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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
4 September 2008 (04.09.2008)

PCT

(10) International Publication Number
WO 2008/106668 A1

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

(21) International Application Number: PCT/US2008/055581

(22) International Filing Date: 1 March 2008 (01.03.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/892,497	1 March 2007 (01.03.2007)	US
11/849,136	31 August 2007 (31.08.2007)	US

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

(72) Inventors: LI, Hua; One Microsoft Way, Redmond, WA 98052-6399 (US). **ZENG, Hua-jun**; One Microsoft Way, Redmond, WA 98052-6399 (US). **HU, Jian**; One Microsoft Way, Redmond, WA 98052-6399 (US). **CHEN, Zheng**; One Microsoft Way, Redmond, WA 98052-6399 (US). **WANG, Jian**; One Microsoft Way, Redmond, WA 98052-6399 (US).

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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),

[Continued on next page]

(54) Title: USER QUERY MINING FOR ADVERTISING MATCHING

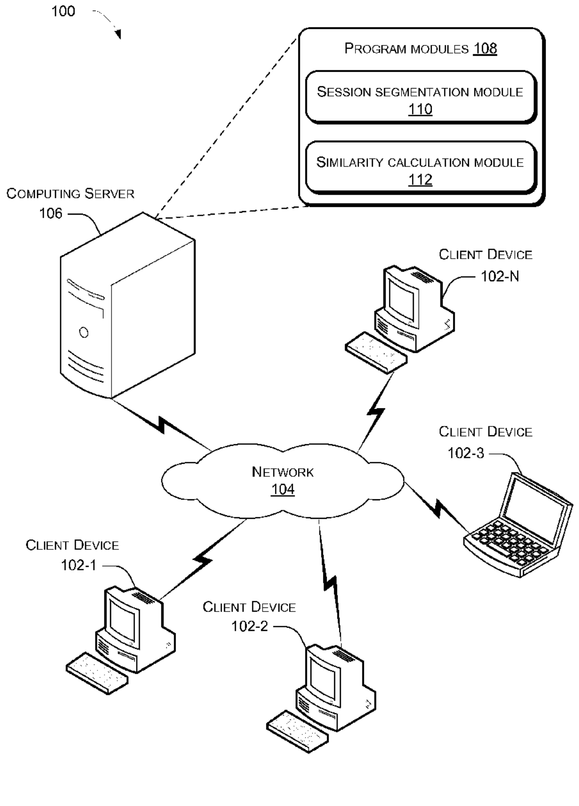

FIG. 1

(57) Abstract: Systems and methods to determine relevant keywords from a user's search query sessions are disclosed. The described method includes identifying search session logs of a user, segmenting the search session logs into one or more search sessions. After the segmentation, the search sessions are analyzed to compose a list of semantically relevant keyword sets including at least a first keyword set and a second keyword set. The described method further includes determining a semantic relevance between the first and second keyword sets according to the frequency at which the first and second keyword sets are reported in the query results and displaying one or more semantically high relevant keyword sets after being filtered by a threshold.

European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

— *with international search report*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

USER QUERY MINING FOR ADVERTISING MATCHING

BACKGROUND

[0001] A global computer network, such as the Internet, provides an opportunity for advertisers to target Internet users within the network for marketing purposes. A user commonly searches for and browses content on the network by entering one or more search terms in a search query, typically implemented in search websites. The content that the user searches can include articles, web pages, emails, or other Internet accessible content containing the search term provided by the user. Many of the advertisements that the user is subjected to are based on search terms entered into a search engine.

[0002] The advertisements are generated by one or more agents, or advertisement engines, operating on web-servers for generating and displaying advertisements to users. Many search engines operate in cooperation with an advertisement engine, to display advertisements (ads) in response to the search terms entered by the user in the search query. The advertisement engine has a database of ads, from which the ads are selected based upon relevance to the search term provided by the user. In the case of paid search auctions, an advertiser bids on one or more search terms that relate to the ads. The paid search auction returns ads of the advertiser that are relevant to user queries according to the bidding search terms. For this, an advertiser needs to accurately match a user’s query to the advertiser’s search terms based on the relevancy of the search terms. Since each user can input widely varying search terms when searching for similar or identical search content, the matching of the user's query to the advertiser's search terms is difficult to achieve.

SUMMARY

[0003] Systems and methods to determine relevant keyword set from a user's search query sessions are disclosed. In one aspect, the methods include parsing a search session log of a user and segmenting the search session log into a search session having a first keyword set and a second keyword set. The methods further include determining a semantic relevance between the first and the second keyword sets according to the frequency at which both the first and second keyword sets are reported in the search results and displaying one or more semantically relevant keyword sets to the user based on the relevance.

[0004] This summary is provided to introduce simplified concepts of user query session mining for matching advertisements, which is further described below in the Detailed Description. This summary is not intended to identify essential features of the claimed subject matter, nor is it intended for use in determining the scope of the claimed subject matter.

BRIEF DESCRIPTION OF THE DRAWINGS

[0005] The detailed description is set forth with reference to the accompanying figures. In the figures, the left-most digit(s) of a reference number identifies the figure in which the reference number first appears. The use of the same reference numbers in different figures indicates similar or identical items.

[0006] FIG. 1 illustrates an exemplary computing environment.

[0007] FIG. 2 illustrates an exemplary computing device for determining keyword set relevance.

[0008] FIG. 3 illustrates an exemplary method for determining keyword set relevance.

[0009] FIG. 4 illustrates an exemplary method for performing similarity analysis.

[0010] FIG. 5 illustrates an exemplary method for calculating a similarity threshold.

DETAILED DESCRIPTION

[0011] This disclosure relates to methods and systems for determining one or more relevant keyword sets in a query provided by a user. The user can search for and browse content on a global computer network, like the Internet, by entering keywords using a search engine. Many search engines work in cooperation with an advertisement agency which further implements advertisements on the Internet through an advertisement engine. The advertisements are generated and displayed to one or more users based on an association with a plurality of keywords, collectively referred to as keyword sets. Based on such an association, the advertisement engine displays advertisements (ads) related to one or more keyword sets provided as input in the query. Moreover, the search engines can create and maintain such keyword sets that are correlated to specific searches or search results. These keyword sets, interchangeably referred to as the advertiser's keyword set, can be used commercially by offering the keyword sets to the advertisers for some consideration. Once such keyword sets are reserved by advertisers; any user entering a search query would generate search results that are associated with the keyword set and in turn the generated result. The advertisers can have ads related to the one or more keyword sets. However, matching the user’s query to the advertiser’s keyword sets can be difficult to achieve.

[0012] The methods and systems disclosed in the following description relate to determining semantically relevant keyword sets from the user’s query to match the advertiser’s keyword sets. The methods include identifying search session logs of the user, segmenting the identified search session logs into one or more search sessions based on a predetermined session time interval. The methods further include determining one or more semantically related keyword sets based on the rate at which the keyword sets occur in the user queries during each search session. In addition, a similarity threshold can be computed to filter the one or more keyword sets with high semantic relevance to match the advertiser's keyword sets.

Exemplary Computing Environment

[0013] Fig.1 illustrates an exemplary computing environment 100 which can be used to implement the techniques described herein, and which may be representative, in whole or in part, of the elements described herein. The computing environment 100 is an example of a computing environment and is not intended to suggest any limitation to the scope of use and/or functionality of the computer and network architectures.

[0014] The computing environment 100 can include a plurality of client devices 102-1, 102-2, 102-3,… 102-n (collectively referred to as client devices 102) communicatively connected through a network 104 with a computing server 106. The client devices 102 may be implemented as any of a variety of conventional computing devices, including, for example, a server, a desktop PC, a notebook or a portable computer, a workstation, a mainframe computer, a mobile computing device, an entertainment device, or an Internet appliance, etc.

[0015] The network 104 may be a wireless or a wired network, or a combination thereof. The network 104 may also be a collection of individual networks, interconnected with each other and functioning as a single large network (e.g., the Internet or an intranet). Examples of such individual networks include, but are not limited to, Local Area Networks (LANs), Wide

Area Networks (WANs), and Metropolitan Area Networks (MANs). Further, the individual networks may be wireless or wired networks, or a combination thereof.

[0016] The computing server 106 can be, for example, a general purpose computing device, a server, a cluster of servers, mainframes, etc. The computing server 106 can be configured to receive one or more queries from the client devices 102. The computing server 106 can be further configured to determine semantically relevant keyword sets from the queries received from the client devices 102. The semantically relevant keyword sets are then displayed to the client devices 102 by the computing server 106.

[0017] In one implementation, the computing server 106 can be configured to execute a session segmentation module 110 and a similarity calculation module 112 to determine semantically relevant keyword sets in the user queries received from any one or more of the client devices 102. Each keyword set can include a single or multiple keywords, a compound term, a hyphenated term, a phrase, or an entire query. Each query can be a single search term or a list of search terms that may or may not have connectors like AND, OR, etc. between the search terms.

[0018] The session segmentation module 110 may be configured to identify a user and one or more search session logs of the user. The search session logs may include a collection of one or more records or information related to search queries during a time interval between the user logging in and logging out of the client device 102. In one implementation, the session segmentation module 110 identifies the user using a globally unique identifier (GUID) or a universally unique identifier (UUID). The session segmentation module 110 further segments the identified search session logs into one or more search sessions. Each of the search sessions can include a series of related message exchanges or exchange of related search queries and respective search results between a computing server and a client device within a predetermined session time interval. The set of user queries occurring within the predetermined session time interval may be segmented as a search session with queries having a semantic relationship with each other. Thereby the session segmentation module 110 determines one or more semantically related keyword sets of the search queries entered during the search session by the user.

[0019] The similarity calculation module 112 can be configured to filter one or more semantically high relevant keyword sets from one or more semantically related keyword sets. The filtration can be performed using a similarity threshold 220 as a filter. In one embodiment, the similarity threshold can be computed by a threshold calculation module. The similarity calculation module 112 can be further configured to filter the semantically high relevant keyword sets after evaluating the similarity threshold against a similarity value. In one embodiment, the similarity calculation module 110 determines the similarity value between the semantically related keyword sets. In one implementation, the similarity value can be computed by either a mutual information analysis or a cosine similarity analysis. The computed similarity value is then compared with the similarity threshold and the keyword sets having similarity value greater than the similarity threshold can be identified as semantically high relevant keyword sets by the similarity calculation module 112.

Exemplary Computing Server

[0020] FIG. 2 is an exemplary computing server 106 for determining keyword set relevance. The computing server 106 includes a processor 202, network interfaces 204, input/output interfaces 206, a memory 208 and input/output devices 212. The input/output interfaces 206 can include, for example, a scanner port, a mouse port, a keyboard port, etc to receive the user queries from the client devices 102. Input/output interfaces 206 can receive data such as, for example, user session log from input/output devices 212. The computing server 106 can be associated with an input/output device 212 either directly or indirectly in a network.

[0021] The memory 208 may be any computer-readable media in the form of volatile memory, such as Random Access Memory (RAM) and/or non-volatile memory, such as Read Only Memory (ROM) or flash RAM. The memory 208 typically includes data and/or program modules for determining keyword set relevance, the data, and modules being immediately accessible to and/or presently operated on by the processor 202. In one implementation, the memory 208 includes program modules 108 and program data 210.

[0022] The program modules 108 may include the session segmentation module 110, the similarity calculation module 112, the threshold calculation module 214, and other modules 216. The program data 210 can store parameters including user session log 218, threshold 220 and other program data 222.

[0023] The session segmentation module 110 identifies the user and the user's search session logs 218. In one implementation, the identification of the user can be accomplished by using either GUID or UUID, where GUID or UUID are stored as other program data 222 in the system memory 208. The session segmentation module 110 further sorts the identified search session logs to align data derived from at least one raw session log with each user. In one implementation, the session segmentation module 110 sorts the identified search session logs using an external sort process. In another implementation, the steps for identifying and sorting user's search session logs can be performed manually based on one or more options presentable to a system administrator or to other entities.

[0024] The session segmentation module 110 segments the sorted search session logs into one or more search sessions. The session segmentation module 110 analyzes each of the search sessions to compile one or more semantically related keyword sets for a given query or a series of the user queries. The search session logs can be segmented into one or more search sessions based on a predetermined session time interval. The segmented search sessions would be such that each session corresponds to a definite interval of time.

[0025] The segmented search session includes a list of similar keyword sets or semantically relevant keyword sets. The keyword sets can be such that each includes a plurality of keywords. The search session can be then analyzed to determine similar keyword sets or keyword sets that are at least related to each other. The determination can be performed using one or more keyword similarity calculation methods.

[0026] The similarity calculation module 112 determines similarity between one or more keyword sets based on frequency of occurrence of the keyword sets. In one implementation, the similarity calculation module 112 determines the similarity of the keyword sets using a similarity calculation method. The similarity calculation method aims at determining whether the keyword sets are semantically relevant or not.

[0027] In such a case, a search session having two queries can be considered as a single keyword set. Each of the two queries can be associated with a keyword set. Both keyword sets, say a first and a second keyword set, can be designated as *u* and *v* respectively. The keyword sets *u* and *v* can be combined together to form a keyword pair *uv.* It would be appreciated that the keyword pair *uv* can be represented as any combination of the first keyword set *u* and the second keyword set *v*, irrespective of the order of the keyword sets. In one implementation, keyword sets *u* and *v* can be combined together to form a keyword pair *vu*.

[0028] In one implementation, the similarity calculation method is based on frequency of occurrence of the keyword sets occurring as a keyword pair. The frequency of occurrence can be represented as f. Hence for the frequency of occurrence of the keyword pair *uv,* the frequency of occurrence for the keyword pair can be denoted as f_{uv}. In another implementation, the frequency of occurrence number f_{uv} of the keyword pair is limited by a threshold value. In one implementation, the threshold value can be f_m, where f_m is the minimum of the frequency occurrence of the first keyword set, say f_u, and frequency occurrence of the second keyword set, say f_v. In a preferred implementation, the semantically relevant keyword sets can be determined based on the relationship between the frequency occurrence number of keyword pair f_{uv} and the minimum frequency occurrence number f_m, represented as:

$$f_{uv} > \sqrt{f_m}$$

(1)

[0029] The respective frequency occurrences are computed and semantically relevant keyword sets are determined based on the above relation (1). If the above relationship is satisfied, the keyword sets are determined as semantically relevant keyword sets. Conversely, if the above relationship is not satisfied, then the keyword sets are determined as semantically non relevant keyword sets.

[0030] In another implementation, the similarity calculation module 112 determines semantically relevant keyword sets based on the relationship between the frequency occurrence number of keyword pair f_{uv} and the minimum frequency occurrence number f_m, represented as:

$$\sqrt{f_m} >= f_{uv} > \sqrt[4]{f_m}$$

(2)

If the above relationship is satisfied, the keyword sets are determined o be semantically relevant. Additionally, the degree of relevance is measured in terms of a similarity value determined through mutual information analysis. Conversely, if the above relationship is not satisfied then the keyword sets may or may not be semantically relevant.

[0031] In one implementation, the calculation of the similarity value between the keyword sets by the mutual information analysis is derivable through the following equation (3):

$$MI(q_u, q_v) = p(q_u, q_v) * \log \frac{p(q_u, q_v)}{p(q_u) * p(q_v)}$$

(3)

[0032] In the above equation (3), $p(q_u, q_v) = \frac{C_{u,v}}{N}, p(q_u) = \frac{C_u}{N}$, $p(q_v) = \frac{C_v}{N}$ for the keyword sets *u* and *v* and N being the total number of query sessions. *C* indicates number of queries present including a keyword set. For example, C_u C_v and C_{uv} would indicate the number of queries that include the keyword set *u, v* and keyword pair *uv* respectively, occurring in all sessions.

[0033] The relation between frequency of occurrence of the keyword sets can be depicted in various forms. In yet another implementation, the similarity calculation module 112 determine the semantically relevant keyword sets based on the relationship between the frequency occurrence number of keyword pair f_{uv} and the minimum frequency occurrence number f_m, represented as:

$$f_{uv} <= \sqrt[4]{f_m}, \qquad (4)$$

On one hand, if the above relationship exists, the keyword sets are determined to be semantically relevant and the degree of relevance is measured in terms of the similarity value by the cosine similarity analysis. On the other hand, if the above relationship does not exist, the keyword sets are determined as semantically non relevant keyword sets.

[0034] In yet another implementation, the similarity value between the keyword sets can be calculated by the cosine similarity analysis, based on the following equation (5):

$$Cos(q_u, q_v) = \frac{\sum_{\forall qj} C_{u,j} \cdot C_{v,j}}{\sqrt{\sum_{\forall qj} C_{u,j}^2} \cdot \sqrt{\sum_{\forall qj} C_{v,j}^2}}$$

(5)

wherein the different variable possess the same meaning as indicated previously.

[0035] As indicated previously, highly relevant keyword set can be evaluated on the basis of the obtained semantically relevant keyword sets. The highly relevant keyword sets can be evaluated by comparing the similarity value for each of the search queries with a threshold value. On the basis of the comparison the respective keyword sets can be classified as highly relevant keyword sets or not.

[0036] In one implementation, the similarity calculation module 112 can further filter the semantically relevant keyword sets to obtain one or more semantically high relevant keyword sets (interchangeably called as the advertiser's keyword sets). The semantically relevant keyword sets are filtered against the similarity threshold referred to as threshold 220. The similarity calculation module 112 filters the semantically relevant keyword sets by comparing the similarity value associated with the semantically relevant keyword sets with the threshold 220. In one implementation, the similarity calculation module 112 can be determined by comparing the similarity value associated with the keyword sets. In cases where the similarity value exceeds the threshold 220, the similarity calculation module 112 classifies the associated keyword sets as semantically highly relevant keyword sets. The semantically highly relevant keyword sets are considered to be filtered only when the similarity value of the semantically relevant keyword sets exceeds the threshold 220.

[0037] In one implementation, the threshold 220 can be determined by the threshold calculation module 214. The threshold 220 can be a minimal value derivable through one or more of the expressions described above. Examples of such threshold value include, but are not limited to, a mutual information threshold or a cosine similarity threshold, and the like. The threshold calculation module 214 identifies one or more search session logs referred to as training search session logs of the user.

[0038] In one implementation, the training search session logs can be identified after the user is identified either by UUID or GUID. The identified training search session logs of the user are segmented into one or more training search sessions based on the predetermined time interval. In particular, the training search session logs are segmented into one or more training search sessions based on the predetermined time interval known as training session time interval, with the expectation that the queries occurring within the training session time interval are semantically relevant queries.

[0039] The threshold calculation module 214 further analyzes the segmented training search sessions to identify one or more semantically related training set queries. In one implementation, the training set queries can be a first predetermined number of commonly occurring queries, say m number of queries, extracted from the query log for, for example corresponding to a definite period like one month of query log data associated with the user. For each and every query in the training set, a second predetermined number of suggested queries, say n, relevant to the query are generated. The relevance information obtained can then be used to construct a candidate suggestion query set. The number of the candidate suggestion query in the set may be with m*n number of queries present in the candidate suggestion query set. In one implementation, the suggested queries relevant to the query can be determined either by the mutual information analysis or the cosine similarity analysis.

[0040] The threshold calculation module 214 classifies the relevant suggested queries into one or more groups based on the relevance level of the suggested queries. In one implementation, the grouping process can be either automated or performed manually. The groups are then labeled based on the level of relevance. In one implementation, the groups can be labeled as relevant, quite relevant, and irrelevant. In other implementations, the groups can be labeled in a different manner. After labeling, the threshold calculation module 214 determines the threshold 220 for each group. Toward this end, the similarity calculation module 112 display the semantically high relevant keyword sets filtered by the threshold 220 for further processing.

Exemplary Processes

[0041] Exemplary methods of determining keyword set relevance to match the advertiser's keywords are described below. Some or all of these methods may, but need not, be implemented at least partially in an architecture such as that shown in FIGS. 1 & 2. Also, it should be understood that certain acts in the methods need not be performed in the order described, may be modified, and/or may be omitted entirely, depending on the circumstances.

[0042] FIG. 3 illustrates an exemplary method 300 for determining the keyword sets that have high semantic relevance. The order in which the exemplary method 300 is described is not intended to be construed as a limitation, and any number of the described method blocks can be combined in any order to implement the method, or an alternate method. Additionally, individual blocks may be deleted from the method without departing from the spirit and scope of the subject matter described herein. Furthermore, the method can be implemented in any suitable hardware, software, firmware, or combination thereof.

[0043] At block 302, a user and respective search session logs of the user are identified. In one implementation, the session segmentation module 110 identifies the search session logs of the user after identifying the user. In one implementation, the user can be identified either by GUID or by UUID. Then, the similarity calculation module 112 can sort the search session logs to align data from at least one raw session log of the user.

[0044] At block 304, the search session logs are segmented into one or more search sessions. The segmentation of the search session logs can be based on numerous factors, like sessions that are active for a definite time period. For example, the session segmentation module 110 segments the sorted search session logs of the user into one or more search sessions based on the predetermined session time interval. In one implementation, the segmentation is performed based on the predetermined session time interval as the queries derived within the predetermined session time interval are estimated to be semantically relevant.

[0045] At block 306, one or more keyword sets from search queries in a search session are identified. In one implementation, the similarity calculation module 112 identifies a first and a second keyword set from the series of queries in a search session. The search session is analyzed to identify the list of semantically relevant keyword set. The list of semantically relevant keyword sets can include at least a first keyword set and a second keyword set. In one implementation, the first and the second keyword sets are stored as the other program data 222 in the memory 208. The first keyword set can be designated as u and the second keyword set can be designated as v and the keyword pair that comprises of both the first and second keyword sets can be designated as uv. The keyword set pair uv can be any combination of u and v regardless of the order of the keyword sets.

[0046] At block 308, the frequency occurrence of the first and the second keyword sets and the keyword set pair is determined. In one implementation, similarity calculation module 112 determines the frequency occurrence of the identified keyword sets and the keyword set pair. The frequency occurrence of the first keyword set u and the second keyword set v can be determined and represented as f_u and f_v respectively. In a similar fashion, the similarity calculation module 112 determines the frequency of occurrence of the keyword set pair uv represented as $f_{uv.}$

[0047] At block 310, a minimum value between the frequency occurrence of the first and the second keyword sets is determined. In one implementation, the similarity calculation module 112 determines minimum frequency occurrence number between f_u and $f_{v.}$ The minimum frequency occurrence number denoted as f_m is determined by the following mathematical expression (6):

$$f_m = min\ (f_u, f_v)$$

(6)

[0048] At block 312, the frequency occurrence of the keyword pair is compared with the value proportional to the minimum frequency occurrence number. The comparison forms a basis for determining semantically relevant keywords corresponding to a keyword set. In one implementation, the similarity calculation module 112 determines whether the keyword sets *u* and *v* are semantically relevant or not, based on the following relationship represented as :

$$f_{uv} > \sqrt{}f_{min}$$

(7)

The determination of the above relationship is made at block 312, i.e., if the "YES" path is traced to block 314, then the keyword set *u* and *v* are determined as semantically relevant and if the "NO" path is traced to block 316, then the keyword sets *u* and *v* may or may not be semantically relevant.

[0049] At block 314, the keyword sets under consideration are displayed to be relevant. For example, the similarity calculation module 112 displays the first and the second keyword sets as semantically relevant keyword sets when the "YES" path is traced from the block 312.

[0050] At block 316, a similarity analysis is performed if the relation as specified in block 312 is not satisfied. In one implementation, the similarity calculation module 112 performs similarity analysis to determine the semantically relevant keyword sets. In one implementation, the similarity analysis performed by the similarity calculation module 112 can be either the mutual information analysis or the cosine similarity analysis to determine whether the keyword sets are semantically relevant or not.

[0051] FIG. 4 illustrates an exemplary method 400 for performing a similarity analysis to determine one or more semantically relevant keyword sets.

[0052] At block 402, values corresponding to minimum frequency occurrence number and the frequency occurrence of the keyword set pair are received. In one implementation, the similarity calculation module 112 receives the frequency occurrence number f_{uv} of the keyword pair uv and the minimum frequency occurrence number f_m.

[0053] At block 404, a determination is made to ascertain whether the given keyword sets are semantically relevant or not on the basis of the minimum frequency occurrence number and the frequency occurrence of the keyword pair. In one implementation, the similarity calculation module 112 determines whether the keyword sets *u* and *v* are semantically relevant or not, based on the following relationship represented as:

$$\sqrt{f_{m}} >= f_{uv} > {}^{4}\sqrt{f_{m}}$$

(8)

If the above relation is satisfied, the "YES" path is traced to block 406 indicating the keyword sets *u* and *v* as semantically relevant. In another implementation, the degree of relevance can be measured in terms of a similarity value indicating the similarity between the keyword sets. The similarity value as indicated can be determined by the mutual information analysis. Conversely, if the "NO" path is traced to block 408, then the keyword sets *u* and *v* may or may not be semantically relevant.

[0054] At block 406, the similarity value is calculated based on number of queries including the search term. In one implementation, the similarity calculation module 112 computes the similarity value between the keyword sets *u* and *v* by the mutual information analysis. The mutual information analysis can be represented by the following equation (9):

$$MI(q_u, q_v) = p(q_u, q_v) * \log \frac{p(q_u, q_v)}{p(q_u) * p(q_v)}$$

(9)

In the above equation (9), $p(q_u, q_v) = \frac{C_{u,v}}{N}, p(q_u) = \frac{C_u}{N}, p(q_v) = \frac{C_v}{N}$, for the keyword sets *u* and *v* and N being the total number of query sessions. *C* indicates number of queries including a keyword set. For example, C_u, C_v and C_{uv} would indicate the number of queries that include the keyword set *u, v* and keyword pair *uv* respectively, occurring in all sessions.

[0055] At block 408, the similarity calculation module 112 determines whether the keyword sets *u* and *v* are semantically relevant or not, based on the following relationship represented as:

$$f_{uv} <= {}^{4}\sqrt{f_{\mathrm{m}}}$$

(10)

The determination of the above relationship is made at block 408,i.e., if the "YES" path is traced to block 410, the keyword sets *u* and *v* are determined as semantically relevant with the degree of relevance being measured in terms of similarity value between the keyword sets. According to one implementation, the similarity value can be determined by the cosine similarity analysis. If the "NO" path is traced to block 414, then the keyword sets *u* and *v* are determined as semantically irrelevant.

[0056] At block 410, the similarity calculation module 112 computes the similarity value between the keyword sets *u* and *v* using cosine similarity analysis. In one implementation, the cosine similarity analysis can be represented by the following equation (11):

$$Cos(q_u, q_v) = \frac{\sum_{\forall q_j} C_{u,j} \cdot C_{v,j}}{\sqrt{\sum_{\forall q_j} C_{u,j}^{2}} \cdot \sqrt{\sum_{\forall q_j} C_{v,j}^{2}}}$$

(11)

wherein the different variables possess the same meaning as indicated previously.

[0057] At block 412, a comparison is made between the similarity value and an associated threshold value. In one implementation, the similarity calculation module 112 compares the similarity value with the threshold 220. In one implementation, the threshold 220 can be determined by the threshold calculation module 214. The similarity calculation module 112 determines whether the similarity value is lesser than the threshold 220 or not. For example, when the similarity value is lesser than the threshold 220, the "YES" path is traced to the block 414 to determine that the keyword sets *u* and *v* are semantically irrelevant or non relevant On the other hand, when the similarity value is greater than or equal to the threshold 220, the "NO" path is traced to block 416 to determine that the keyword sets *u* and *v* are semantically relevant.

[0058] At block 414, keyword sets that are determined as irrelevant are displayed. In one implementation, the similarity calculation module 112 displays the keyword sets *u* and *v* as semantically irrelevant when the "YES" path is traced from the block 412, i.e., when the similarity value between the keyword sets *u* and *v* falls below the threshold 220, the keyword sets are determined as semantically non relevant and displayed as irrelevant keyword sets.

[0059] At block 416, keyword sets determined as relevant are displayed. In one implementation, the similarity calculation module 112 displays the keyword sets *u* and *v* as semantically relevant when the "NO" path is traced from the block 412 i.e. when the similarity value between the keyword sets *u* and *v* exceeds the threshold 220. For all similarities values that are greater than the threshold 220, the keyword sets are semantically relevant and displayed as relevant keyword sets.

[0060] FIG. 5 illustrates an exemplary method for calculating the threshold 220.

[0061] At block 502, the threshold calculation module 214 identifies one or more training search session logs of the user. In one implementation, the user is identified by the threshold calculation module using either GUID or UUID. After identifying the user, the threshold calculation module 214 identifies the training search session logs of the user.

[0062] At block 504, training search session logs are segmented to obtain training search sessions. In one implementation, the threshold calculation module 214 segments the training search session logs of the user into one or more training search sessions based on a predetermined training session time interval.

[0063] At block 506, a first predetermined number of queries are generated. The first predetermined number of queries can be considered to be a training set form the training search sessions query logs. For example, the threshold calculation module 214 generates a first predetermined number of commonly occurring queries from the training search session logs, for example m, collectively referred as a training set. In one implementation, the training search session logs can be extracted from a query log that is pertinent to a definite time period and associated with the user.

[0064] At block 508, a second predetermined number of suggested queries are generated for each of the queries in the training set. The generated suggested queries are such that they are relevant to the query in consideration. In one implementation, the threshold calculation module 214 derives a second predetermined number of suggested queries, as indicated by a value say n, relevant to each query in the training set. The relevancy of the query with the second predetermined number of the suggested queries n can be obtained by various relevance or similarity analysis methods. In one implementation, the similarity analysis can be performed either by the mutual information analysis or by the cosine similarity analysis. The relevance information obtained can then be used to construct a candidate suggestion query set, with m*n number of queries present in the candidate suggestion query set. Based on the relevance information present in the candidate suggestion query set, the relevant suggested queries are obtained.

[0065] At block 510, the suggested queries are classified into one or more groups on the basis of their relevance. In one implementation, the threshold calculation module 214 classifies the relevant suggested queries into one or more groups. The groups are then labeled based on the level of relevance. In one implementation, the groups are labeled as "quite relevant", "relevant", and "irrelevant". In other implementations, the groups are labeled in a different fashion. The process of labeling the groups can be either automated or performed manually, for example, by a system administrator.

[0066] At block 512, the similarity threshold value associated with each of the groups is determined. In one implementation, the threshold calculation module 214 determines the threshold 220 for each of the group or each of the relevance level. The threshold 220 can be either the mutual information threshold or the cosine similarity threshold as determined by the threshold calculation module 214.

[0067] In another implementation, the threshold 220 can be used to filter the keyword sets that have high semantic relevance. The filtered semantically high relevant keyword sets interchangeably called as advertiser's keyword sets are then used by advertisers or others for further processing.

[0068] Any of the acts described above with respect to any method may be implemented by a processor or other computing device based on instructions stored on one or more computer-readable media. Computer-readable media can be any available media that can be accessed locally or remotely by the resource modeling application. By way of example, and not limitation, computer-readable media may comprise computer storage media and communication media. Computer storage media includes volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information such as computer-readable instructions, data structures, program modules, or other data. Computer storage media includes, but is not limited to, RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to store the desired information and which can be accessed by the resource modeling application. Communication media typically embodies computer-readable instructions, data structures, program modules, or other data on any information delivery media. Combinations of the any of the above should also be included within the scope of computer-readable media.

Conclusion

[0069] Although the invention has been described in language specific to structural features and/or methodological acts for implementing an exemplary method for determining relevant keyword set, it is to be understood that the invention is not necessarily limited to the specific features or acts described. Rather, the specific features and acts are disclosed as exemplary forms of implementing the invention.

What is claimed is:

1. A method comprising:

identifying search session logs of a user;

determining a first keyword set and a second keyword set from a search session in the search session logs;

calculating semantic relevance between the first and second keyword sets based on frequencies at which the first and second keyword sets occur; and

displaying one or more semantically relevant keyword sets based on the calculation.

2. The method of claim 1, wherein the identifying comprises:

identifying the user using either a globally unique identifier (GUID) or a universally unique identifier (UUID); and

aligning data from at least one raw session log with the user.

3. The method of claim 1, wherein the determining comprises:

segmenting the search session logs into search sessions; and

determining a semantic relationship between user queries made during the search session to compose a set of related keyword sets, the set of related keyword sets comprising at least the first keyword set and the second keyword set.

4. The method of claim 3, wherein the search sessions are defined based upon a search session time interval.

5. The method of claim 1, wherein the semantic relevance is determined by following relationship:

$$f_{uv} > \sqrt{f_{\min}}\ (f_u, f_v)$$

where f_u is the frequency of the first keyword set,

f_v is the frequency of the second keyword set,

f_{uv} is the frequency of the first and second keyword sets together as a pair, and

$f_{\min}$ is equal to the smaller of the values f_u and f_v.

6. The method of claim 1, wherein the calculating comprises calculating a similarity value between the first and second keyword sets, further wherein the similarity value is calculated based on either a mutual information analysis or a cosine similarity analysis.

7. The method of claim 6, wherein the one or more semantically relevant keyword sets are determined by the mutual information analysis based on the following relationship:

$$\sqrt{f_m} >= f_{uv} > \sqrt[4]{f_m}$$

where f_{uv} is the frequency of the first and second keyword sets together as a pair, and

f_{m} is equal to the smaller of the values for the frequency of the first keyword set and the frequency of the second keyword set,

wherein the mutual information analysis is conducted based on the following equation,

$$MI(q_u,q_v) = p(q_u,q_v)*\log\frac{p(q_u,q_v)}{p(q_u)*p(q_v)}$$

where $p(q_u,q_v) = \frac{C_{u,v}}{N}$,

$p(q_u) = \frac{C_u}{N}$,

$p(q_v) = \frac{C_v}{N}$,

N is the total number of query sessions,

C_u is the number of queries for keyword set u occurring in all sessions,

C_v is the number of queries for keyword set v occurring in all sessions, and

C_{uv} is the number of queries for keyword set u and v together as a pair occurring in all sessions.

8. The method of claim 6, wherein the one or more semantically relevant keyword sets are determined by the cosine similarity analysis based on the following relationship:

$$f_{uv} =< \sqrt[4]{f_{\mathrm{m}}},$$

where f_{uv} is the frequency of the first and second keyword sets together as a pair, and f_{m} is equal to the smaller of the values f_u and $f_{v,}$

wherein the cosine similarity analysis is conducted based on the following equation,

$$Cos(q_u, q_v) = \frac{\sum_{j=1}^{n} C_{u,j} * C_{v,j}}{\sqrt{\sum_{j=1}^{n} C_{u,j}^2} * \sqrt{\sum_{j=1}^{n} C_{v,j}^2}}$$

C_u is the number of queries for keyword set u occurring in all sessions, and

C_v is the number of queries for keyword set v occurring in all sessions.

9. The method of claim 6, wherein the calculating further comprises comparing the similarity value with a similarity threshold.

10. The method of claim 9, wherein the similarity threshold is determined by:

generating a training set, wherein the training set includes a first predetermined number of queries obtained from a training search session;

identifying a second predetermined number of suggested queries for each query in the training set based on a relevance of the suggested queries to the query, wherein the relevance is determined either by the mutual information analysis or the cosine similarity analysis;

categorizing the suggested queries into one or more groups based on the relevance; and

determining the similarity threshold for each of the groups.

11. A computing device comprising:

a memory;

one or more processors operatively coupled to the memory;

a session segmentation module configured to segment search sessions the search session logs of a user into search sessions and to identify a first keyword set and a second keyword set from a search session; and

a similarity calculation module configured to calculate a similarity value for the first and second keyword sets based on frequencies at which the first and second keyword sets occur in the search session.

12. The computing device of claim 11, wherein the similarity value is determined based on one or more of a total number of query sessions, a first number of queries including the first keyword set, a second number of queries including the second keyword set and a third number of queries including the first and second keyword sets together as a pair.

13. The computing device of claim 11, wherein the similarity calculation module is further configured to compare the similarity value with a similarity threshold and to display one or more semantically relevant keyword sets based on the comparison.

14. The computing device of claim 13 further comprising:

a threshold calculation module configured to compute the similarity threshold by:

organizing one or more training keyword sets into one or more groups based on a relevance measure of the training keyword sets and suggested queries; and

determining the similarity threshold for each of the groups.

15. The computing device of claim 14, wherein the threshold calculation module is configured to organize the training keyword sets by:

generating a training set which includes a first predetermined number of queries obtained from a training search session;

identifying a second predetermined number of the suggested queries for each query in the training set based on a relevance of the suggested queries to the query; and

categorizing the suggested queries into the one or more groups based on the relevance.

16. A computer-readable data storage medium having a set of computer readable instructions that, when executed by a processor, perform acts comprising:

identifying search session logs of a user;

computing semantic relevance between a first keyword set and a second keyword set in the search session logs based on frequencies at which the first and second keyword sets occur ; and

displaying one or more semantically relevant keyword sets based on the computation.

17. The computer-readable data storage medium of claim 16, wherein the identifying comprises:

parsing the search session logs into search sessions, wherein the search sessions are determined based upon a search session time interval; and

deriving semantically related user queries made during the search sessions to compose a set of related keyword sets wherein the set of related keyword sets comprise of at least the first keyword set and the second keyword set.

18. The computer-readable data storage medium of claim 16, wherein the computing comprises:

measuring a similarity value between the first and second keyword sets based on the frequencies; and

evaluating the similarity value against a similarity threshold.

19. The computer-readable data storage medium of claim 18, wherein the similarity value is calculated based on one or more of a total number of query sessions, a first number of queries including the first keyword set, a second number of queries including the second keyword set and a third number of queries including the first and second keyword sets together as a pair.

20. The computer-readable data storage medium of claim 18, wherein the similarity threshold is computed by:

generating a training set wherein the training set includes a first predetermined number of queries obtained from a training search session;

recognizing a second predetermined number of suggested queries for each query in the training set based on a relevance of the suggested queries to the query, wherein the relevance is determined either by a mutual information analysis or a cosine similarity analysis; and

determining a similarity threshold for each group of suggested queries, wherein the suggested queries are grouped into one or more groups based on the relevance.

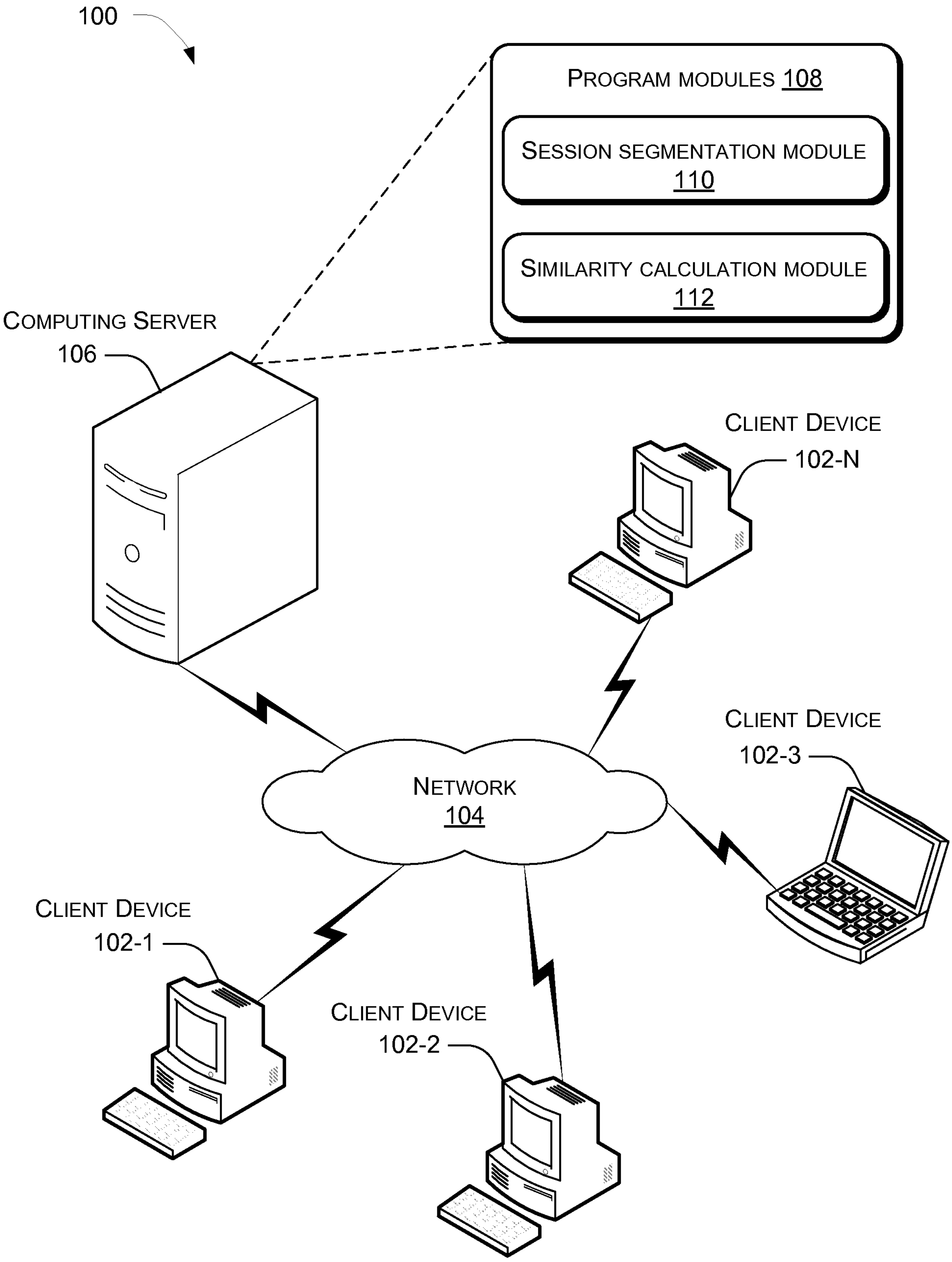
100
PROGRAM MODULES 108
SESSION SEGMENTATION MODULE 110
SIMILARITY CALCULATION MODULE 112
COMPUTING SERVER 106
CLIENT DEVICE 102-N
CLIENT DEVICE 102-3
NETWORK 104
CLIENT DEVICE 102-1
CLIENT DEVICE 102-2

FIG. 1

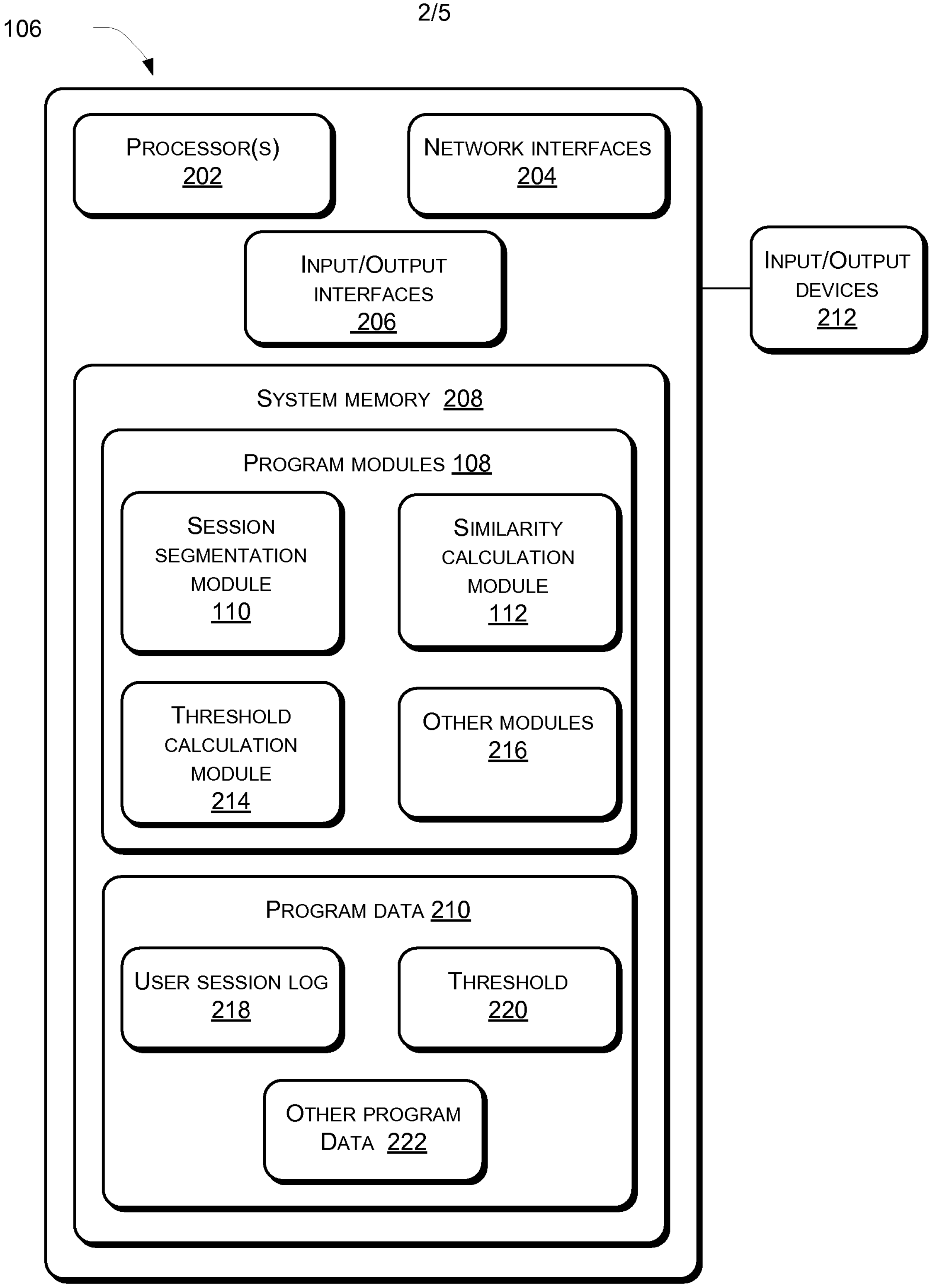
106
PROCESSOR(S) 202
NETWORK INTERFACES 204
INPUT/OUTPUT INTERFACES 206
INPUT/OUTPUT DEVICES 212
SYSTEM MEMORY 208
PROGRAM MODULES 108
SESSION SEGMENTATION MODULE 110
SIMILARITY CALCULATION MODULE 112
THRESHOLD CALCULATION MODULE 214
OTHER MODULES 216
PROGRAM DATA 210
USER SESSION LOG 218
THRESHOLD 220
OTHER PROGRAM DATA 222

FIG. 2

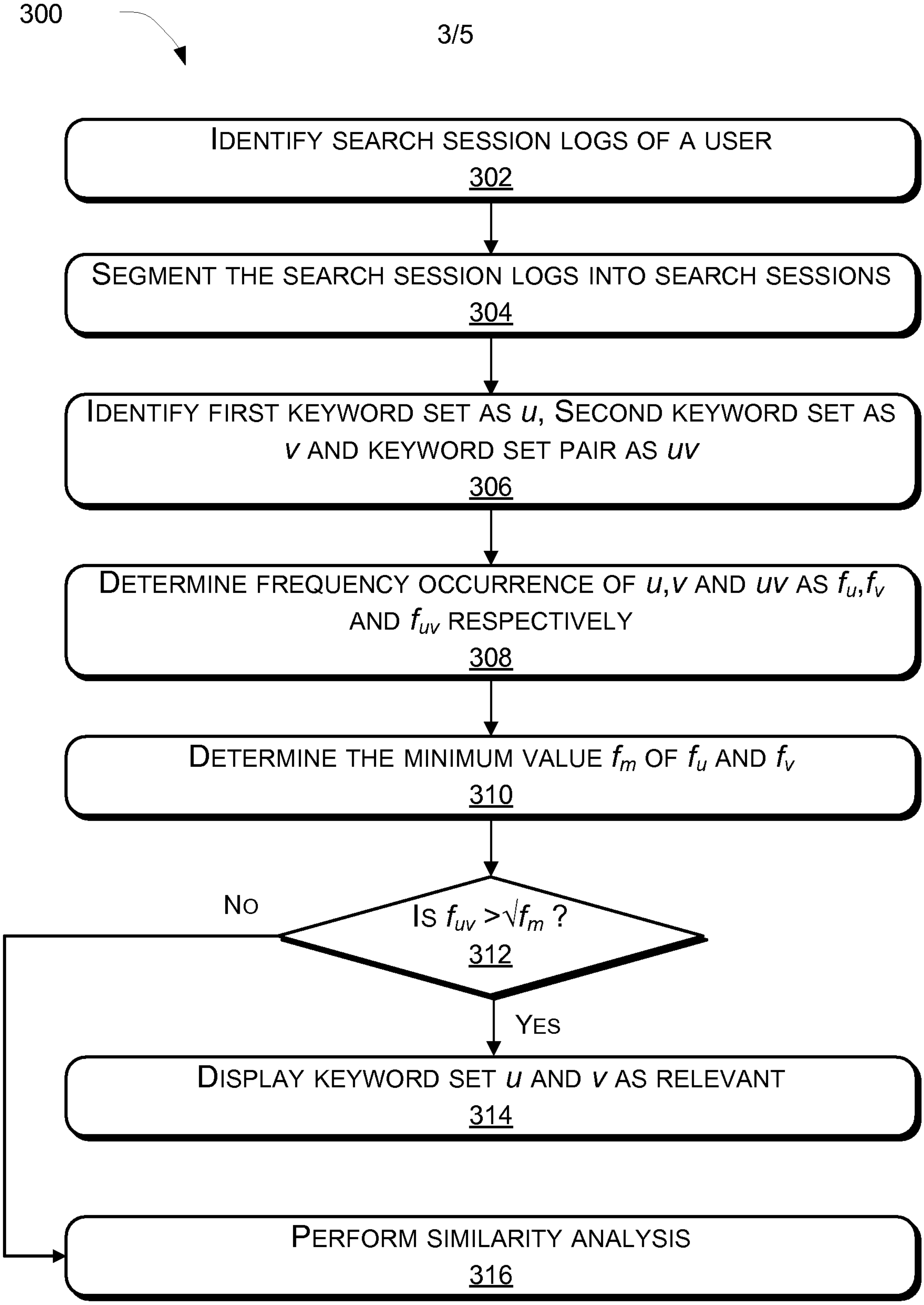
300
IDENTIFY SEARCH SESSION LOGS OF A USER
302
SEGMENT THE SEARCH SESSION LOGS INTO SEARCH SESSIONS
304
IDENTIFY FIRST KEYWORD SET AS u, SECOND KEYWORD SET AS v AND KEYWORD SET PAIR AS uv
306
DETERMINE FREQUENCY OCCURRENCE OF u,v AND uv AS f_u,f_v AND f_{uv} RESPECTIVELY
308
DETERMINE THE MINIMUM VALUE f_m OF f_u AND f_v
310
NO
IS $f_{uv} > \sqrt{f_m}$?
312
YES
DISPLAY KEYWORD SET u AND v AS RELEVANT
314
PERFORM SIMILARITY ANALYSIS
316

FIG. 3

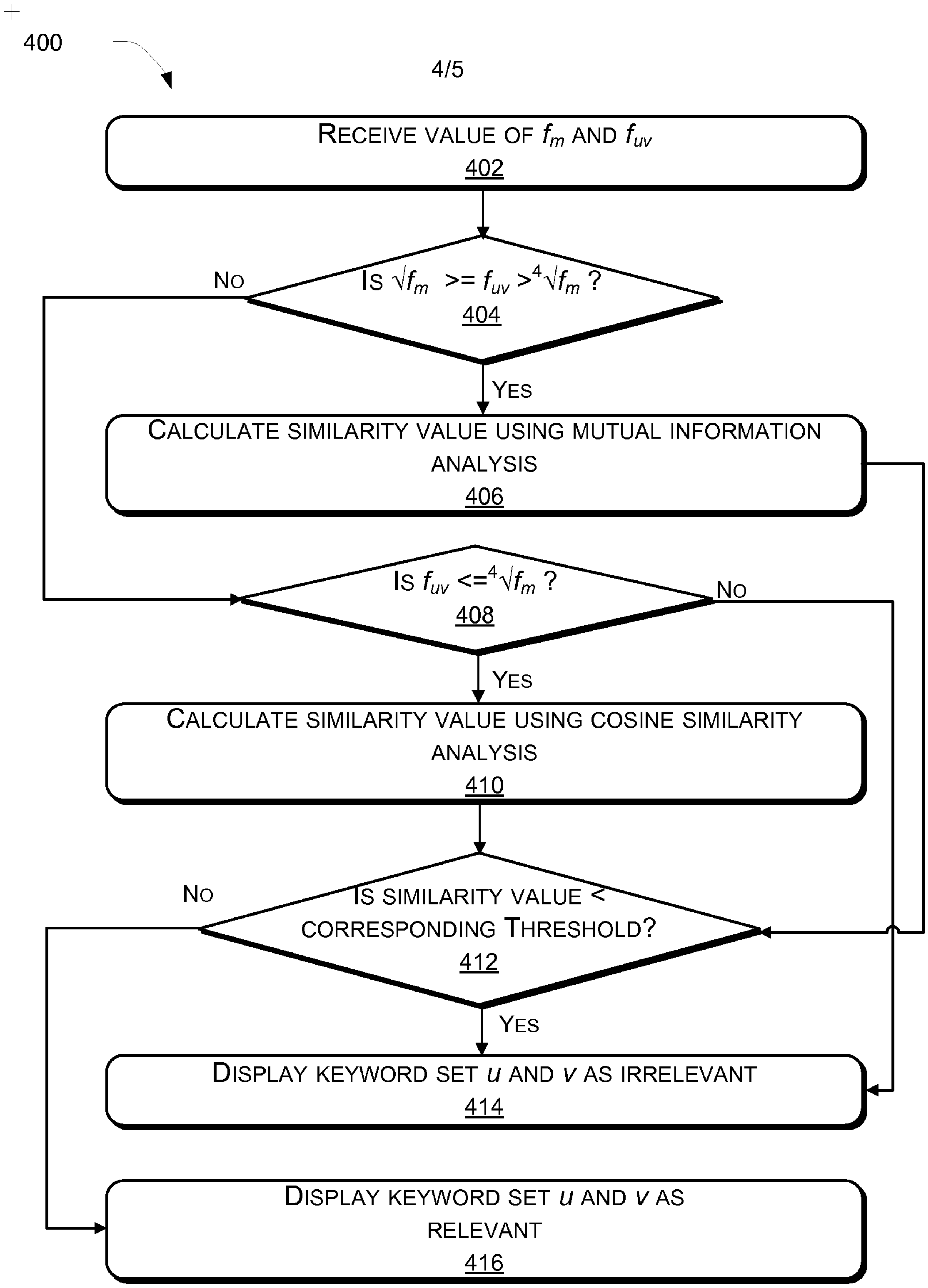
400
RECEIVE VALUE OF f_m AND f_{uv}
402
IS $\sqrt{f_m} >= f_{uv} > \sqrt[4]{f_m}$?
404
NO
YES
CALCULATE SIMILARITY VALUE USING MUTUAL INFORMATION ANALYSIS
406
IS $f_{uv} <= \sqrt[4]{f_m}$?
408
NO
YES
CALCULATE SIMILARITY VALUE USING COSINE SIMILARITY ANALYSIS
410
IS SIMILARITY VALUE < CORRESPONDING THRESHOLD?
412
NO
YES
DISPLAY KEYWORD SET u AND v AS IRRELEVANT
414
DISPLAY KEYWORD SET u AND v AS RELEVANT
416

FIG. 4

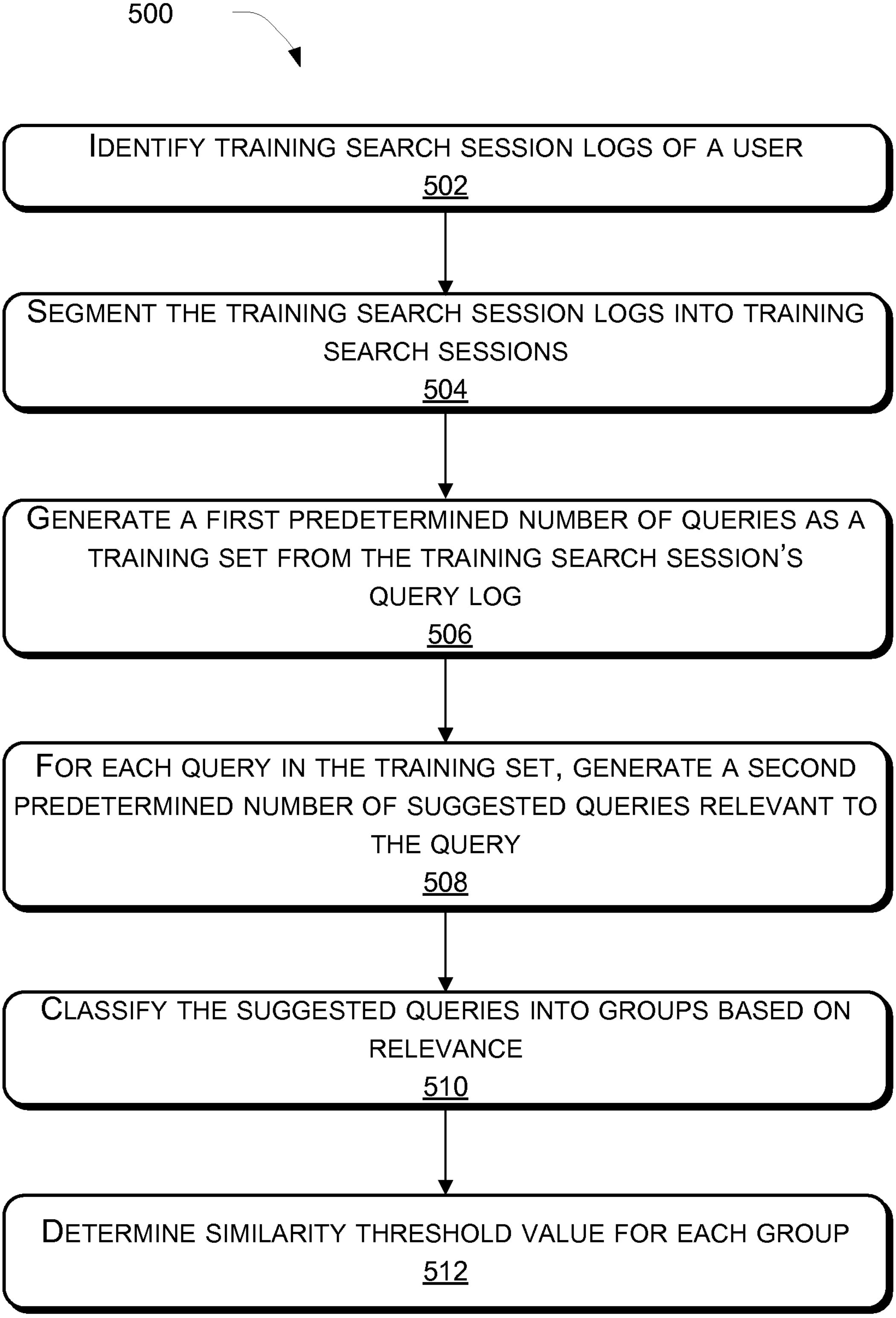
500
IDENTIFY TRAINING SEARCH SESSION LOGS OF A USER
502
SEGMENT THE TRAINING SEARCH SESSION LOGS INTO TRAINING SEARCH SESSIONS
504
GENERATE A FIRST PREDETERMINED NUMBER OF QUERIES AS A TRAINING SET FROM THE TRAINING SEARCH SESSION'S QUERY LOG
506
FOR EACH QUERY IN THE TRAINING SET, GENERATE A SECOND PREDETERMINED NUMBER OF SUGGESTED QUERIES RELEVANT TO THE QUERY
508
CLASSIFY THE SUGGESTED QUERIES INTO GROUPS BASED ON RELEVANCE
510
DETERMINE SIMILARITY THRESHOLD VALUE FOR EACH GROUP
512

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2008/055581

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30(2006.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC8 : G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Korean Utility models and applications for Utility models since 1975
Japanese Utility models and applications for Utility models since 1975

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
eKIPASS(KIPO internal) "session", "keyword", "advertisement", "similarity"

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 6,269,361 B1 (DARREN J. DAVIS, et al.) 31 July 2001 see abstract, column 4 line 51 - column 5 line 34, and figure 2.	1-20
A	US 7,181,447 B2 (ANDY CURTIS, et al.) 20 February 2007 see abstract, and figure 1.	1-20
A	US 5,987,464 A (ERIC SCHNEIDER) 16 November 1999 see abstract, column 4 line 54 - column 6 line 24, and figure 3.	1-20

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

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents,such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search 31 JULY 2008 (31.07.2008)	Date of mailing of the international search report **31 JULY 2008 (31.07.2008)**
Name and mailing address of the ISA/KR Korean Intellectual Property Office Government Complex-Daejeon, 139 Seonsa-ro, Seo-gu, Daejeon 302-701, Republic of Korea Facsimile No. 82-42-472-7140	Authorized officer LEE, Seok Hyung Telephone No. 82-42-481-8507

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2008/055581

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 6269361 B1	31.07.2001	AU 2000-51714 A1	18.12.2000
		AU 2000-51714 B2	18.12.2000
		AU 2000-51714 A5	18.12.2000
		AU 769955 B2	12.02.2004
		BR 200011035 A	26.02.2002
		CA 2375132 AA	07.12.2000
		CA 2375132 A1	07.12.2000
		CN 1378674 A	06.11.2002
		DE 20023291 U1	07.08.2003
		EP 1208500 A1	29.05.2002
		EP 1208500 A4	14.04.2004
		IL 146706 A0	25.07.2002
		JP 2003-501729 A	14.01.2003
		JP 3676999 B2	27.07.2005
		KR 10-2002-0019042 A	09.03.2002
		PA 01012340 A	21.07.2003
		NZ 515534 A	29.08.2003
		US 2001-042064 A1	15.11.2001
		US 2001-047354 A1	29.11.2001
		US 2001-051940 A1	13.12.2001
		US 2002-165849 A1	07.11.2002
		US 2003-033292 A1	13.02.2003
		US 2003-055816 A1	20.03.2003
		US 2003-149622 A1	07.08.2003
		US 2003-208474 A1	06.11.2003
		US 2005-223000 AA	06.10.2005
		US 2005-289120 A9	29.12.2005
		US 2006-136404 AA	22.06.2006
		US 2006-143096 AA	29.06.2006
		US 2006-190328 AA	24.08.2006
		US 2006-190354 AA	24.08.2006
		US 2006-212447 AA	21.09.2006
		US 2006-247981 AA	02.11.2006
		US 6978263 BB	20.12.2005
		US 6983272 BB	03.01.2006
		US 7035812 BB	25.04.2006
		US 7065500 BB	20.06.2006
		US 7092901 BB	15.08.2006
		US 7110993 BB	19.09.2006
		US 7225182 BB	29.05.2007
		US 7231358 BB	12.06.2007
		WO 2000-73960 A1	07.12.2000
		ZA 200109564 A	17.02.2003

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2008/055581

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 7181447 B2	20.02.2007	CA 2546492 A1	23.06.2005
		CA 2546494 A1	23.06.2005
		EP 1697865 A2	06.09.2006
		EP 1706816 A2	04.10.2006
		EP 4813564 TD	03.05.2007
		EP 4813565 TD	05.04.2007
		EP 1697865 A2	06.09.2006
		EP 1697865 A4	10.10.2007
		EP 1706816 A2	04.10.2006
		JP 2007-513439 T2	24.05.2007
		JP 2007-513440 T2	24.05.2007
		US 2005-125374 A1	09.06.2005
		US 2005-125376 A1	09.06.2005
		US 2005-125392 A1	09.06.2005
		US 7152061 BB	19.12.2006
		WO 2005-057366 A2	23.06.2005
		WO 2005-057366 A3	29.12.2005
		WO 2005-057367 A2	23.06.2005
		WO 2005-057367 A3	09.03.2006
		WO 2005-057368 A2	23.06.2005
		WO 2005-057368 A3	02.03.2006
		WO 2005-057369 A2	23.06.2005
		WO 2005-057369 A3	01.06.2006
US 5987464 A	16.11.1999	US 6442549 BA	27.08.2002