut-off values for new pages.

The word relevance scores are then saved to the word relevance table 210 in the training database 125 at a state 536. A determination is then made at a decision state 540 whether more pages that need to be trained are available. If no more pages are available, the process 500 terminates at an end state 544. If a determination is made that more pages do exist, the process 500 moves to a state 550 wherein the next page to be analyzed is selected. The number of words are then counted on the page at the state 520 and the process continues as described above.

Through the process 500 described above, a word relevance table is developed which includes normalized word relevances for every word pair and word adjacency that might be found in an Internet page. By analyzing new pages and by adding together the relevances of each word within the page, an automated system is provided for assigning a page relevance score for a particular page to each of the predetermined categories within the system. Thus, once a particular category has been trained by analysis of a large number of pages, the system can rapidly analyze new pages for their relevance to each of the predetermined categories.

As described above in Figure 2, a page retrieval module 110 is utilized for retrieving new Internet pages and sending them to the analysis module 120 for scoring. Figure 7 provides an illustration of a process 600 for retrieving pages from the Internet.

The process 600 begins at a start state 602 and then moves to a state 606 wherein the address of the first site to categorize is determined by random access of an address from the categorized web database 30. Once an address of a first site to categorize is determined at the state 606, the process 600 moves to a state 610 wherein the first page of the Internet site is read. The process then moves to a state 612 wherein the page that has been read is forwarded to the analysis module 120 so that the word pairs and word adjacencies on the page are analyzed for their relevance to a predetermined category.

The process 600 then moves to a decision state 616 in order to determine whether more pages exist on the current site to be analyzed. If no more pages exist on the current site, the process 600 moves to a decision state 620 to determine whether any sites on the Internet reference the currently analyzed site. If no more sites reference the current site, the process 600 terminates at an end state 624.

If more pages do exist to be analyzed at the decision state 616, the process 600 moves to a state 630 wherein the next page on the current site is read. The process then continues to state 612 wherein the new page is sent to the analysis module 120.

If a determination is made at the decision state 620 that there are sites that reference the current site, the process 600 moves to a state 632 wherein the system points to the address of the first referenced site. The process 600 then returns to the state 610 in order to read the first page on the newly retrieved Internet site.

EXAMPLE 1

Normalizing Training Data

As discussed above, the source pages of different web sites have different numbers of words on them. This can affect the word pair and word adjacency training process since Internet sites with fewer words on them can force higher relevancies on word pairs and word adjacencies than sites with fewer words. For instance, consider two pages, A and B, with 10 and 500 words pairs on their source pages respectively. Assuming each site has a current page score (***Sc***) of 0 and a target page score (***St***) of 1000. The current training algorithm takes the form of the following equation:

(E1) $Wrn = Wrc + I$,

where ***Wrn*** is the new word pair relevance and ***Wrc*** is the current word pair relevance and ***I*** is the amount that the each word pair relevance should be incremented such that if the page were immediately re-scored its score would equal the target score. ***I*** can be found by taking the current score, subtracting it from the target score and dividing it by the total number of word pairs (***Wt***) on the page. The equation is as follows:

(E2) $I = [(St - Sc) / Wt$

Finding the new word pair relevance requires adding the current relevance to the increment value. The new word pair relevance equation then becomes:

(E3) $Wrn = Wrc + [(St - Sc) / Wt]$

Using the equation above to calculate the word pair relevances for sites A and B we find:

(E4) $Wrn(A) = 0 + [(1000 - 0) / 10] = 100$ (note: $I = 100$)

(E5) $Wrn(B) = 0 + [(1000 - 0) / 500] = 2$ (note: $I = 2$)

Interpreting these results, the 10 word pairs on site A would each have a relevance of 100 while the 500 word pairs on site B would each have a relevance of 2 to the chosen category after one round of training.

If these two sites were determined to be equally "qualified" to train a particular category, then logically they should influence word pairs from other pages to a similar degree. However, at this point, this is not the case. Instead, a site with 10 word pairs can influence the weight of words found up to as much as 5000% more than a site with 500 word pairs. Instead, a system that increments word pairs "evenly", regardless of the number of words that occur on the page is desired.

A method for normalizing the amount that each word pair is incremented is advantageous. Using the results from E4 and E5, the minimum and maximum amount that each word pair can be incremented is 100 and 2 respectively. Since, we want the minimum relevance score and the maximum relevance score to approach each other, we can take their average using the midpoint theorem:

Mp = (p1 + p2) / 2 , where Mp is midpoint, p1 is point 1, and p2 is point 2

We find that the midpoint between the min and max increment is:

(E6) $Mp = [I(A) + I(B)] / 2$

Using the values from E4 and E5,

(E7) $Mp = [100 + 2] / 2 = 102 / 2 = 51$

Thus, determining the *"adjustment constants"* that should be used to adjust the relevance scores towards the midpoint score for each site relies on the following two equations:

(E8) $I(A) * AdjCon(A) = Mp$ or $AdjCon(A) = Mp / I(A)$

(E9) $I(B) * AdjCon(B) = Mp$ or $AdjCon(B) = Mp / I(B)$

Substituting in,

(E10) $AdjCon(A) = 51 / 100 = .51$

(E11) $AdjCon(B) = 51 / 2 = 25.5$

Therefore, with ten words, the increment should be multiplied by .51 to reach the midpoint value of 51. Similarly, with 500 words, the increment value needs to be multiplied by 25.5 to reach the midpoint value of 51. This logic can be used to formulate the training normalization constant, ***Nt***. The equation for calculating ***Nt*** is:

(E12) $Wt(X) * Nt = AdjCon(X)$ or $Nt = AdjCon(X) / Wt(X)$

With a min of 10 words ($Wt(A) = 10$) and max of 500 words ($Wt(B) = 500$), the training normalization constant is:

(E13) $Nt = AdjCon(A) / Wt(A) = .51 / 10 = .051$

(E14) $Nt = AdjCon(B) / Wt(B) = 25.5 / 500 = .051$

The training normalization constant with a range of words between 10 and 500 words is .051. The importance of this constant can now be illustrated. The total score, ***Sn***, for the pages in our example after one round of training can be found using the equation:

(E15) $Sn = Wt * Nt * (St - Sc) / Tp$,

where ***Tp*** is the total number of possibilities of word combinations.

It should be noted that the total number of possibilities is dependent upon such things as groupings and the manner in which the words are cycled through. For example, if the page has 100 words, we can take groups of 10 words and cycle through them in increments of 5. Taking such things into account the equation for ***Tp*** becomes:

$Tp = (Wt / Wi - 1) * (Wg)! / [(Wg - k)! (k)!]$

Where k is the k-set: k =1 for single words, k = 2 for word pairs, k = 3 for word triplets, etc. ***Wg*** is word groupings, ***Wt*** is word total, and ***Wi*** is word increment (or cycling). In the examples in discussed below, ***Wt*** is equal to ***Tp***. While this simplifies the examples provided herein, it is not necessarily the case when k > 1.

In the special case where ***Wt*** = ***Tp***, the amount that the relevance score for each word will be raised is:

(E16) ***Nt * (St - Sc) or .051 * (St – Sc)***

This is a simplified example, but illustrates the basic principles of normalizing word scores in the training process. Note that for k > 1 (or anything other than single word counts), ***Wt is not equal to Tp.***

It should also be appreciated that this normalization process can be used to not only train lexical elements to be associated with a particular sites (up-train), it can also be used to train lexical elements to not be associated with a particular site (down-train). During an up-training session, the word relevance scores of lexical elements on a page are increased within the designated category to indicate that they are more strongly associated with the category.

During a down-training session, the word relevance scores of lexical elements on a page are reduced to indicate that they are less strongly associated with a chosen category. Accordingly, it should be realized that to down train a page, the normalization constant would be calculated to move the score of each page downward to, for example, a score of 500. Thus, each lexical element on the page would be multiplied by a normalization constant that resulted in a lowered value for the page relevance score.

However, in either case, it is advantageous to normalize the amount that each word relevance score changes so that a page with fewer lexical elements does not more greatly affect the word relevancies found on that page.

Example 2

Normalizing Internet Page Scoring

If words, word pairs and word adjacencies are "trained up" by approximately the same value so that each has a gradually greater relevance score, then how does that affect the page scoring process. Assume two sites A and B, have 10 and 500 words on them respectively. Each has a score of 0 before one round of training and the target score is 1000. Since we are dealing with single words, k = 1, then ***Wt*** = ***Tp***. Using equation 16, we find that the amount each word will be incremented is:

(E17) ***.051 * (St – Sc) = .051 * (1000 – 0) = 51***

If each word was raised 51 points, then the score of each page after one round of training would be 51 times the number of words on that page. The score for each page is:

(E18) ***Score(A) = 10 * 51 = 510***

(E19) ***Score(B) = 500 * 51 = 25500***

Obviously, these scores are not close to each other. Judging solely upon the numbers, it would seem that site B was much more relevant to a category than site A. However, we used them both to train the same category. Consequently, they should have similar values after one round of training. We need a system that takes into account the skew that pages with varying numbers of words can create.

What we want to accomplish is to create some means of normalizing scores of pages based on the number of words that occur on them. Using equations 18 and 19, we can approximate the maximum and minimum scores for sites. Since we want the min and max to approach each, we can find their midpoint using the midpoint formula:

(E20) ***(510 + 25500) / 2 = 13005***

Finding the *"adjustment variables"* for this set of data requires dividing the midpoint score by the real score:

(E21) ***Ns(A) = 13005 / 510 = 25.5*** ***(note: Wt = 10)***

(E22) ***Ns(B) = 13005 / 25500 = .51*** ***(note: Wt = 500)***

We now know the points (10 words, 25.5) and (500 words, .51). If we find a few more points (255, 1), (132, 1.931818), and (378, 0.674603) and plot them, we get
an ordered data set with a trendline that has the equation:

(E23) ***y = 255 * x ^ -1***

Substituting in the ***Ns(Wt)*** for ***y*** (which is the score normalizer, given a set number of words) and ***Wt*** (total words) for ***x***. We get the equation:

(E24) ***Ns(Wt) = 255 * (Wt) ^ -1***

For our sites A and B with 10 and 500 words:

(E25) ***Ns(10) = 255 * (10) ^ -1 = 25.5***

(E26) ***Ns(500) = 255 * (500) ^ -1 = .51***

In general, the scoring equation becomes:

(E27) ***Normalized Score(Site X) = Ns(Wt(Site X))*Original Score(Site X) .***

Using the results from equations 18 and 19, the scores of site A and site B were 510 and 25,500, respectively. Using the normalized score technique, after one round of training the scores of these sites would be:

(E28) ***Normalized Score(A) = Ns(Wt(A)) * Score(A) = 25.5 * 510 = 13005***

(E29) ***Normalized Score(B) = Ns(Wt(B)) * Score(B) = .51 * 25500 = 13005***

The sites have the same score after training. This supports the logic that sites that are used to train a category should have similar scores. These equations, in combination with the normalization of training data, as shown in Example 1, minimizes the error caused by having sites with different numbers of words on them in a training set.

Example 3

Scoring a Page

Approximately 8000 samples were collected from sites from the Category Two (or Sex 2) of the Suggest Database. These potential category two sites had previously been checked by Web Analysts to determine whether they were, in fact, Internet sites that were primarily sexual or pornographic in nature. A score of 8 was assigned to a site that was verified as a sex site and a score of 7 to those sites that were determined not to be sex sites. The categorization system had assigned a category rating for category two to all 8000 Sites.

The purpose of the study was to determine whether the categorization system could distinguish between sites rated as 8's and 7's, or accepted sites and deleted sites, respectively. It should be noted that a deleted site is one that should not have been categorized within the Sex category and an accepted site is one that was confirmed to be within the category. The hypothesis was that the mean score for the sites rate as 8's would be statistically different from the mean score for sites rated as 7's. As suspected, the mean for the accepted sites (8's) were significantly higher than the mean for the deletions (7's). However, there was an overlap between the two groups. This result suggests that the use of a cutoff point could be used to minimize the error involved.

	Mean Score	**Standard Deviation**	**Median**
7's (deletions)	929	482	842

Alpha Point = **Ap** = **M7** + 4 **(SD7)** = 929 + 4 (482) = **2857**

Beta Point = **Bp** = **M7** + 1 **(SD7)** = 929 + 1 (482) = **1411**

Using an alpha point of 2857 we found a sorting accuracy of 99% or above. There were only 9 sites that were above the alpha score, but did not belong within the Sex category. Seven of them were simple errors, perhaps attributable to poor training of the Category 2 sites.

Two of them were purposeful tricks, meaning that the Internet sites used sex-related terms to attract attention in their metatags. The exact percentage for the sorting accuracy, using the alpha point of 2857, was therefore 99.30%. Thus, according to this test, if a thousand sites were entered with a score above this alpha point there will be, on average, only 7 mistakes and 993 correctly sorted sites.

However, because the alpha point is set very high, many sites that are, in fact, sexually oriented, will not be categorized at all. Using an alpha point of 2857, the inclusion level of accepted sites is only 49.80%. This means that out of a thousand sites that should be placed in category two, 498 would be found and 502 missed.

For this reason, the system also monitors sites that have a lower relevance to each category through creation of a beta point. Using a beta point of 1411, the inclusion level rises from 49.80% to 81.76%. The number of sites missed falls from 502 to 183 sites, and the number caught rises from 498 to 817. Thus, the use of both the alpha and beta points results in more accurate scoring of any new site.

Example 4

Normalizing Training Data by Increments

Another embodiment of a method for normalizing training data is explained below. First, we define Is = initial score and Ts = target score for the page being trained.

1) Begin with a test increment value of, for example, 1. Increment the values of the relevance of all lexical values by the test value. (e.g.: all lexical values existing on the page).

2) Calculate the resulting page relevance score after this test addition.

3) If the new score = Ms., the increment value, I, (for all lexical elements) =

$$I = (Ts - Is) / (Ms - Is)$$

Thus, the difference between the target score and the current score, divided by the effect on the score when each elements relevance is incremented by 1 is the correct number to Increment each element to achieve the target score.

Accordingly, if the Is = 500, and Ts = 1000 incrementing all relevancies by 1 will result in a page score of 550 and:

$$I = (1000 - 500) / (550 - 500).$$

Therefore, to increment the page to result in a page score of 1000, we need to use an increment value is 10 for each lexical element.

In general the relevance for a value will be incremented by the Increment constant (I) * the # of occurrences of that element on the page. This follows from the notion that the more often an element appears on a page the more relevant it is. However, this process resulted in large fluctuations in the relevance of elements that would occur frequently, but were not common words. For this reason, in one embodiment, each value was only allowed to increment by a maximum 5 * increment constant (I).

WHAT IS CLAIMED IS:

1. A computerized system for controlling access to Internet sites, comprising:

a first module that categorizes an Internet site into predefined subject-matter categories;

a second module that stores the address of said Internet site and its associated category to a database; and

a third module that controls access to said Internet site, said third module comprising instructions that block access to said Internet site if said Internet site is within a particular category.

2. The computerized system of Claim 1, wherein said first module categorizes said Internet site by determining the relevance of lexical elements within said site to a subject-matter category.

3. The computerized system of Claim 2, wherein said lexical elements are selected from the group consisting of: words, word pairs and word adjacencies.

4. The computerized system of Claim 2, wherein the relevance of said lexical elements to said subject matter category is determined by comparing said lexical elements to a training database comprising a plurality of lexical elements and their associated relevancies to subject matter categories.

5. The computerized system of Claim 1, wherein said subject matter categories are selected from the group consisting of: sex, drugs, sports, shopping, alcohol, gambling, politics, religion, travel and violence.

6. The computerized system of Claim 1, wherein said third module comprises a permissions table for determining whether a user has access rights to Internet sites within a particular category.

7. A database of categorized Internet addresses, produced by the method of:

retrieving a first Internet page;

parsing said Internet page to determine the lexical elements on said page;

comparing said lexical elements to a table of category relevancies to determine the relevance of each lexical element to a subject matter category;

determining the subject matter category of said Internet page based on the relevance of each lexical element to a category; and

storing the address and subject matter category of said Internet page to a database.

8. The database of Claim 7, wherein said lexical elements are selected from the group consisting of: words, word pairs and word adjacencies.

9. The database of Claim 7, wherein the relevance of said lexical elements to said subject matter category is determined by comparing said lexical elements to a training database comprising a plurality of lexical elements and their associated relevancies to subject matter categories.

10. The database of Claim 7, wherein the table of category relevancies is produced by obtaining an Internet page from within a known subject matter category and storing lexical elements from said page to said table along with relevance scores indicating that said lexical elements are associated with said category.

11. A method of controlling access to an Internet site, comprising:

categorizing a first Internet site into a predefined subject-matter category;

storing the address of said Internet site and its associated category to a database;

capturing a user request to view said site; and

determining whether said user has permission to view said category of sites and, responsive to said determination, controlling access to said site.

12. The method of Claim 11, wherein said Internet site is categorized by determining the relevance of lexical elements within said site to a subject-matter category.

13. The method of Claim 12, wherein said lexical elements are selected from the group consisting of: words, word pairs and word adjacencies.

14. The method of Claim 12, wherein the relevance of said lexical elements to said subject matter category is determined by comparing said lexical elements to a training database comprising a plurality of lexical elements and their associated relevancies to subject matter categories.

15. The method of Claim 11, wherein said subject matter category is selected from the group consisting of: sex, drugs, sports, shopping, alcohol, gambling, politics, religion, travel and violence.

16. The method of Claim 11, wherein determining whether said user has permission to view said category of sites comprises referring to a permissions table that stores user identifications and access rights to subject matter categories.

17. A method of controlling access to Internet sites, comprising:

providing a training database, said training database comprising lexical elements and their relevance to subject matter categories;

determining the relevance of a plurality of Internet pages to subject matter categories;

storing the address of said plurality of Internet pages and their relevance to subject matter categories to a categorized database; and

comparing the Internet address of pages requested by users with said categorized database to determine if said users have permission to view said pages.

18. The method of Claim 17, wherein said Internet pages are categorized by determining the relevance of lexical elements within said site to a subject-matter category.

19. The method of Claim 18, wherein said lexical elements are selected from the group consisting of: words, word pairs and word adjacencies.

20. The method of Claim 18, wherein the relevance of said lexical elements to said subject matter category is determined by comparing said lexical elements to a training database comprising a plurality of lexical elements and their associated relevancies to subject matter categories.

21. The method of Claim 17, wherein said subject matter categories are selected from the group consisting of: sex, drugs, sports, shopping, alcohol, gambling, politics, religion, travel and violence.

22. The method of Claim 17, wherein determining whether said user has permission to view said category of sites comprises referring to a permissions table that stores user identifications and access rights to subject matter categories.

5

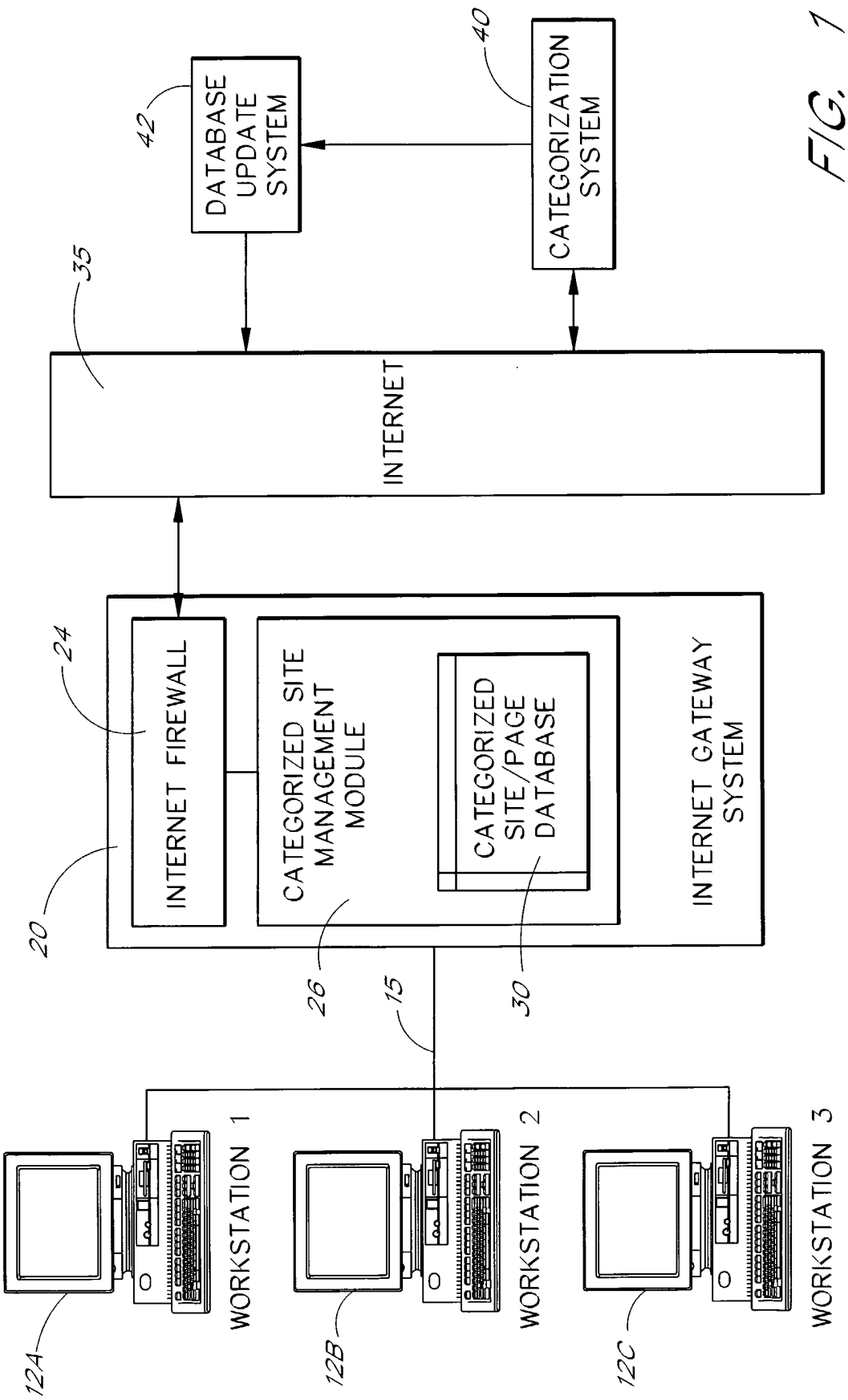

DATABASE UPDATE SYSTEM
42
CATEGORIZATION SYSTEM
40
35
INTERNET
24
INTERNET FIREWALL
CATEGORIZED SITE MANAGEMENT MODULE
CATEGORIZED SITE/PAGE DATABASE
INTERNET GATEWAY SYSTEM
20
26
30
15
WORKSTATION 1
WORKSTATION 2
WORKSTATION 3
12A
12B
12C

FIG. 1

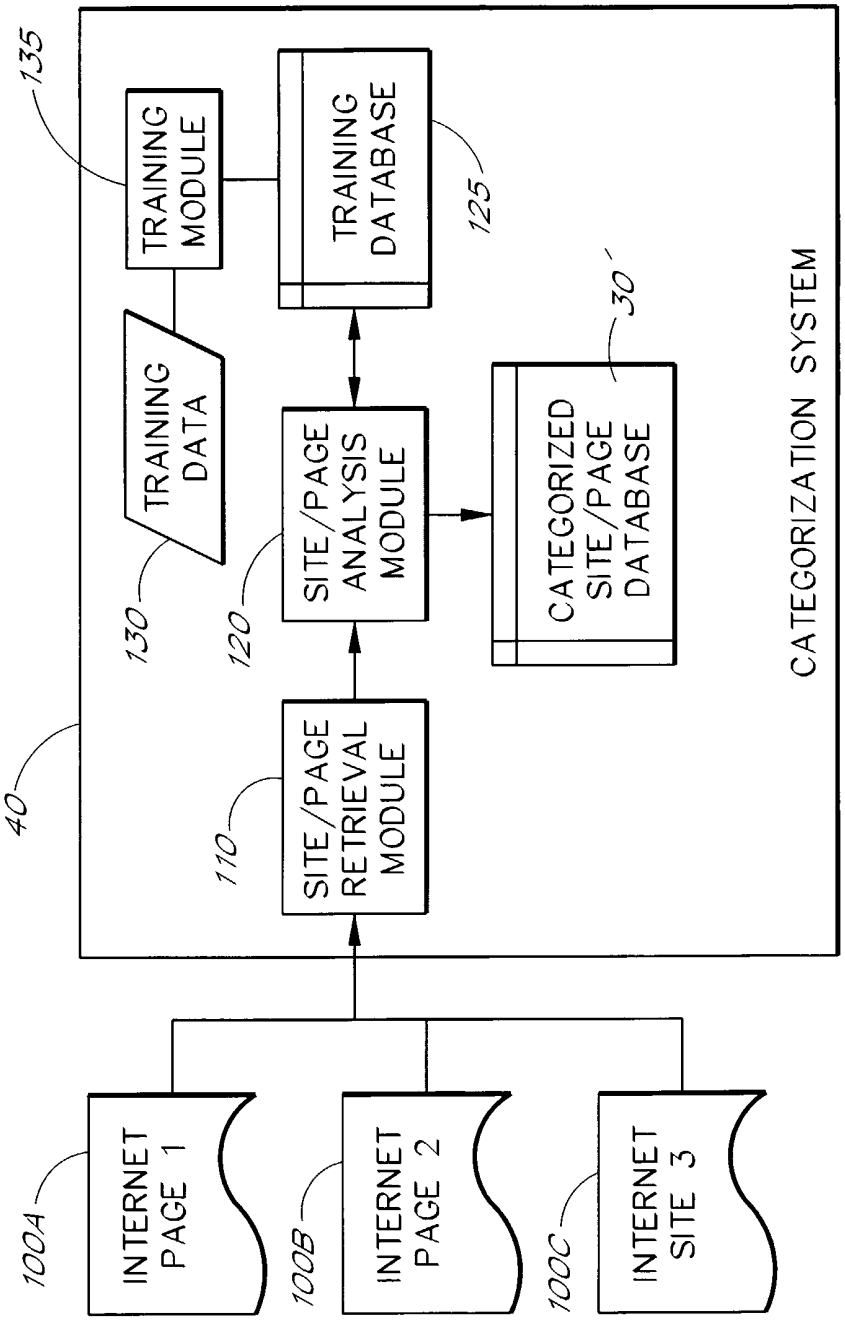

135
TRAINING MODULE
TRAINING DATABASE
125
TRAINING DATA
130
120
SITE/PAGE ANALYSIS MODULE
CATEGORIZED SITE/PAGE DATABASE
30'
CATEGORIZATION SYSTEM
40
110
SITE/PAGE RETRIEVAL MODULE
100A
INTERNET PAGE 1
100B
INTERNET PAGE 2
100C
INTERNET SITE 3

FIG. 2

125

WORD ID TABLE 200

WORD	ID NUMBER
THE	1
AND	2
CLEVELAND	234
BROWNS	198
DIAMOND	755
JEWELRY	1345

CATEGORY ID TABLE 205

CATEGORY	ALPHA SCORE	BETA SCORE	DESCRIPTION
1	920	810	SPORTS
2	860	705	SEX
3	880	740	SHOPPING
4	710	630	PORNOGRAPHY
5	940	865	LEISURE
6	860	690	RELIGION

WORD RELEVENCE TABLE 210

FIRST WORD ID	SECOND WORD ID	CATEGORY	RELEVENCE SCORE	PAIR	ADJACENT
234	198	1	900	N	Y
755	1345	1	290	N	Y
755	1345	3	940	Y	N
234	1345	1	250	N	Y

TRAINING DATABASE

FIG. 3

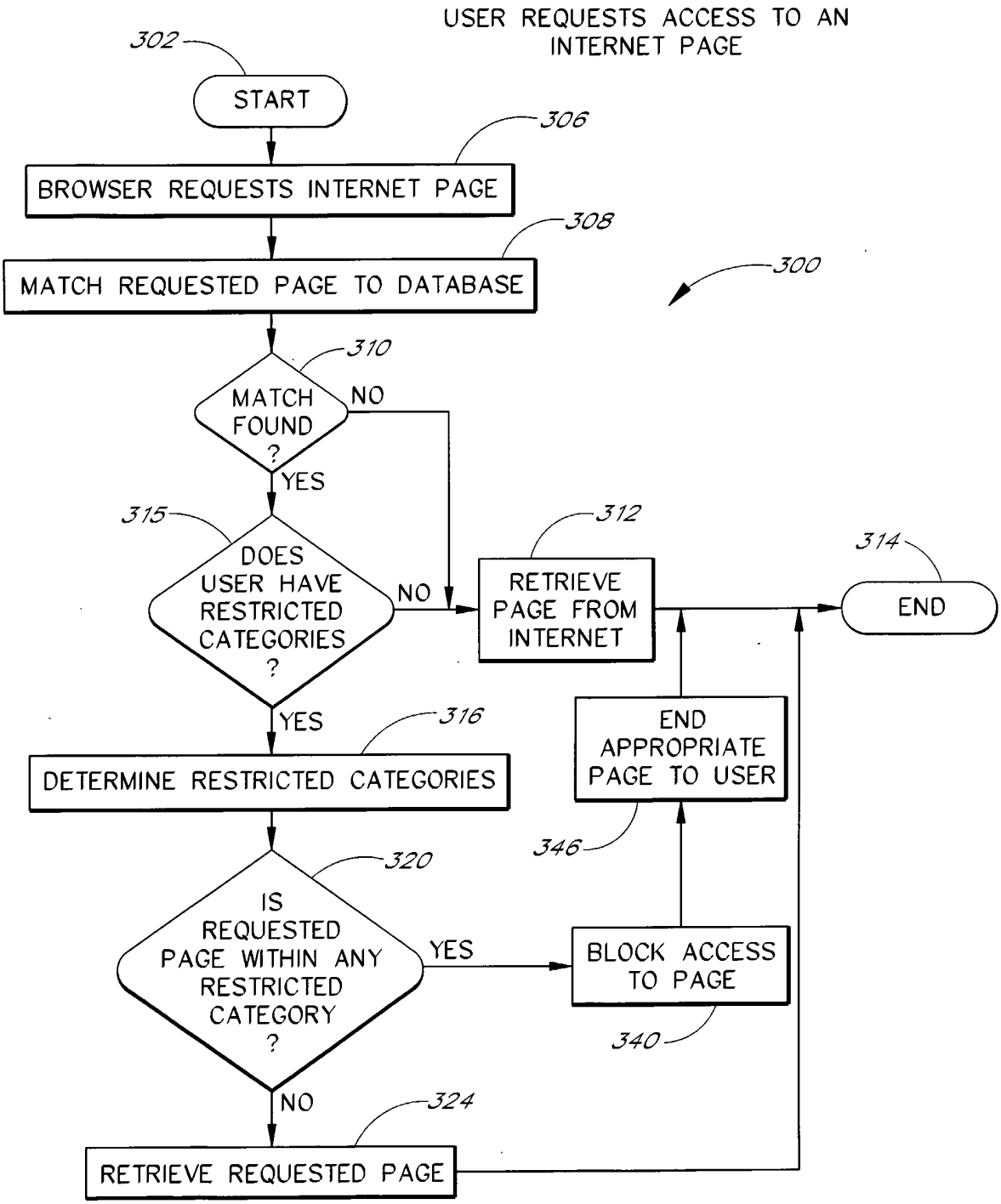

USER REQUESTS ACCESS TO AN INTERNET PAGE
302
START
306
BROWSER REQUESTS INTERNET PAGE
308
MATCH REQUESTED PAGE TO DATABASE
300
310
MATCH FOUND ?
NO
YES
315
DOES USER HAVE RESTRICTED CATEGORIES ?
NO
312
RETRIEVE PAGE FROM INTERNET
314
END
YES
316
DETERMINE RESTRICTED CATEGORIES
END APPROPRIATE PAGE TO USER
346
320
IS REQUESTED PAGE WITHIN ANY RESTRICTED CATEGORY ?
YES
BLOCK ACCESS TO PAGE
340
NO
324
RETRIEVE REQUESTED PAGE

FIG. 4

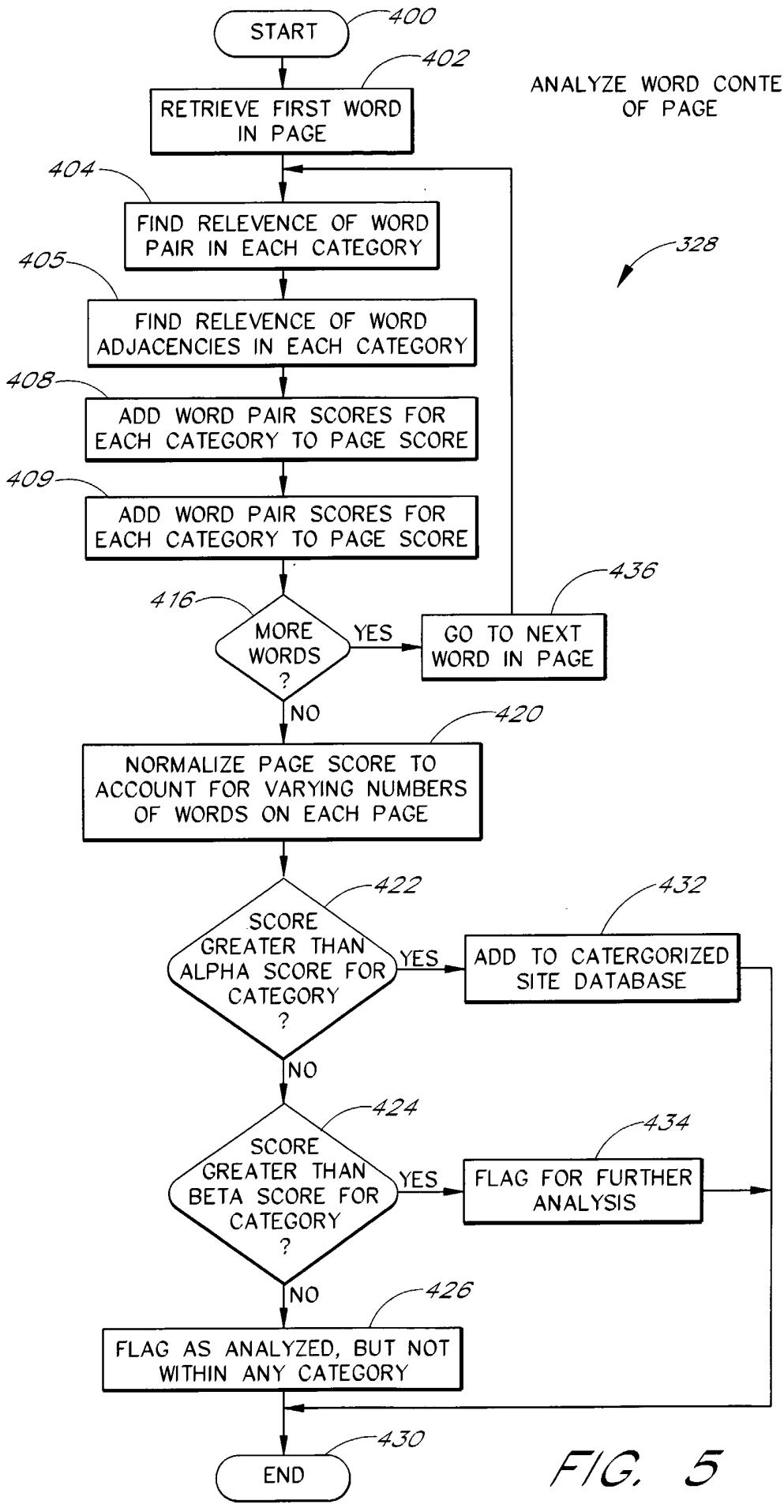

ANALYZE WORD CONTENT OF PAGE
328
START
400
RETRIEVE FIRST WORD IN PAGE
402
FIND RELEVENCE OF WORD PAIR IN EACH CATEGORY
404
FIND RELEVENCE OF WORD ADJACENCIES IN EACH CATEGORY
405
ADD WORD PAIR SCORES FOR EACH CATEGORY TO PAGE SCORE
408
ADD WORD PAIR SCORES FOR EACH CATEGORY TO PAGE SCORE
409
MORE WORDS ?
416
YES
GO TO NEXT WORD IN PAGE
436
NO
NORMALIZE PAGE SCORE TO ACCOUNT FOR VARYING NUMBERS OF WORDS ON EACH PAGE
420
SCORE GREATER THAN ALPHA SCORE FOR CATEGORY ?
422
YES
ADD TO CATERGORIZED SITE DATABASE
432
NO
SCORE GREATER THAN BETA SCORE FOR CATEGORY ?
424
YES
FLAG FOR FURTHER ANALYSIS
434
NO
FLAG AS ANALYZED, BUT NOT WITHIN ANY CATEGORY
426
END
430

FIG. 5

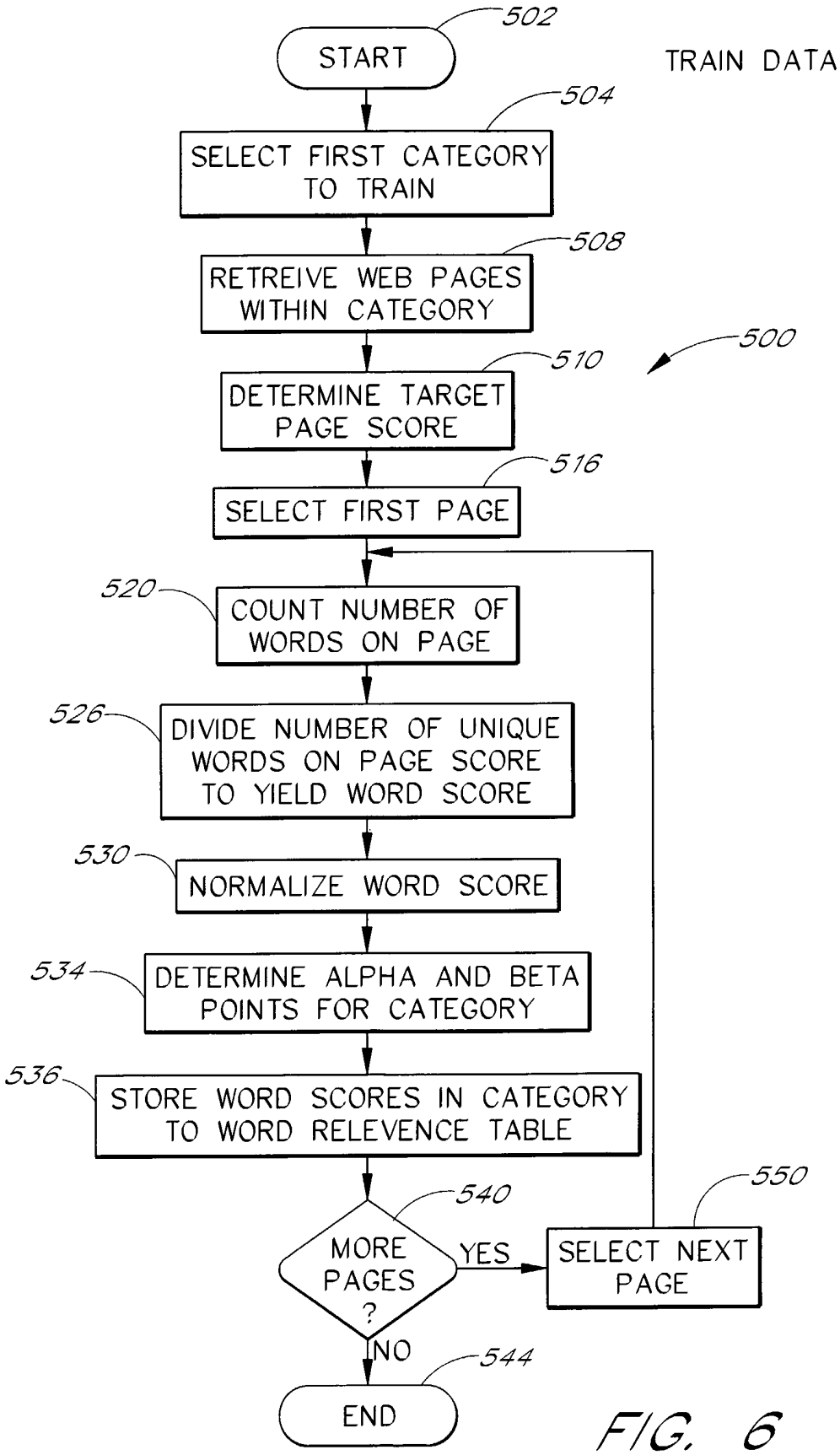

TRAIN DATA
500
502
START
504
SELECT FIRST CATEGORY TO TRAIN
508
RETREIVE WEB PAGES WITHIN CATEGORY
510
DETERMINE TARGET PAGE SCORE
516
SELECT FIRST PAGE
520
COUNT NUMBER OF WORDS ON PAGE
526
DIVIDE NUMBER OF UNIQUE WORDS ON PAGE SCORE TO YIELD WORD SCORE
530
NORMALIZE WORD SCORE
534
DETERMINE ALPHA AND BETA POINTS FOR CATEGORY
536
STORE WORD SCORES IN CATEGORY TO WORD RELEVENCE TABLE
540
MORE PAGES ?
YES
550
SELECT NEXT PAGE
NO
544
END

FIG. 6

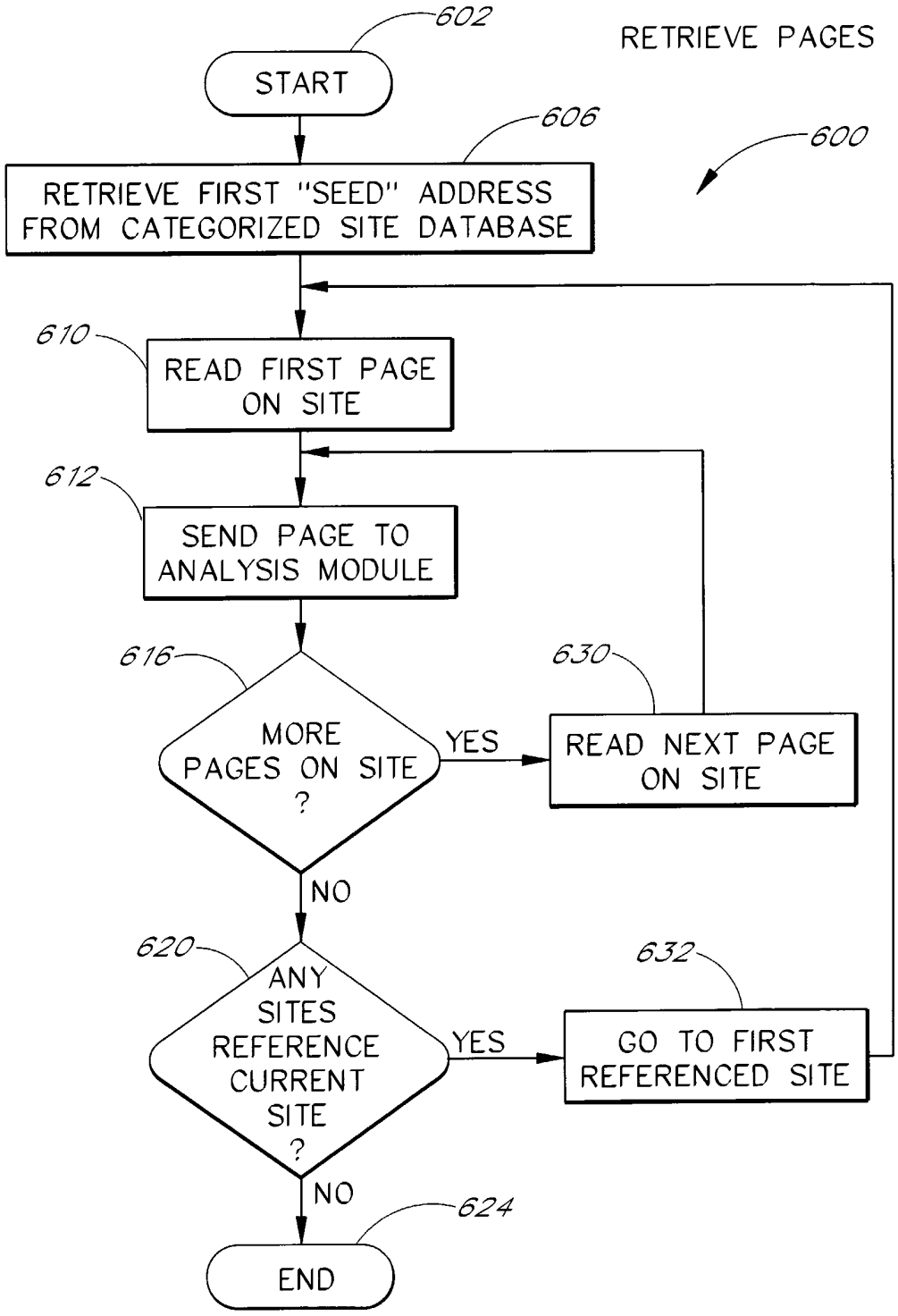

RETRIEVE PAGES
600
602
START
606
RETRIEVE FIRST "SEED" ADDRESS FROM CATEGORIZED SITE DATABASE
610
READ FIRST PAGE ON SITE
612
SEND PAGE TO ANALYSIS MODULE
616
MORE PAGES ON SITE ?
YES
630
READ NEXT PAGE ON SITE
NO
620
ANY SITES REFERENCE CURRENT SITE ?
YES
632
GO TO FIRST REFERENCED SITE
NO
624
END

FIG. 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/02280

A. CLASSIFICATION OF SUBJECT MATTER
IPC(7) :G06F 17/30, G06F 11/00
US CL : 707/9, 10; 709/203, 217, 221, 255
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)
U.S. : 707/9, 10; 709/203, 217, 221, 255

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5,832,212 A (CRAGUN et al) 03 November 1998, abstract, fig.1, col.3 line 30 to col.4 line 30, col.5 line 35 to col.6 line 10.	1-22
Y	US 5,835,722 A (BRADSHAW et al) 10 November 1998, abstract, figs.1, 2, col.7 lines 8-65, col.9 lines 1-31, col.10 lines 4-46.	1-22
A	US 5,848,412 A (ROWLAND et al) 08 December 1998, fig.1, col.3 line 36 to col.4 line 29.	1-22

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"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
"A"	document defining the general state of the art which is not considered to be of particular relevance	"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
"E"	earlier document published on or after the international filing date	"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
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"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		

Date of the actual completion of the international search	Date of mailing of the international search report
19 APRIL 2000	25 MAY 2000
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer AHMAD MATAR Telephone No. (703) 305-4731

Form PCT/ISA/210 (second sheet) (July 1998) ★