



(51) International Patent Classification:
G06F 17/30 (2006.01)

(21) International Application Number:
PCT/US2016/048552

(22) International Filing Date:
25 August 2016 (25.08.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
14/863,925 24 September 2015 (24.09.2015) US

(71) Applicant: INTEL CORPORATION [US/US]; 2200
Mission College Boulevard, Santa Clara, CA 95054 (US).

(72) Inventors: RASHID, Al, M.; 964 Hildebrand Circle, Fol-
som, CA 95630 (US). ZHANG, Sushu; 9543 S. Stanley
Place, Tempe, AZ 85284 (US). KUHN, Robert, H.; 1420
Riverbend Lane, White Heath, IL 61884 (US).

(74) Agents: ROZMAN, Mark, J. et al.; Trop, Pruner & Hu,
P.C., 1616 S. Voss Rd., Ste. 750, Houston, TX 77057-
2631 (US).

(81) Designated States (unless otherwise indicated, for every
kind of national protection available): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY,
BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM,
DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,
HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR,
KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG,
MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM,
PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC,
SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN,
TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every
kind of regional protection available): ARIPO (BW, GH,
GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ,
TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU,
TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE,
DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU,
LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK,
SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ,
GW, KM, ML, MR, NE, SN, TD, TG).

Published:

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

(54) Title: CLIENT-SIDE WEB USAGE DATA COLLECTION

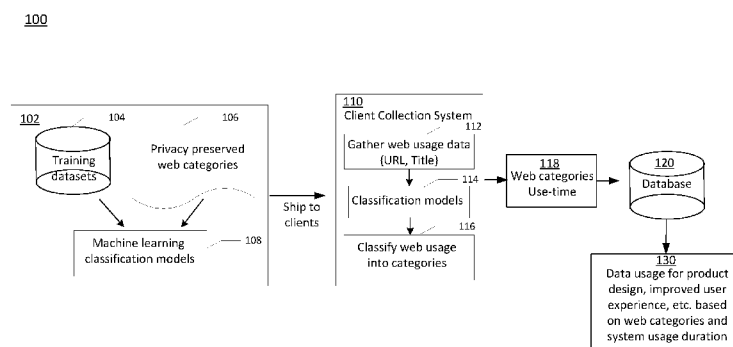


FIG. 1

(57) Abstract: In an embodiment, a system includes a processor that includes at least a first core that includes collection logic to re-
cord a history of website accesses of a plurality of websites by a user. The first core also includes classification logic to assign the
website accesses to corresponding categories by application of a plurality of models, where each model corresponds to a respective
category, and to determine a classification summary that includes a plurality of category metrics, each category metric associated
with the respective category, each category metric based on a corresponding measure of the website accesses within the respective
category. The classification summary suppresses a corresponding identity of each website accessed. The system also includes a non-
volatile memory coupled to the processor. Other embodiments are described and claimed.

CLIENT-SIDE WEB USAGE DATA COLLECTIONTechnical Field

[0001] Embodiments pertain to client side web usage data collection.

Background

[0002] To design systems competitively, some original equipment manufacturers (OEMs) use data collected on end-user systems. Increasingly, browser usage constitutes a significant part of personal computer usage, and therefore understanding how various types of users use browsers differently may be of importance to understand market segment requirements of personal computers.

[0003] Some web services collect raw data on servers including browser cookie tracking, for data-mining on the servers. However, raw browser usage data is private information, and collecting personal computer (PC) users' browsing behavior data in a privacy-preserving and unobtrusive way may be difficult.

[0004] Some solutions may be web service-based, requiring raw uniform resource locators (URLs) to be captured between users' requests and websites visited, potentially leaving the user system with a privacy/security risk. Additionally, the web service may log the user's Internet Protocol (IP) address and the URL may even contain personal information such as user name. Further, some solutions are intrusive in that they require a browser plugin or network sniffing.

[0005] Many secure browsing web services offer only binary classes, e.g., "child-friendly or not," "malicious or not," and are geared toward providing specific services to customers, e.g., parental control. Some solutions work for only broad categorization such as a top level URL domain, e.g., www.youtube.com, which may produce little to no useful information.

Brief Description of the Drawings

[0006] FIG. 1 is a block diagram of a process, according to embodiments of the present invention.

[0007] FIG. 2 is a block diagram of a system, according to an embodiment of the present invention.

[0008] FIG. 3 is a flow diagram of a method, according to an embodiment of the present invention.

[0009] FIG. 4 is a flow diagram of a method, according to another embodiment of the present invention.

[0010] FIG. 5 is a flow diagram of a method according to another embodiment of the present invention.

[0011] FIG. 6 is a block diagram of an example system with which embodiments can be used.

Detailed Description

[0012] In embodiments, if a user opts in, a system can collect the user's browsing history and classify entries into high level system impact categories, e.g., using machine learning techniques. The usage by categories may be sent to a server to represent browser usage of system components. In embodiments, the site names do not leave the client system, to prevent URLs selected by the user from becoming public knowledge.

[0013] The following set of guidelines may be used in embodiments:

1. Privacy. Raw URLs do not leave a user's system. Instead, raw URLs are turned into web categories using decentralized classification (also categorization herein) models. Private information does not leak from one site to another, as with cookies.

2. Unobtrusiveness.

- Avoid browser plugins, which may pose a security risk.

- Avoid packet sniffing. In an embodiment, categories may reference computer system function and performance characteristics rather than users' specific actions on the web. For example, multiple forms of online video watching, including even objectionable content, may be mapped to a 'video streaming' category. Sites that typically use secure communication may be mapped into a 'security required' category, e.g., a shopping site or a bank site. In embodiments, a classifier may transform information about the user into data that pertains to architectural requirements, in order to design more effective systems. The classifier may output an estimated error rate (e.g., confidence level), which can be used in data analysis.

[0014] The approach presented herein is capable of classifying a broad range of web site categories by computer system behavior, and may be utilized to determine system component usage for PC designers. Classification may be based on the entire URL, so that most frequently used pages within a domain can be characterized.

[0015] Embodiments include machine learning models that can be tuned to any number of categories so as to be appropriate to a privacy sensitivity of each user, addressing common privacy guidelines. For example, specialized user experience studies may make use of machine learning models that correspond to a detailed list of fine-grained categories, e.g., to be applied with users who opt in to a detailed usage collection. On a general usage system, “fuzzier” and smaller number of categories may be used, e.g., resulting in on-client models that may be much smaller and faster. Because cookies are not used in the embodiments presented herein, the models in the embodiments presented would be difficult to be co-opted for unintended purposes, e.g., for information gathering such as specific URLs accessed by a user.

[0016] Another benefit of the client side decentralized approach is that the overall computation can be treated as massively parallel, in contrast to a web services-based approach where a number of page hits to the web service from all the clients can be huge, potentially requiring an expensive server infrastructure investment.

[0017] FIG. 1 is a block diagram of a process, according to an embodiment of the present invention. Process 100 includes three phases: model building 102, data collection and classification 110, and server data processing 130.

[0018] A first phase 102 is model-building. This is an offline model preparation phase that uses machine learning and text mining. Models generated are able to predict one or more web-categories, given a URL and some page title information.

[0019] In an embodiment, phase 102 proceeds as follows:

1. Construct training data. Sample URL data (title, description included) may be gathered from website classification sites, e.g., dmoz.org, parsed, and stored in an analyzable format. Also to be downloaded is data about website popularities, e.g., numerical ranking of URLs according to popularity (e.g., frequency of hits in a defined time period).

2. Determine/prune category names. There are too many (>14,000) categories in a dmoz dataset. However, a typical description in a dmoz dataset may be intended to characterize user usage rather than system usage. As an example, a user may not wish to report the following categories: tobacco (subset of shopping); Minnesota (subset of banking); gambling (subset of games). Instead, more generic categories such as “shopping” and “games” may be preferable (e.g., less revealing of user lifestyle) over “tobacco” and “gambling”.

Categories may be pruned using the following algorithm:

Initially, categories are organized in a hierarchy/tree. Each path through nodes from root to a leaf in this tree forms a category. For example, by calling the root of the tree “top,” the following is a category: top→arts→animation→anime→titles→d→digimon→characters. Each of the “top,” “arts,” “animation,” ..., “characters” represents a node in the tree. A goal is to eliminate most of these nodes, and treat the set of the remaining leaf nodes as the pruned set.

Consider URLs from dmoz that matches with URL popularity dataset and build a hierarchy of the categories, as present in dmoz. Initially, there are typically >14,000 nodes in the tree, as found in the dmoz dataset. Each node includes two computed statistics. The first statistic is an average weight of the URLs it is associated with the node.

Weight W_u of a URL u may be expressed according to the following:

$$W_u = -\log_2 (R_u/2N)$$

where R_u is the rank of the URL, and N is the total number of popular URLs considered,

e.g., $N \approx 10^6$. The most popular URL has $R_u=1$. The second statistic of each node is how many URLs fall under the node.

The hierarchy tree can be pruned recursively based on the number of URLs covered and average weight (importance/popularity) of the URLs in the sub-tree, until a desired number of categories are left, e.g., 10-50 categories. That is, starting from the root, traversing through a branch, and stops proceeding through that branch if the last node toward the leaf does not have enough average weight or large enough number of URLs. The last node visited on that branch is one of the categories. This iterative process also considers category-filtering,

eliminating a set of categories that might be too sensitive to include, e.g., “Adult,” “LGBT,” etc. Finally, review of the categories is conducted and a subset, e.g., 10-30 different categories are selected from the approximately 14,000 categories, to use as a set of categories for classification.

3. Build models. Model building may include preparation of a dataset of {URL, textual description, category} using the selected categories. The dataset is effectively a set of examples from which to learn. Each example has some textual information, e.g., URL and description of the website, and the category. The textual information is tokenized to derive features, which provide hints to the corresponding category. For example, for the URL “linkedin.com,” the description may be: “a networking tool to find connections to recommended job candidates, industry experts and business partners.” One way to tokenize example is to split by words, which gives the following features for this example: linkedin, networking, tool, find, connection, recommend, job, candidate, industry, expert, business, partner. The original category of this URL was “top/computers/internet/on_the_web/online_communities/social_networking,” which after pruning becomes “online_communities.” The tokenized features in each example are treated as (feature) vectors. A total number of features can be huge, and too many features or variables can lead to inferior models. Therefore, the feature space is then reduced using L1 regularization (also known as Lasso Penalty regularization). In L1 regularization, the best model is the one that minimizes prediction error, and has fewer features (variables).

The classification models are then built via linear support vector machine (SVM) or logistic regression with regularization to keep the models generalizable and effective. Typically one model is built for each category. The models may be tested with cross-validation for any improvement required. In cross validation, the available data is randomly split into n-ways, and models are built using (n-1) splits, and the learned model is tested against the remaining split. Each model is to be saved as a corresponding file. Since each model is a linear combination of textual features for a category, each model may include all coefficients (or weights) learned for all of the textual features. For example, in one embodiment in the case of logical regression, the learned model for a category c_j may be expressed as

$$P(Y=c_j) = 1/(1 + e^{-(\beta_0 + \sum \beta_i f_i)}) ,$$

where the learned coefficients β_i corresponding to the tokenized textual features, f_i , are saved as models. Maximization of distinction between categories (e.g., selection of non-overlapping categories) can enhance utility of the categories.

The models are to be shipped to the client systems along with a collector (e.g., software to perform the data collection).

[0020] A second phase 110 includes data collection and classification. A low intensity collector in the client system, e.g. personal computer (PC), gathers web usage data 112 that includes minimal browsing history data (e.g., URLs and page titles) and system utilization, e.g., CPU consumption, by the web sites visited. The history data is then tokenized and passed into a classifier 116 to perform a classification, e.g., determine a corresponding category in which to place each URL. The classifier 116 uses the classification models 114 learned in phase 102 to determine output 118 that includes a quantitative classification of the web site accesses, to be sent to a database 120. The classification suppresses the identity of each website, and instead presents a quantitative measure of website access (e.g., based on website access frequency and website access durations) according to each category.

[0021] A third phase 130 is server data processing. Anonymous and de-identified information is uploaded to the server from the database 120, e.g., for analysis. The analysis may be used as system use feedback in analytics that may, e.g., influence product improvement of components, design specifications of hardware or software, etc.

[0022] The above-described approach includes a trained/learned information transformation algorithm that produces compression of information with intentional loss of precision, while focusing on de-identifying personal information. Categories can be coarse and privacy-preserving. An algorithm may be invoked to automatically prune thousands of fine-grained categories (e.g., retrieved from dmoz.org) into a smaller number of categories. A further refinement process may be invoked to preserve privacy of categories, e.g., through a filter that provides “sanity checks” constructed according to privacy principles e.g., developed by privacy experts and via user studies. The user studies or surveys can be conducted periodically, e.g., annually, semi-annually, etc., and may be automated. In one embodiment, the final number of categories to be used for classification is between 10 and 100.

[0023] In embodiments, classification (e.g., category determination) of URLs happens locally on the user's system, unlike many solutions where the explicit URLs are sent to a web service that potentially exposes the user's IP address and where the web server can store sensitive web usage data server.

[0024] In embodiments, a non-intrusive, secure collector is used. The collector is neither a plug-in to the browsers that can make browsers unstable and pose security risks, nor it is a network packet sniffer.

[0025] FIG. 2 is a block diagram of a system according to embodiments of the present invention. System 200 is a personal computer that includes a processor 210 and a non-volatile memory 218. The processor 210 includes one or more cores 212₁ to 212_N. Core 212₁ may include collection logic 214 and classification logic 216. In embodiments, the nonvolatile memory 218 may store classification models 220, each model corresponding to a category. The system 200 may be coupled to a server 230.

[0026] In operation, the collection logic 214 (e.g., hardware, software, firmware, or a combination thereof) may be executed in the core 212₁ and upon execution may collect, during a usage period, a history of URLs (optionally including a title on a corresponding title page of each URL) accessed by a user and corresponding elapsed access times. The collection logic 214 can pass the collected history to the classification logic 216, which can classify the URLs according to the classification models 220 (e.g., developed according to model building described above) that are typically stored in the nonvolatile memory 218. For example, each classification model can indicate, based on URL information received, whether the URL in question falls in the category corresponding to the classification model. Generally, categories are constructed to be non-overlapping. Additionally, the categories are constructed so as to suppress detailed personal preference information, e.g., the URL of each website accessed.

[0027] A classification report that is output from the classification logic 216 may include a relative importance of each category determined from the URL access history received, e.g. a numerical value associated with the category for the particular access history being analyzed. The complete classification report (also classification summary, or categorization summary herein) for the particular URL access history typically may include a corresponding value for

each category based on, e.g., a count of URLs and access time of each URL. The classification report output suppresses (e.g., omits) the identity of each URL in order to protect privacy of the user. The classification report may be output to server 230.

[0028] The server 230 may store the classification report. The classification report may be used to determine modification of a future generation of the system 202. For example, the server 230 may collect many classification reports from various users and may analyze the classification reports received to produce an analysis that may point to inferences based on the populations of each of the categories. The analysis may be used as a basis, e.g., in analytics, to implement design changes, e.g., to effect improvement in utility of the system by users.

[0029] Referring to FIG. 3, shown is a flow diagram of a method according to an embodiment of the present invention. Method 300 is a method of developing classification models. Method 300 begins at block 302, where URL data is sampled and stored in an analyzable format. For example, the URL data may come from a source of URLs such as dmoz.com. Continuing to block 304, a URL ranking for each URL sampled may be determined based on a source of URL popularity rankings, e.g., from www.alexa.com. Advancing to block 306, categories may be determined based on URL rankings and a desired granularity of the categories. The desired granularity (e.g. number of categories) is an input to the algorithm. For example, in embodiments, a count of the categories created will be less than a count of URLs sampled, and the categories selected are intended to preserve privacy by suppressing URL titles and characteristics deemed too personal to be shared. For example, an expert filter (e.g., software, hardware, firmware, or a combination thereof) may be applied to the categories to filter out those categories deemed too personal to be shared (e.g., filtering out categories such as “adult movies”) and instead include more general categories (e.g., “movies”). The filter may be constructed by following common privacy guidelines, and from the outcome of user surveys that may reveal sensitivity to categories.

[0030] Moving to block 308, a subset of the determined categories may be selected, depending on the granularity specified. Proceeding to block 310, a classification model may be built for each category using L1 regularization, linear regression, etc. Each model is associated with a corresponding category and can provide a quantitative measure of a fit of a

URL to the corresponding particular category. The models may be used to determine in which category to place a URL that is logged, e.g., in a URL access summary of a user.

[0031] FIG. 4 is a flow diagram of a method according to another embodiment of the present invention. Method 400 begins at block 402, where a user's browsing history (e.g., list of URLs visited and length of time visited) is collected over a defined time period. Continuing to block 404, at the user's device, the URLs are classified into high level categories through use of classification models, the categories suppressing identities of the URLs and associated page titles. Suppression of the URL identities and titles pages is intended to protect privacy of the user. Advancing to block 406, a classification summary (e.g., system usage by category) is sent to a server. The classification summary is a representation of browser usage of a user by category (e.g., based on instances of website access and duration of each access), and may, along with other classification summaries sent from other users' PCs, be analyzed to provide as input for product design and/or modification, e.g., to effect improvement of system components of the user's PC.

[0032] FIG. 5 is a flow diagram of a method according to another embodiment of the present invention. Method 500 begins at block 502, where a server collects system usage classification data from each of a plurality of users (e.g., users that are participants in a usage study) via the user's personal computer. In embodiments, the classification data includes a category population count of websites accessed by a user over a defined time period, and may also include access duration of each access instance. Each accessed website is to be classified within one of a defined set of categories (e.g., non-overlapping) that are privacy-preserving. Privacy preservation is achieved through initial selection of the defined categories. For instance, the categories may be selected so as to suppress an identity (e.g., URL) of the websites to be classified, and categories may be selected so that a classification (e.g., classification data from a user) reflects system usage of the personal computer (PC) of the user, e.g., categories may be determined in part through use of a filter to filter out categories that reveal personal preferences, the filter constructed based on expert input.

[0033] Continuing to block 504, the server analyzes the plurality of classifications received from the various PCs to determine system usage trends among the participants of the study. Advancing to block 506, the server can use the analysis of the classifications in analytics that

can, e.g., provide input to update design requirements of PCs and PC components, improve user experience, etc.

[0034] Referring now to FIG. 6, shown is a block diagram of an example system with which embodiments can be used. As seen, system 600 may be a smartphone or other wireless communicator. A baseband processor 605 is configured to perform various signal processing with regard to communication signals to be transmitted from or received by the system. In turn, baseband processor 605 is coupled to an application processor 610, which may be a main CPU of the system to execute an OS and other system software, in addition to user applications such as many well-known social media and multimedia applications. Application processor 610 may further be configured to perform a variety of other computing operations for the device. The application processor 610 may include collection logic 614 to collect a user's browsing history, e.g., URLs visited by the user. The application processor 610 may also include classification logic 616 to classify the browsing history according to high level categories (e.g. the categories suppress identities of the URLs) using models that have been provided, according to embodiments of the present invention. The application processor 610 may provide classification data, e.g., the usage information classified according to category (e.g., suppressing the raw usage data, such as actual URLs and titles, from transmission) to a server, e.g., via RF transceiver 670, according to embodiments of the present invention. The server may store the received usage information. In an embodiment, the usage information can be combined with usage information received from other users, analyzed, and used in analytics that may influence future modification of hardware, software, operating systems, etc. to improve user experience, enhance efficiency in information retrieval, etc.

[0035] In turn, the application processor 610 can couple to a user interface/display 620, e.g., a touch screen display. In addition, application processor 610 may couple to a memory system including a non-volatile memory, namely a flash memory 630 and a system memory, namely a dynamic random access memory (DRAM) 635. As further seen, application processor 610 further couples to a capture device 640 such as one or more image capture devices that can record video and/or still images.

[0036] Still referring to FIG. 6, a universal integrated circuit card (UICC) 640 comprising a subscriber identity module and possibly a secure storage and cryptoprocessor is also coupled

to application processor 610. System 600 may further include a security processor 650 that may couple to application processor 610. A plurality of sensors 625 may couple to application processor 610 to enable input of a variety of sensed information such as accelerometer and other environmental information. An audio output device 695 may provide an interface to output sound, e.g., in the form of voice communications, played or streaming audio data and so forth.

[0037] As further illustrated, a near field communication (NFC) contactless interface 660 is provided that communicates in a NFC near field via an NFC antenna 665. While separate antennae are shown in FIG. 6, understand that in some implementations one antenna or a different set of antennae may be provided to enable various wireless functionality.

[0038] To enable communications to be transmitted and received, various circuitry may be coupled between baseband processor 605 and an antenna 690. Specifically, a radio frequency (RF) transceiver 670 and a wireless local area network (WLAN) transceiver 675 may be present. In general, RF transceiver 670 may be used to receive and transmit wireless data and calls according to a given wireless communication protocol such as 3G or 4G wireless communication protocol such as in accordance with a code division multiple access (CDMA), global system for mobile communication (GSM), long term evolution (LTE) or other protocol. In addition a GPS sensor 680 may be present. Other wireless communications such as receipt or transmission of radio signals, e.g., AM/FM and other signals may also be provided. In addition, via WLAN transceiver 675, local wireless communications can also be realized.

[0039] Additional embodiments are described below.

[0040] A first embodiment is a system that includes a processor including at least a first core that includes collection logic to record a history of website accesses of a plurality of websites by a user. The processor also includes classification logic to assign the website accesses to corresponding categories by application of a plurality of models, where each model corresponds to a respective category, and to determine a classification summary that includes a plurality of category metrics, each category metric associated with the respective category, each category metric based on a corresponding measure of the website accesses within the respective category, where the classification summary suppresses a corresponding

identity of each website accessed. The system also includes a nonvolatile memory coupled to the processor.

[0041] A 2nd embodiment includes elements of the 1st embodiment, where the nonvolatile memory is to store a representation of each of the plurality of models.

[0042] A 3rd embodiment includes elements of the 1st embodiment, where each category metric is to include a respective frequency statistic that is based on a count of the website accesses of the websites assigned to the corresponding category during a determined time period.

[0043] A 4th embodiment includes elements of the 1st embodiment. Additionally, each category metric is to include a respective temporal statistic that is based on a cumulative time duration of the website accesses of the websites assigned to the corresponding category during a determined time period.

[0044] A 5th embodiment includes elements of the 1st embodiment, where a category count of the categories is less than approximately 100.

[0045] A 6th embodiment includes elements of any one of embodiments 1 – 5, where each category corresponds to a unique set of websites and each website is to be included a single corresponding category.

[0046] A 7th embodiment is a method that includes gathering, by a server, website identification data of a plurality of websites and corresponding popularity data; determining by the server an initial set of categories based on the website identification data and the corresponding popularity data; applying a category reduction filter to the initial set of categories to exclude a subset of categories that corresponds to private information of a user that is to access websites via a user system, to produce a reduced set of categories; constructing a final set of categories from the modified set of categories according to a specified count of categories in the final set of categories; building a plurality of models, each model associated with a corresponding category of the final set of categories, each model to provide a quantitative measure of a fit of a particular website for inclusion in the corresponding category; and providing a classification tool to the user system, where the classification tool includes the plurality of models and the final set of categories, where each model is identified with its corresponding category.

[0047] An 8th embodiment includes elements of the 7th embodiment, where constructing the final set of categories includes combining two or more categories of the modified set of categories to reduce a count of distinct categories to be included in the final set of categories.

[0048] A 9th embodiment includes elements of the 7th embodiment, where building the models includes applying training data to the final set of categories using one or more machine learning techniques.

[0049] A 10th embodiment includes elements of the 9th embodiment, where each model is formed based at least in part on universal resource locators (URLs) and corresponding page titles of the training data.

[0050] An 11th embodiment includes elements of the 7th embodiment, and further includes periodically updating the classification tool by repeating gathering the website data, determining the initial set of categories, applying the category reduction filter, constructing the final set of categories, and forming the plurality of models.

[0051] A 12th embodiment includes elements of the 7th embodiment, where periodically updating the classification tool further comprises periodically updating the category reduction filter.

[0052] A 13th embodiment includes elements of the 7th embodiment, where at least some of the categories in the final set of categories pertain to system usage of the user system.

[0053] A 14th embodiment includes elements of the 7th embodiment, where the classification tool is to output a classification summary that includes a measure of website accesses for each category of the final set of categories.

[0054] A 15th embodiment includes elements of the 14th embodiment, where the classification summary is to suppress an identity of each universal resource locator (URL) of each website represented within a particular category.

[0055] A 16th embodiment includes elements of any one of the 7th to the 15th embodiments further includes constructing the category reduction filter based on expert input received from at least one expert source.

[0056] A 17th embodiment is a machine readable medium having stored thereon instructions, which if performed by a machine cause the machine to perform a method that includes receiving, by a server from each of a plurality of user systems, a respective classification summary that includes, for each category of a set of categories, a category metric that includes a frequency statistic including a measure of website accesses of websites assigned to the category during a defined time period, where the classification summary is to suppress a corresponding identity of each of the websites assigned to each category; performing an analysis of the classification summary received; and determining modifications of user system design requirements based at least in part on the analysis.

[0057] An 18th embodiment includes elements of the 17th embodiment, where at least some of the categories of the set of categories pertain to system usage of each user system from which the classification summaries are received.

[0058] A 19th embodiment includes elements of the 17th embodiment, where suppression of the corresponding identity of each of the websites assigned to each category includes prevention of determination of a corresponding universal resource locator (URL) and a corresponding page title of each of the websites reflected in the classification summary.

[0059] A 20th embodiment includes elements of any one of the 17th to the 19th embodiments, where each category metric further includes a time duration statistic determined based on a sum of time durations of access, during the defined time period, of each of the websites within the corresponding category.

[0060] A 21st embodiment is a method that includes receiving, by a server from each of a plurality of user systems, a respective classification summary that includes, for each category of a set of categories, a category metric that includes a frequency statistic including a measure of website accesses of websites assigned to the category during a defined time period, where the classification summary is to suppress a corresponding identity of each of the websites assigned to each category; performing an analysis of the classification summary received; and determining modifications of user system design requirements based at least in part on the analysis.

[0061] A 22nd embodiment includes elements of the 21st embodiment, where at least some of the categories of the set of categories pertain to system usage of each user system from which the classification summaries are received.

[0062] A 23rd embodiment includes elements of the 21st embodiment, where suppression of the corresponding identity of each of the websites assigned to each category is to prevent determination of a corresponding universal resource locator (URL) and a corresponding page title of each of the websites reflected in the classification summary.

[0063] A 24th embodiment includes elements of any one of the 21st to the 23rd embodiments, where each category metric further includes a time duration statistic determined based on a sum of time durations of access, during the defined time period, of each of the websites within the corresponding category.

[0064] A 25th embodiment is a system that includes a server including at least one processor to: receive from each of a plurality of user systems, a respective classification summary that includes, for each category of a set of categories, a category metric that includes a frequency statistic including a measure of website accesses of websites assigned to the category during a defined time period, where the classification summary is to suppress a corresponding identity of each of the websites assigned to each category; perform an analysis of the classification summary received; and recommend modifications of user system design requirements based at least in part on the analysis.

[0065] A 26th embodiment includes elements of the 25th embodiment, where at least some of the categories of the set of categories pertain to system usage of each user system from which the classification summaries are received.

[0066] A 27th embodiment includes elements of the 25th embodiment, where suppression of the corresponding identity of each of the websites assigned to each category includes to prevent determination of a corresponding universal resource locator (URL) and a corresponding page title of each of the websites reflected in the classification summary.

[0067] A 28th embodiment includes elements of any one of embodiments 25 – 27, where each category metric further includes a time duration statistic determined based on a sum of time durations of access, during the defined time period, of each of the websites within the corresponding category.

[0068] A 29th embodiment is a method that includes recording a history of website accesses of a plurality of websites by a user; assigning the website accesses to corresponding categories by application of a plurality of models, where each model corresponds to a respective category; and determining a classification summary that includes a plurality of category metrics, each category metric associated with the respective category, each category metric based on a corresponding measure of the website accesses within the respective category, where the classification summary suppresses a corresponding identity of each website accessed.

[0069] A 30th embodiment includes elements of the 29th embodiment, where each category metric is to include a respective frequency statistic that is based on a count of the website accesses of the websites assigned to the corresponding category during a determined time period.

[0070] A 31st embodiment includes elements of the 29th embodiment, where each category metric is to include a respective temporal statistic that is based on a cumulative time duration of the website accesses of the websites assigned to the corresponding category during a determined time period.

[0071] A 32nd embodiment includes elements of the 29th embodiment, where a category count of the categories is less than approximately 100.

[0072] A 33rd embodiment includes elements of any one of embodiments 29 - 32, where each category corresponds to a unique set of websites and each website is to be included a single corresponding category.

[0073] Embodiments may be used in many different types of systems. For example, in one embodiment a communication device can be arranged to perform the various methods and techniques described herein. Of course, the scope of the present invention is not limited to a communication device, and instead other embodiments can be directed to other types of apparatus for processing instructions, or one or more machine readable media including instructions that in response to being executed on a computing device, cause the device to carry out one or more of the methods and techniques described herein.

[0074] Embodiments may be implemented in code and may be stored on a non-transitory storage medium having stored thereon instructions which can be used to program a system to

perform the instructions. Embodiments also may be implemented in data and may be stored on a non-transitory storage medium, which if used by at least one machine, causes the at least one machine to fabricate at least one integrated circuit to perform one or more operations. The storage medium may include, but is not limited to, any type of disk including floppy disks, optical disks, solid state drives (SSDs), compact disk read-only memories (CD-ROMs), compact disk rewritables (CD-RWs), and magneto-optical disks, semiconductor devices such as read-only memories (ROMs), random access memories (RAMs) such as dynamic random access memories (DRAMs), static random access memories (SRAMs), erasable programmable read-only memories (EPROMs), flash memories, electrically erasable programmable read-only memories (EEPROMs), magnetic or optical cards, or any other type of media suitable for storing electronic instructions.

[0075] While the present invention has been described with respect to a limited number of embodiments, those skilled in the art will appreciate numerous modifications and variations therefrom. It is intended that the appended claims cover all such modifications and variations as fall within the true spirit and scope of this present invention.

What is claimed is:

- 1 1. A system including:
2 a processor including at least a first core that includes:
3 collection logic to record a history of website accesses of a plurality of
4 websites by a user; and
5 classification logic to assign the website accesses to corresponding categories
6 by application of a plurality of models, wherein each model corresponds to a respective
7 category, and to determine a classification summary that includes a plurality of category
8 metrics, each category metric associated with the respective category, each category metric
9 based on a corresponding measure of the website accesses within the respective category,
10 wherein the classification summary suppresses a corresponding identity of each website
11 accessed; and
12 a nonvolatile memory coupled to the processor.
- 1 2. The system of claim 1, wherein the nonvolatile memory is to store a representation of
2 each of the plurality of models.
- 1 3. The system of claim 1, wherein each category metric is to include a respective
2 frequency statistic that is based on a count of the website accesses of the websites assigned to
3 the corresponding category during a determined time period.
- 1 4. The system of claim 1, wherein each category metric is to include a respective
2 temporal statistic that is based on a cumulative time duration of the website accesses of the
3 websites assigned to the corresponding category during a determined time period.
- 1 5. The system of claim 1, wherein a category count of the categories is less than
2 approximately 100.
- 1 6. The system of any one of claims 1 - 5, wherein each category corresponds to a unique
2 set of websites and each website is to be included a single corresponding category.

- 1 7. A method comprising:
2 gathering, by a server, website identification data of a plurality of websites and
3 corresponding popularity data;
4 determining by the server an initial set of categories based on the website
5 identification data and the corresponding popularity data;
6 applying a category reduction filter to the initial set of categories to exclude a subset
7 of categories that corresponds to private information of a user that is to access websites via a
8 user system, to produce a reduced set of categories;
9 constructing a final set of categories from the modified set of categories according to
10 a specified count of categories in the final set of categories;
11 building a plurality of models, each model associated with a corresponding category
12 of the final set of categories, each model to provide a quantitative measure of a fit of a
13 particular website for inclusion in the corresponding category; and
14 providing a classification tool to the user system, wherein the classification tool
15 includes the plurality of models and the final set of categories, wherein each model is
16 identified with its corresponding category.
- 1 8. The method of claim 7, wherein constructing the final set of categories includes
2 combining two or more categories of the modified set of categories to reduce a count of
3 distinct categories to be included in the final set of categories.
- 1 9. The method of claim 7, wherein building the models includes applying training data
2 to the final set of categories using one or more machine learning techniques.
- 1 10. The method of claim 9, wherein each model is formed based at least in part on
2 universal resource locators (URLs) and corresponding page titles of the training data.
- 1 11. The method of claim 7, further comprising periodically updating the classification
2 tool by repeating gathering the website data, determining the initial set of categories,
3 applying the category reduction filter, constructing the final set of categories, and forming the
4 plurality of models.

12. The method of claim 7, wherein periodically updating the classification tool further comprises periodically updating the category reduction filter.

13. The method of claim 7, wherein at least some of the categories in the final set of categories pertain to system usage of the user system.

14. The method of claim 7, wherein the classification tool is to output a classification summary that includes a measure of website accesses by the user system for each category of the final set of categories.

15. The method of claim 14, wherein the classification summary is to suppress an identity of each universal resource locator (URL) of each website represented within a particular category.

16. The method of any one of claims 7 to 15, further comprising constructing the category reduction filter based on expert input received from at least one expert source.

17. A machine readable medium having stored thereon instructions, which if performed by a machine cause the machine to perform a method comprising:
receiving, by a server from each of a plurality of user systems, a respective classification summary that includes, for each category of a set of categories, a category metric that includes a frequency statistic including a measure of website accesses of websites assigned to the category during a defined time period, wherein the classification summary is to suppress a corresponding identity of each of the websites assigned to each category;
performing an analysis of the classification summary received; and
determining modifications of user system design requirements based at least in part on the analysis.

18. The computer readable medium of claim 17, wherein at least some of the categories of the set of categories pertain to system usage of each user system from which the classification summaries are received.

1 19. The computer readable medium of claim 17, wherein suppression of the
2 corresponding identity of each of the websites assigned to each category includes preventing
3 determination of a corresponding universal resource locator (URL) and a corresponding page
4 title of each of the websites reflected in the classification summary.

1 20. The computer readable medium of any one of claims 17 -19, wherein each category
2 metric further includes a time duration statistic determined based on a sum of time durations
3 of access, during the defined time period, of each of the websites within the corresponding
4 category.

1 21. A method comprising:
2 recording a history of website accesses of a plurality of websites by a user;
3 assigning the website accesses to corresponding categories by application of a
4 plurality of models, wherein each model corresponds to a respective category;
5 determining a classification summary that includes a plurality of category metrics,
6 each category metric associated with the respective category, each category metric based on a
7 corresponding measure of the website accesses within the respective category, wherein the
8 classification summary suppresses a corresponding identity of each website accessed.

1 22. The method of claim 21, wherein each category metric is to include a respective
2 frequency statistic that is based on a count of the website accesses of the websites assigned to
3 the corresponding category during a determined time period.

1 23. The method of claim 21, wherein each category metric is to include a respective
2 temporal statistic that is based on a cumulative time duration of the website accesses of the
3 websites assigned to the corresponding category during a determined time period.

1 24. The method of claim 21, wherein a category count of the categories is less than
2 approximately 100.

1 25. The method of any one of claims 21 - 24, wherein each category corresponds to a
2 unique set of websites and each website is to be included a single corresponding category.

100

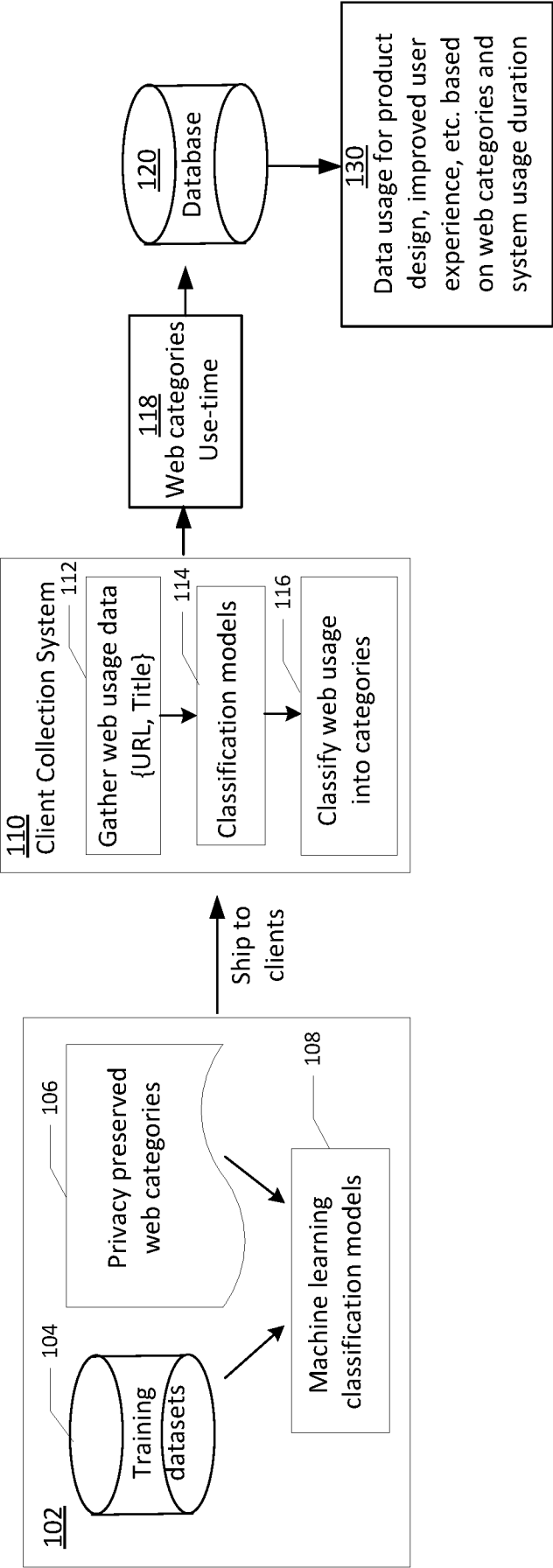


FIG. 1

2/6

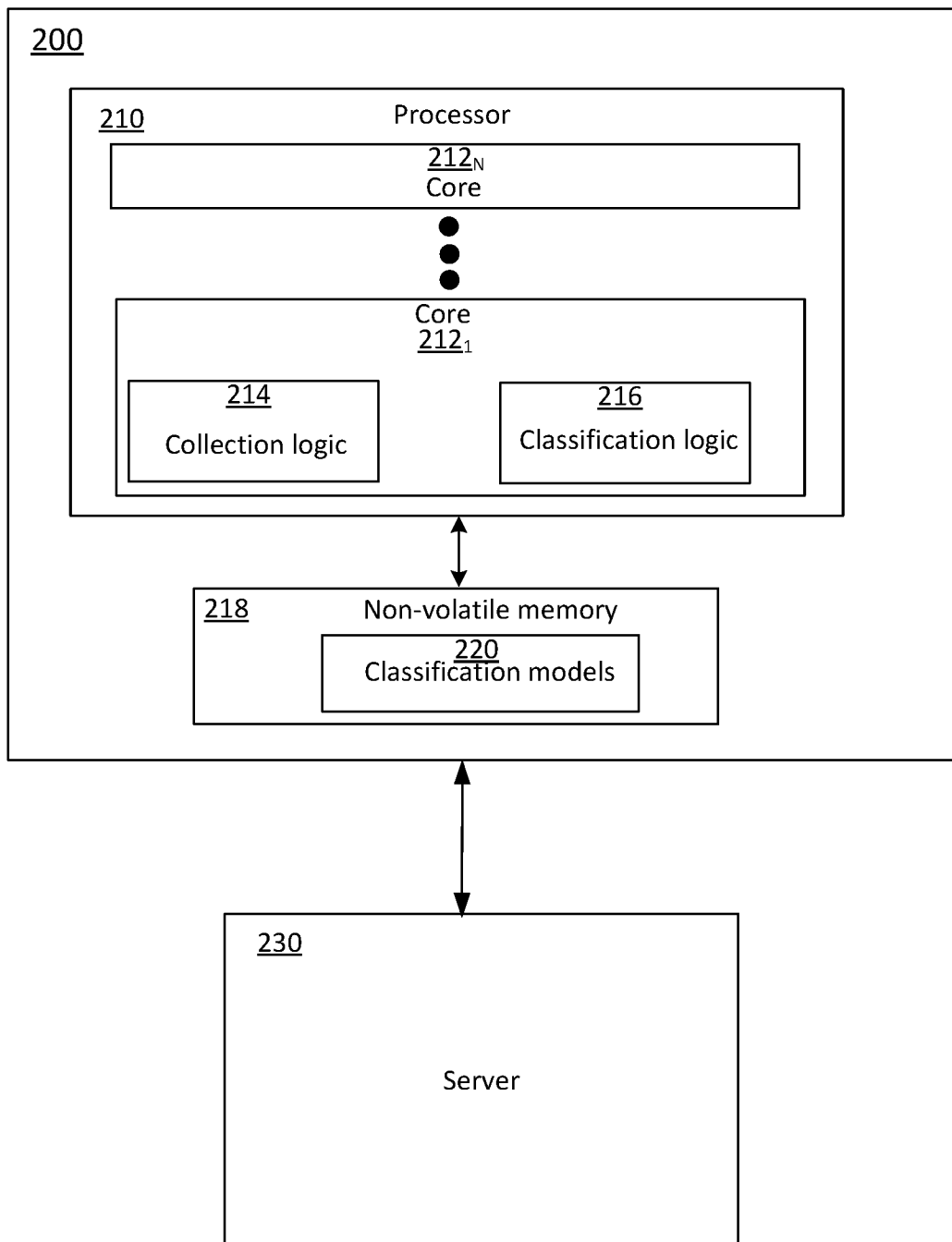


FIG. 2

3/6

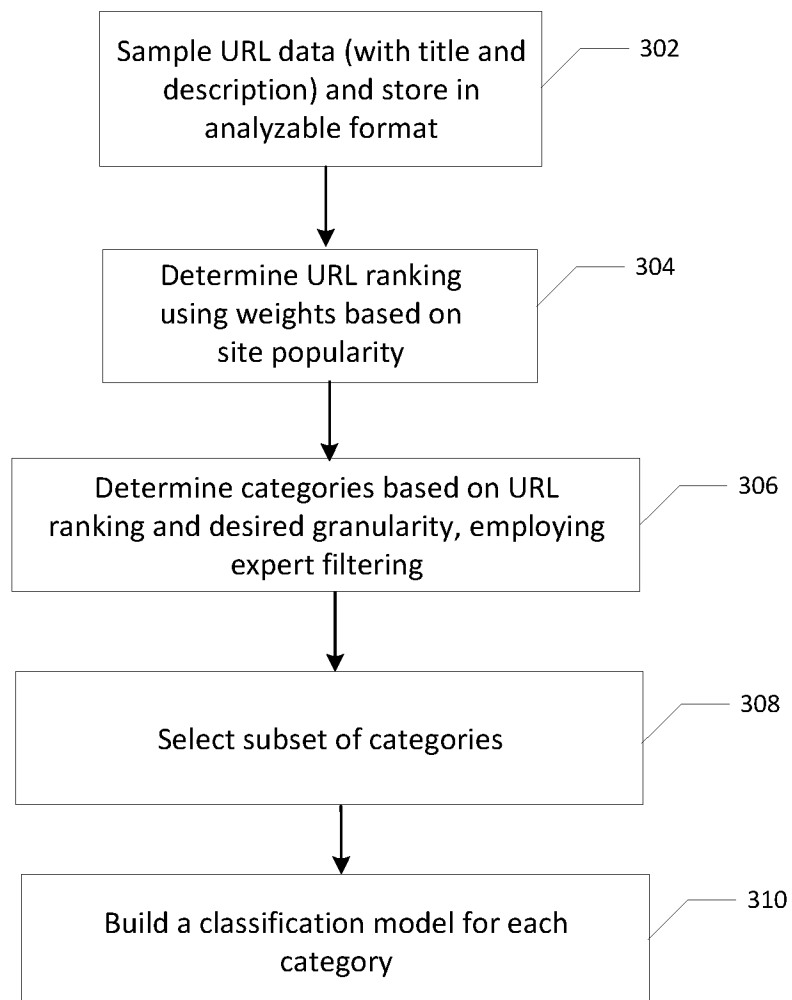
300

FIG. 3

4/6

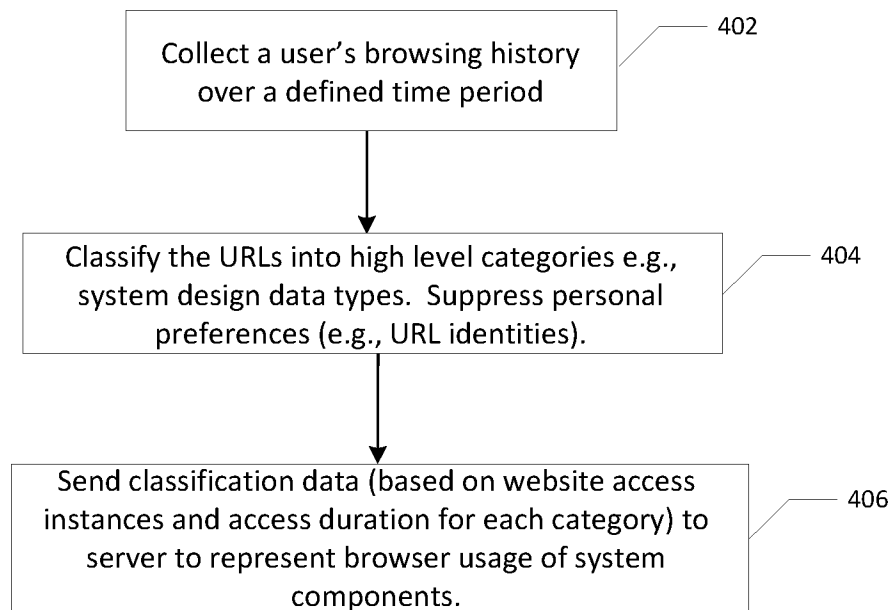
400

FIG. 4

5/6

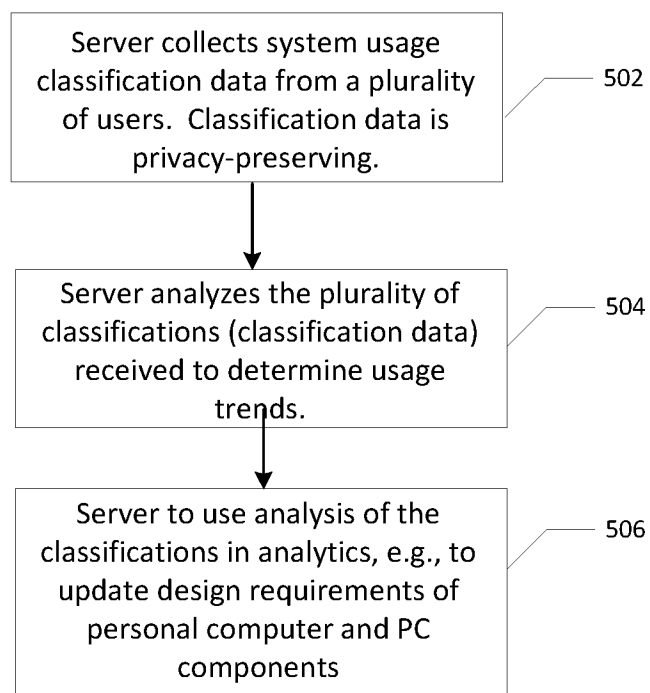
500

FIG. 5

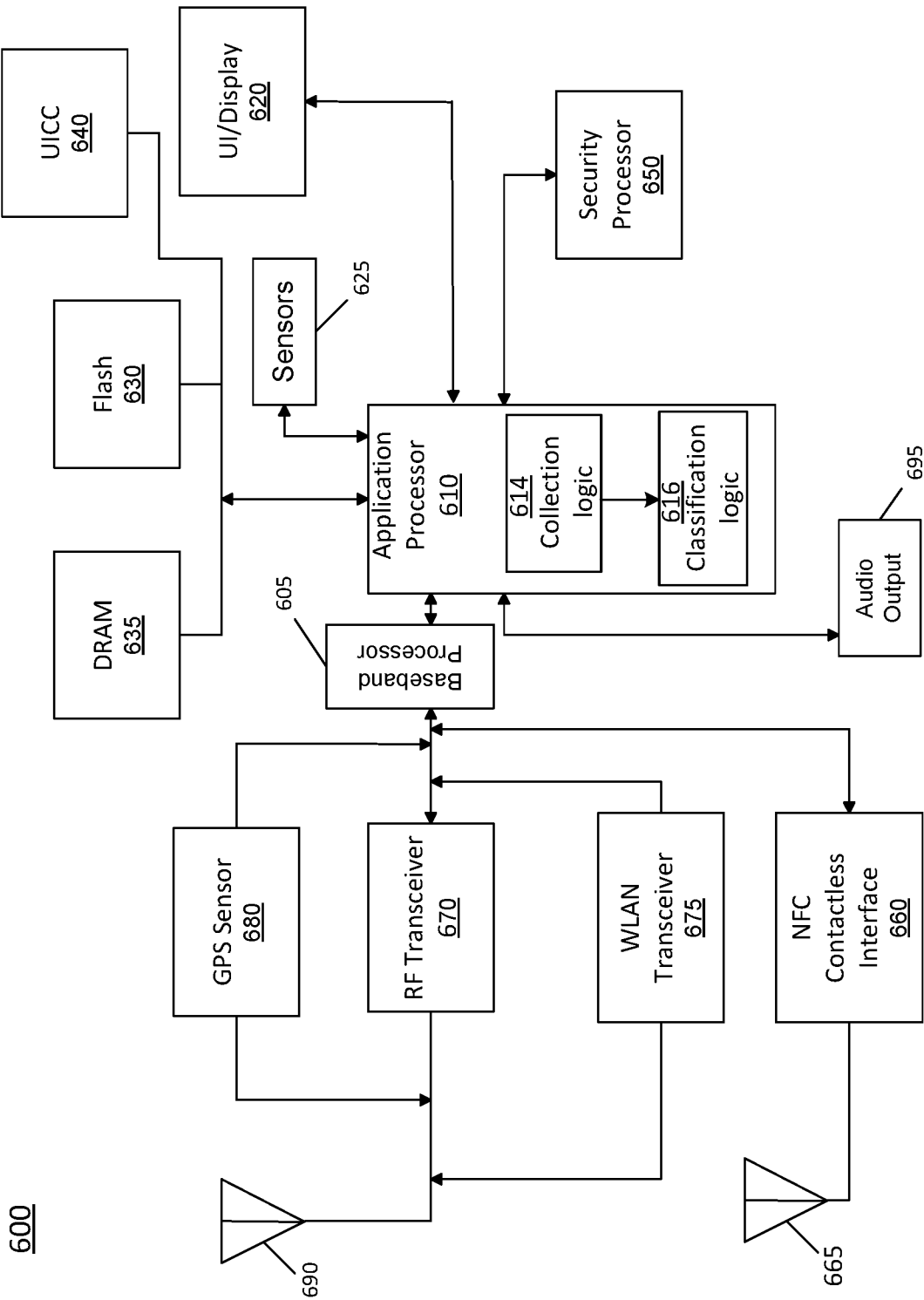


FIG. 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2016/048552**A. CLASSIFICATION OF SUBJECT MATTER****G06F 17/30(2006.01)i**

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHEDMinimum documentation searched (classification system followed by classification symbols)
G06F 17/30; G06F 15/16; G06F 21/24; H04N 5/76; G06F 21/00Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Korean utility models and applications for utility models
Japanese utility models and applications for utility modelsElectronic data base consulted during the international search (name of data base and, where practicable, search terms used)
eKOMPASS(KIPO internal) & Keywords: website access history, website categories, category models, classification summary, category metrics, suppressed website identity, machine learning, category reduction filter**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2006-0059225 A1 (TIMOTHY P. STONEHOCKER et al.) 16 March 2006 See paragraphs [0003], [0024], [0045]-[0046], [0050]-[0055], [0059]-[0060], [0064]-[0065], [0075], [0081], [0100], [0106], [0111]-[0112], and [0120]; and figures 5B, 6B, and 7.	1-25
Y	US 2011-0137904 A1 (SHYAM SUNDAR RAJARAM et al.) 09 June 2011 See paragraphs [0010], [0020], [0027], [0030], [0044], [0047], [0049], [0052]-[0053], [0058], [0060], and [0062]; and figures 2 and 4.	1-25
A	JP 2006-323546 A (MATSUSHITA ELECTRIC IND. CO., LTD.) 30 November 2006 See paragraphs [0003], [0025], [0053], [0055], [0070], and [0079]; and figure 4.	1-25
A	US 2012-0297017 A1 (BENJAMIN LIVSHITS et al.) 22 November 2012 See paragraphs [0020], [0023]-[0024], [0026], [0028], and [0068].	1-25
A	US 2014-0059658 A1 (MARTIN STECHER) 27 February 2014 See paragraphs [0001], [0030], [0038]-[0039], and [0043].	1-25

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

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

06 December 2016 (06.12.2016)

Date of mailing of the international search report

07 December 2016 (07.12.2016)

Name and mailing address of the ISA/KR

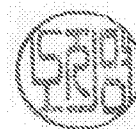
International Application Division
Korean Intellectual Property Office
189 Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

NHO, Ji Myong

Telephone No. +82-42-481-8528



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2016/048552

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2006-0059225 A1	16/03/2006	CA 2579312 A1 CA 2579312 C CN 101432714 A CN 101432714 B JP 2008-513887 A JP 4782790 B2 WO 2006-031864 A2 WO 2006-031864 A3	23/03/2006 25/06/2013 13/05/2009 16/10/2013 01/05/2008 28/09/2011 23/03/2006 16/04/2009
US 2011-0137904 A1	09/06/2011	US 9256692 B2	09/02/2016
JP 2006-323546 A	30/11/2006	None	
US 2012-0297017 A1	22/11/2012	None	
US 2014-0059658 A1	27/02/2014	US 9268933 B2	23/02/2016