

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
8 November 2007 (08.11.2007)

PCT

(10) International Publication Number
WO 2007/127676 A1

(51) International Patent Classification:

G06F 17/30 (2006.01)

(21) International Application Number:

PCT/US2007/067075

(22) International Filing Date: 20 April 2007 (20.04.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

11/414,405 28 April 2006 (28.04.2006) US

(71) Applicant (for all designated States except US): **YAHOO!, INC.** [US/US]; 701 First Avenue, Sunnyvale, CA 94089 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **SUN, Gordon** [US/US]; 322 St. Martin Drive, Redwood Shores, CA 94065 (US). **ZHENG, Zhaohui** [CN/US]; 505 Cypress Point Drive, Mountain View, CA 94043 (US).

(74) Agent: **OSTROW, Seth, H.**; Dreier LLP, 499 Park Avenue, New York, NY 10022 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, 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, 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.

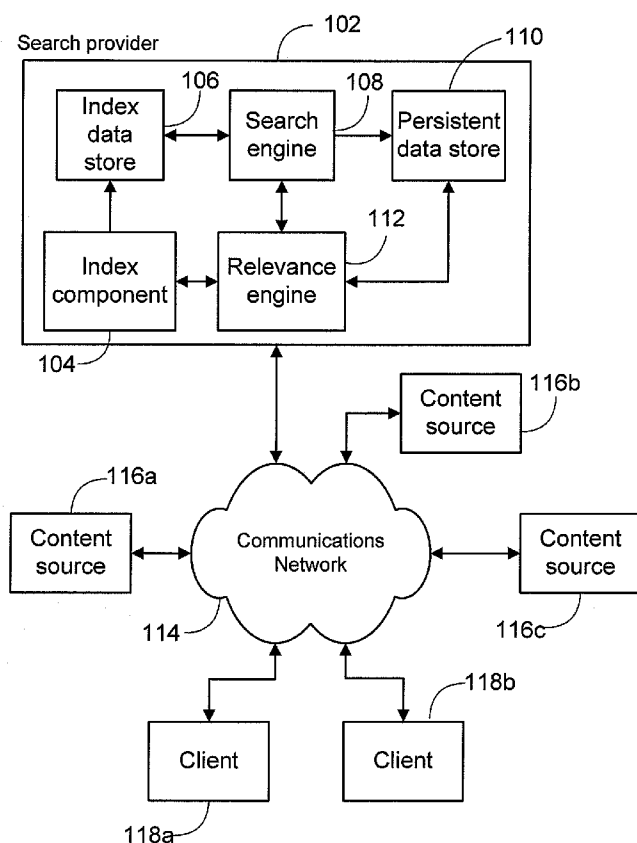
(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— with international search report

[Continued on next page]

(54) Title: SYSTEM AND METHOD FOR INDEXING WEB CONTENT USING CLICK-THROUGH FEATURES



(57) Abstract: The system and method of the present invention allows for the determination of the relevance of a content item to a query through the use of a machine learned relevance function that incorporates click-through features of the content items. A method for selecting a relevance function to determine a relevance of a query-content item pair comprises generating training set having one or more query-URL pairs labeled for relevance based on their click-through features. The labeled query-URL pairs are used to determine the relevance function by minimizing a loss function that accounts for click-through features of the content item. The computed relevance function is then applied to the click-through features of unlabeled content items to assign relevance scores thereto. An inverted click-through index of query-score pairs is formed and combined with the content index to improve relevance of search results.



-
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*
- For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

**SYSTEM AND METHOD FOR INDEXING WEB CONTENT USING CLICK-
THROUGH FEATURES**

COPYRIGHT NOTICE

[0001] A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure, as it appears in the Patent and Trademark Office patent files or records, but otherwise reserves all copyright rights whatsoever.

RELATED APPLICATIONS

[0002] This application is related to commonly owned U.S. patent application serial no. 10/424,170, filed on April 25, 2003 and titled "METHOD AND APPARATUS FOR MACHINE LEARNING A DOCUMENT RELEVANCE FUNCTION," attorney docket no. 5598/36, and U.S. patent application serial no. 11/343,910, filed on June 30, 2006 and titled "LEARNING RETRIEVAL FUNCTIONS INCORPORATING QUERY DIFFERENTIATION FOR INFORMATION RETRIEVAL," attorney docket no. 7346/76, both of which are hereby incorporated by reference herein in their entirety.

FIELD OF THE INVENTION

[0003] The invention disclosed herein relates generally to information searching and retrieval and, more specifically, to the determination of relevance of content items retrieved in response to a search query based on the click-through features of the content items.

BACKGROUND OF THE INVENTION

[0004] The Internet, which allows access to billions of content items stored on host computers around the world, represents a particularly diverse and large collection of content items. Development of a search engine that can such index a large and diverse collection of content items, yet provide the user a short, relevant result set of content items in response to a query has long been recognized as a problem in information retrieval. For example, a user of a search engine typically supplies a query to the search engine that contains only a few terms and expects the search engine to return a result set comprising relevant content items. Although a search engine may return a result set comprising tens, hundreds, or more content items, most users are likely to only view the top several content items in the result set. Thus, to be useful to a user, a search engine should determine those content items in a given result set that are most relevant to the user, or that the user would be most interested in, on the basis of the query that the user submits.

[0005] A user's perception of the relevance of a content item to a query is influenced by a number of factors, many of which are highly subjective. These factors are generally difficult to capture in an algorithmic set of rules represented by a relevance function. Furthermore, these subjective factors may change over time, as for example when current events are associated with a particular query term. As another example, changes over time in the aggregate content of the content items available through the Internet may also alter a user's perception of the relative relevance of a given content item to a given query. Users who receive search result sets that contain results not perceived to be highly relevant become frustrated and potentially abandon the use of the search engine. Designing effective and efficient retrieval functions is therefore of high importance to information retrieval.

[0006] In the past, search engine designers have attempted to construct relevance functions that take a query and a content item as a set of inputs and return a relevance value, which indicates the relevance of the content item to the query. The relevance value may be used,

for example, to order by relevance a set of content items that are responsive to a given query. For the ordering to be useful, however, the underlying relevance function should accurately and quickly determine the relevance of a given content item to a given query. Many retrieval systems and methods are known to those of skill in the art, including vector space models, probabilistic models and language modeling methods. In constructing its relevance functions, however, existing retrieval systems do not effectively use information regarding user-made judgments of the relevance of a content item to a given queries expressed as clickthrough information, which enables formulation of relevance functions with improved accuracy and effectiveness over existing systems and techniques.

SUMMARY OF THE INVENTION

[0007] Click-through information collected by the search engines during searching and retrieval of content items in response to user queries reflects user preferences and can be considered as implicit user judgment on the relevance of a content item to a query. To that end, disclosed herein are systems and methods for incorporating click-through-feature-based indexing for the identification of content items that are relevant to a given query. The system is operative to extract click-through features of query-content item pairs and apply machine learning, e.g., regression methods and decision trees, to learn a click score function. The sample space for feature-oriented methods is the collection of the click-through feature vectors in the corpus of content items. A click score function trained based on click-through feature vectors provides an information retrieval system, such as an Internet or intranet search engine, with information that may be used in determining the relevance of a content item to a query based on the click-through features thereof.

[0008] According to one embodiment, the present invention is directed towards a method for selecting a click score function to determine a click score of one or more query-content item pairs. The method comprises generating a training set comprising one or more query-content item pairs labeled with click scores by a human judge based on the relevance of the pair. For

each labeled query-content item pair, a click score function is determined and a loss function is used to modify the click score function based on the click-through features of the content item. A click score function that produces the smallest loss is selected. The selected click score function may then be applied to the click-through feature vectors for a plurality of unlabeled query-content item pair for assigning click scores thereto.

[0009] According to one embodiment, the method for selecting the click score function comprises minimizing a difference between a click score for the given query-content item pair assigned by human judgments and an output of the click score function. According to another embodiment, the method for selecting the click score function comprises using a loss function selected from a set of loss functions including a least mean square loss function, a query pairwise loss function and a loss function that compares a desired ranking order with an output ranking order.

[0010] According to another embodiment, a method for indexing content items based on the application of a click score function to the click-through features of the content items is disclosed. According to one embodiment, the method comprises generating an inverted click-through index of the unlabeled content items and the associated query-score pairs, wherein a key to the index is a URL of a given content item. The inverted click-through index may be combined with a general content index by associating the query-score pairs with the content items in the content index. The combined content index provides an improved searching and document retrieval facility.

[0011] According to one embodiment, the method for indexing content items comprises retrieving one or more content items in a result set in response to receipt of the query from the user. For a given content item in the result set, a click-through feature vector is determined and a click score function is applied to the click-through feature vector to generate a click score, which is used to generate one or more <query, content-item, click-score> index entries for the given

content item. One or more one or more <query, click-score> pairs are indexed for the content item.

[0012] According to one embodiment, the method for indexing content items comprises retrieving one or more content items in a result set in response to receipt of a query from the user. A determination is made whether an inverted click-through index exists for a given content item. If the inverted click-through index exists, the given query is compared with the query-score pairs in the inverted click-through index for the given content item. If the given query differs from the query-score pairs in the inverted click-through index by one or more terms, a partial click score is computed for the given query based on the indexed query-score pairs associated with the given content item.

[0013] According to one embodiment, the present invention is directed towards a system for indexing content items based on click-through features. The system comprises a index component operative to determine a click score function based on a training set of labeled query-content item pairs and the click-through features thereof, assign click scores to a plurality of unlabeled query-content item pairs through application of the click score function to the one or more click-through features and generate an inverted click-through index of the unlabeled content items and the associated query-score pairs. A relevance engine is operative for each query-content item pair to receive one or more click scores for one or more clicked queries in the index and generate a relevance score therefore. The relevance engine may also receive and utilize additional features of the one or more content items in generating a given relevance score. A search engine is operative to retrieve one or more content items in a result set in response to receipt of the query from the user and order the content items in the result set according to the relevance scores from the relevance engine.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] The invention is illustrated in the figures of the accompanying drawings which are meant to be exemplary and not limiting, in which like references refer to like or corresponding parts, and in which:

[0015] Fig. 1 is a block diagram presenting a system for indexing and searching web content based on the click-through features of the content items according to one embodiment of the present invention;

[0016] Fig. 2 is a flow diagram illustrating a method for computing a click score function according to one embodiment of the present invention;

[0017] Fig. 3 is a flow diagram illustrating a method for indexing unlabeled content items based on click-through features according to one embodiment of the present invention; and

[0018] Fig. 4 is a flow diagram illustrating a method for searching indexed content items according to one embodiment of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

[0019] Embodiments of a method, system, and article of manufacture comprising software programs for utilizing feature-based retrieval functions for information retrieval systems in accordance with the present invention are described herein with reference to the drawings in Figs. 1 through 4.

[0020] Turning to Fig. 1, a system comprising hardware and software components configured according to one embodiment of the present invention is illustrated.

According to the embodiment of FIG. 1, a search provider 102 comprises one or more software and hardware components operative to facilitate the identification of content items responsive to a query and calculate the relevance of the content items to the query, which may be used for ranking purposes. The hardware and software components include, but are not limited to, an index component 104, a search engine 108 and a relevance engine 112. The search provider 102 also comprises one or more data stores for providing persistent storage of information used in

searching and calculating relevance values for content items including, but not limited to, an index data store 106 and a persistent data store 110. Those of skill in the art should note that although the embodiment of Fig. 1 illustrates information retrieval in the context of a search engine, the systems and methods of the present invention are generally applicable to information retrieval tasks.

[0021] The search provider 102 is communicatively coupled with a network 114, which may include a connection to one or more local and wide area networks, such as the Internet and may provide access to linked networks of content items, e.g., the World Wide Web. Using the network 114, the search provider 102 is capable of accessing content sources 116a, 116b and 116c that host content items a user may wish to locate through use of the search engine 108 at the search provider 102. The search provider 102 may communicate with one or more content sources 116a, 116b and 116c for maintaining cached copies of content items that the content sources 116a, 116b and 116c host. The collection of content items, as well as information regarding content items, is referred to as “crawling”, and is the process by which the search provider 102 collects information upon which the search engine 108 performs searches.

Exemplary methods for crawling information on a network is described in commonly owned U.S. Patent No. 6,021,409, entitled “METHOD FOR PARSING, INDEXING AND SEARCHING WORLD-WIDE-WEB PAGES,” the disclosure of which is hereby incorporated by reference in its entirety.

[0022] The search provider 102 crawls content sources 116a, 116b and 116c in communication with the network 114, such as the Internet, which may comprise collecting combinations of content items and information regarding the same. An index component 104 parses and indexes the content items and related information that the search provider 102 collects through the crawling process. The index component 104 generates a content index that defines a structure for the content items and related information that allows for location and retrieval of the content items and related information. According to one embodiment of the invention, the index

component 104 creates a content index of word-location pairs that allows a search engine 108 to identify specific content items and information regarding the same in response to a query, which may be from a user, software component, automated process, etc. Exemplary methods for indexing information is described in commonly owned U.S. Patent Application No. 5,745,889, entitled "METHOD FOR PARSING INFORMATION OF DATABASE RECORDS USING WORD-LOCATION PAIRS AND METAWORD-LOCATION PAIRS," the disclosure of which is hereby incorporated by reference in its entirety. The one or more indexes that the index component 104 generates are written to an index data store 106 for persistent storage and use by other components of the search provider 102.

[0023] A user of a client device 118a and 118b who desires to retrieve a content item from a content source 116a, 116b and 116c that is relevant to a particular topic, but who is unsure or ignorant regarding the address or location of the content item, submits a query to the search engine 108. According to one embodiment, a user utilizes a client device 118a and 118b to connect over the network 114 to the search engine 108 at the search provider 102 and provide a query. A typical query has one or more terms. For example, the query "2005 Lincoln continental" contains three terms and is referred to as a three-term query. Similarly, queries containing only one term are referred to as one-term queries, queries containing two terms are two-term queries, etc. A space or other delimiter character that the search engine 108 comprehends delimits individual terms comprising a query.

[0024] According to one embodiment of the invention, a client device 118a and 118b is a general purpose personal computer comprising a processor, transient and persistent storage devices, input/output subsystem and bus to provide a communications path between components comprising the general purpose personal computer. For example, a 3.5 GHz Pentium 4 personal computer with 512 MB of RAM, 40 GB of hard drive storage space and an Ethernet interface to a network. Other client devices are considered to fall within the scope of the present invention including, but not limited to, hand held devices, set top terminals, mobile handsets, etc. The

client device 118a and 118b typically run software applications (such as a web browser) that provide for transmission of queries to the search provider 102, as well as receipt and display of result sets responsive to the queries.

[0025] Upon receipt of the query, the search engine 108 examines the content index using the terms that comprise the query in an attempt to identify a result set that contains those content items that are responsive to the query. The search engine 108 formulates the result set for transmission over the network 114 and presentation to the user through use of the client device 118a and 118b. Where the result set comprises a link to content items, such as a uniform resource locator (URL), the user may select a link in the result set to navigate to the content source that is hosting the content item that the link identifies. The search engine 108 utilizes a persistent data store 110 for storage of an historical log of the queries that users submit, which may include an indication of the selection of items in results sets that the search engine 108 transmits to users.

[0026] As discussed previously, users become increasingly frustrated when presented with a result set that does not identify content items are more relevant to a given query prior to less relevant items. Accordingly, the present embodiment provides an indexing component 104 that is operative to utilize machine learning that is trained based on click-through features of the content items to determine a click score of a given content item to a given query for use in determining relevance of the query-content item pair. The index component 104 receives pairs of queries and content items, applying a click score function, the selection of which is described in greater detail herein, to determine a click score for the content item vis-à-vis the query.

[0027] The index component 104 utilizes click-through “features” of a given query-URL pair to determine click score for the URL with regard to the query. According to one embodiment, a feature is a quantification of an aspect of a relationship between a query and content item represented by its URL, which may include quantifying aspects of the query, the content item, or both. Such query-dependent features may be represented as one or more values

for a given query, a content item or both, and is referred to as a feature vector. In one embodiment, the index component 104 generates a click score based on one or more click-through features. The click-through features may include, but are not limited to, aggregated first-clicks, aggregated last-clicks, the average amount of time users stay on a website associated with the given URL, the spam score of each URL, the position of each URL in a given search results page, the expected clicks on that position, the frequency of that query in the query log, the popularity of each content item, etc. Using the click through features, the index component is operative to generate a click score that may be stored in the content index as another feature of a URL that the relevance engine may utilize in determining the relevance of a URL to a given query, e.g., for ranking purposes.

[0028] The index component applies a click score function to a given query-URL pair determine a click score for the pair. According to one embodiment, the click function that the index component utilizes is determined using machine learning techniques. The index component 104 receives a training set of query-content item pairs and a vector of click-through features associated therewith. The content items may be represented by their respective URLs. Content items from the training set to may be presented to one or more human subjects for labeling, which is the assignment of training click scores indicating the relevance of content items in the training set to the query. Alternatively, the index component 104 may access the persistent data store 110 to retrieve a past query (training query) and corresponding result set (training set), utilizing selection information from a user regarding the selection of items in the result set in response to the query to determine training click scores for the content items in the training set. On the basis of the training click scores, content items in the training set and the training query, the index component 104 is operative to compute a click score function, which the index component 104 applies to the click-through features of unlabeled content items stored in the index data store 106 to assign click scores to each query-content item pair.

[0029] Once the click scores have been assigned to the query-content item pairs, the index component 104 is operative to generate an inverted click-through index for each content item in its index, which the index data store 106 maintains. The URL of a given content item may be used as a key to the inverted click-through index. The inverted index may then be persistently stored along with the content index in the index data store 106 and used by other components of the search provider 102. Alternatively, index component 104 may supplement the content index at the index data store 106 with click score information.

[0030] According to one embodiment, when a user of client device 118a or 118b submits a query to the search engine 108, the relevance engine 112 may utilize information in both the content and inverted click-through indexes, e.g., click score, for content items that are responsive to the query to generate relevance scores for the content items. These content items, or their URLs, may be retrieved from the index at the index data store 106 and presented to the user in order of relevance as indicated by the relevance engine 112.

[0031] In the event the query submitted by the user through a client device 118a and 118b does not have a matching query-score pair in the inverted click-through index for a responsive content item, the index component 104 computes a partial click score for the query-content item pair, which accounts for any extra or missing words in the given query that differ it from the indexed queries. The search engine 108 receives the click scores, or click scores, for the content items in the result set and passes these data to the relevance engine 112, which utilizes click scores (which may include utilization of other features) for calculating relevance scores. The search engine may use the relevance scores for ranking or ordering purposes, e.g., presenting the content items or links to content items with the highest relevance scores (more relevant) prior to the content items or links to content items in the result set with lesser relevance scores (less relevant). The search engine 108 transmits the ranked result set to the client device 118a and 118b for viewing by the user.

[0032] Fig. 2 presents a flow diagram illustrating one embodiment of a method for selecting a click score function for determining a click score of a query-content pair. The determination of a click score function is predicated on the generation of a training set, step 205. The training set may comprise a plurality of query-URL pairs. Each pair in the training set may have an associated log of click-through features, such as aggregated first-clicks, aggregated last-clicks, the average amount of time users stay on the website associated with the give URL, the spam score of each URL, the position of each URL in the content index, the expected clicks on that position, the frequency of that query in the query log, the popularity of the website identified by the URL, e.g., number of clicks, and the like. The click through features may be used to evaluate the relevance of a content item to a query. According to one embodiment, human candidates may evaluate the relevance of a content item to a query and to label each query-content pair in the training set according, step 210. A label may be a numeric value, for example, on the scale of zero to four, where four can indicate the highest relevance and zero the lowest. In other embodiments, various other forms of labeling known to those of skill in the art may be used to assign relevance values to the training set.

[0033] Having assigned labels to the training set of query-content item pairs, click-through features associated with the labeled query-content item pair may be extracted, step 215. Click-through feature vectors may be generated for each labeled pair in the training set, step 220. For example, the following vector representation may be used to represent K click-through features for a given labeled query-content item pair: $\langle ctrf_1, ctrf_2, \dots, ctrf_K \rangle$. If additional click-through features are added to the given query-content item pairs, the feature vectors for such pairs may be extended in the following manner:

$\langle ctrf_1, ctrf_2, \dots, ctrf_K, ctrf_{K+1}, ctrf_{K+2}, \dots, ctrf_{K+M} \rangle$, where M is the total number of click-through related features newly added, step 225.

[0034] Having identified the vector representations of the click-through features for each query-content item pair in the training set, a regression function may be determined for mapping

from various click-through features to relevance judgments on the labeled query-content item pairs, step 230. According to one embodiment, such a regression function mapping may be expressed in the following manner:

where x_i is the vector representation of i th query-content item pair in the training set in terms of its click-through features; y_i is its numerical value reflecting the degree of relevance, such as the values of 0, 1, 2, 3, or 4 that correspond to the editorial grades of perfect, excellent, good, fair and bad, respectively; N is the totally number of query-content item pairs in the training set; and h is a click score function.

[0035] A click score function, h , may be selected that results in the smallest possible loss over all query-content item pairs in the training set. According to one embodiment, such a click score function may be computed by minimizing the following empirical loss function:

The value of the loss function L is the summation of the square of the difference between a training click score for the query-content item pair and the output of a click score function h for determining relevance on the basis of the feature vector x_i . An optimal click score function results in a smallest value for L , e.g., zero. According to one embodiment, any regression method known to those of skill in the art may be applied to minimize the above loss function, such as linear regression, logistic regression, support vector machine (SVM), gradient boosting tree, or the like, step 230. The selected click score function is stored for use in determining a click score for unlabeled query-content item pairs.

[0036] Fig. 3 presents one embodiment of a method for using a selected click score

$$L(h) = \sum_{i=1}^N (y_i - h(x_i))^2$$

function to generate and index click scores for the unlabeled content items, e.g., web pages. A URL identifies each content item, which may also have one or more queries and click-through feature logs associated therewith. The click-through features for each content item may include, but are not limited to, aggregated first-clicks, aggregated last-clicks, the average amount of time

users stay on the website associated with the given URL, the spam score of each URL, the position of each URL, the expected clicks on that position, the frequency of that query in the query log, the popularity of each URL, etc.

[0037] According to one embodiment, the indexing component is operative to extract click-through features for each unlabeled query-content item pair in an index, step 305, and evaluate the extracted click-through features, which may be one or more features. The set of features that are selected may be made in advance of the evaluation, e.g., by a system administrator identifying the features for inclusion, or may be made at the time of the evaluation. The index component applies a click score function, which may be computed in accordance with a method described herein, to the selected click-through features to calculate a click score for the given query-content item pair, step 310. The computed click score may be stored along with the associated query and URL of the given content item in the respective content source in the following manner: <query, URL, click-through relevance score>. The index component checks if additional query-content item pairs require processing, step 315, and if so, repeats steps 305 and 310.

[0038] Having assigned click scores to query-content item pairs stored in the index, the index component generates an inverted click-through index of the content items, step 320. According to one embodiment, the index component indexes all query-score pairs into an inverted click-through index, a key to which is a URL associated therewith. The resultant index comprises a plurality of items having the following format: URL→ (query1, score1), (query2, score2), (query3, score3), ..., (query *n*, score *n*). The index component stores the inverted click-through index in the index data store, step 325, which may then be combined with the content index previously stored in the index data store to generate an improved content index, step 330. Thus, in the combined content index, each content item is designated by its URL and has a one or more click-through feature-based query-score pairs associated therewith.

[0039] The click-through index may be periodically checked to determine whether click scores associated with the indexed query-content item pairs require recalculation in view of newly available click-through features, e.g., from query logs. If additional click-through features have been added to the given query-content item pairs since the last index update, the feature vector for the given pair can be extended in the following manner:

$\langle ctrf_1, ctrf_2, \dots, ctrf_K, ctrf_{K+1}, ctrf_{K+2}, \dots, ctrf_{K+M} \rangle$, where M is the total number of click-through related features newly added. The index component may determine whether the new click-through features effect the click score of the given query-content item pair and, if so, recomputed the click score for the given query-content item pair by applying a click score function to the new feature vector.

[0040] Fig. 4 presents one embodiment of a method for ordering search results that accounts for click score information. A user submits a query to the search engine, step 405, which causes the search engine to identify one or more entries in an index in the index data score that are responsive to the query, step 410, e.g., relevant to the query. In some embodiments, only content items including or associated with one or more terms in the query are included in the result set, e.g., that contain user supplied tags that contain the terms. Alternatively, or in conjunction with the foregoing, the search engine may utilize other criteria to select content items for inclusion in the result set, such as a click score for the query-URL pair.

[0041] The search engine retrieves information from the index for the content items the result set, step 415. The index from which the search engine retrieves the information may include, but is not limited to, query-click score pairs, anchor text-weighting score pairs, title text, body text, etc. The search engine applies a relevance function to determine the relevance of a given content item in the result set to the query from the user, step 420. The relevance function receives these data from the index, which the relevance function utilizes to generate a relevance score for a given content item in the result set. The relevance function may be determined using a variety of information retrieval techniques known to those of skill in the art. For example, the

relevance function may be determined by hand-constructing the function through the combinations of text similarity techniques including, but not limited to, term frequency, inverse document frequency, proximity, document popularity measures (e.g., page rank), etc. The relevance function may be also determined using a variety of machine learning techniques known to those of skill in the art including, but not limited to, Support Vector Machine, Boosted Trees, other machine learning techniques.

[0042] The search engine applies the ranking function to data from the index for a given content item in the result set to generate a relevance score for the given content item, step 420. A check is performed to determine if any additional content items exist in the result set for which the search engine is to generate a relevance score, step 425. Where additional content items exist in the result set for processing, step 425, processing returns to step 420 with the search engine generate a relevance score for a given content item. The search engine generates relevance score for the items in the result set, which the search engine utilizes to rank or order the result set, step 430. The ordered result set is transmitted to the client device for presentation to the user, step 435.

[0043] The foregoing description, for purposes of explanation, uses specific nomenclature and formula to provide a thorough understanding of the invention. It should be apparent to those of skill in the art that the specific details are not required in order to practice the invention. The embodiment have been chosen and described to best explain the principles of the invention and its practical application, thereby enabling others of skill in the art to utilize the invention and various embodiments with various modifications as are suited to the particular use contemplated. Thus, the foregoing disclosure is not intended to be exhaustive or to limit the invention to the precise forms disclosed and those of skill in the art recognize that many modifications and variations are possible in view of the above teachings.

WHAT IS CLAIMED IS:

1. A method for indexing content items based on click-through features, the method comprising:

generating a training set comprising one or more query-content item pairs, wherein a given query-content item pair has one or more click-through features associated therewith;

labeling one or more query-content item pairs in the training set by assigning click score thereto based on the one or more click-through features thereof;

determining a click score function based on the click scores of the labeled query-content item pairs and the click-through features thereof;

applying the click score function to a plurality of unlabeled query-content item pairs to determine click scores thereof based on the one or more click-through features of the unlabeled query-content item pairs.

2. The method of claim 1 further comprising the step of identifying a content item using a Uniform Resource Locator ("URL").

3. The method of claim 1 wherein labeling a given content item in the training set comprises providing a given query-content item pair to a human judge to assign a click score.

4. The method of claim 1 wherein determining the click score function comprises minimizing a difference between a click score for the given query-content item pair and an output of the click score function.

5. The method of claim 1 wherein determining the click score function comprises using a loss function selected from a set of loss functions including least mean square loss function, a query pairwise loss function and a loss function that compares a desired ranking order with an output ranking order.

6. The method of claim 1 comprising storing the determined click score function for application to an unlabeled query-content item pair.

7. The method of claim 1 comprising generating an inverted click-through index of the unlabeled content items and the associated query-score pairs, wherein a key to the index is a URL of the content item.

8. The method of claim 7 comprising the step of combining the inverted click-through index with a content index by associating the query-score pairs with the content items identified in the content index.

9. The method of claim 8 comprising the steps of:

retrieving one or more content items in a result set in response to receipt of the query from the user;

for a given content item in the result set, determining a click-through feature vector for the given content item;

applying the click score function to the click-through feature vector for the given content item to generate a click score for the given content item;

generating one or more <query, content-item, click-score> data structures for the given content item;

indexing one or more <query, click-score> pairs for the content item from the <query, content-item, click-score> data structures.

10. A system for indexing and searching content items based on its click-through features, the system comprising:

an index component operative to determine a click score function based on a training set of labeled query-content item pairs and the click-through features thereof, assign click scores to a plurality of unlabeled query-content item pairs through application of the click score function to the one or more click-through features and generate an inverted click-through index of the unlabeled content items and the associated query-score pairs;

a relevance engine operative to receive one or more query scores for one or more content items and generate one or more relevance scores therefore; and

a search engine operative to retrieve one or more content items in a result set in response to receipt of the query from the user and order the content items in the result set according to the relevance scores from the relevance engine.

11. The system of claim 10 wherein a Uniform Resource Locator (URL) of a content item is used as a key to locate the given content item the inverted click-through index.

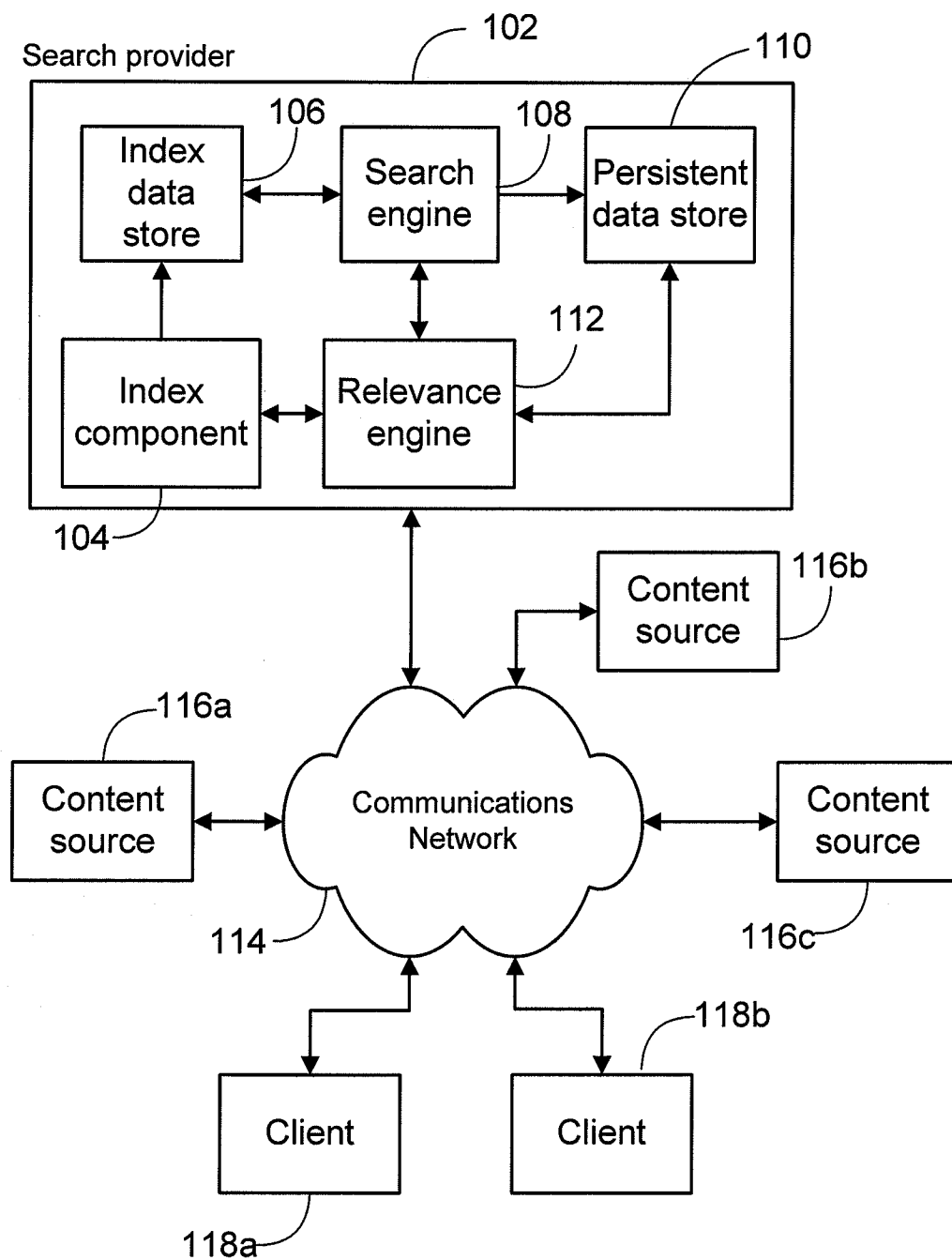
12. The system of claim 10 wherein the training set of labeled query-content items comprises a plurality of query-content items having click scores assigned thereto by a human judge.

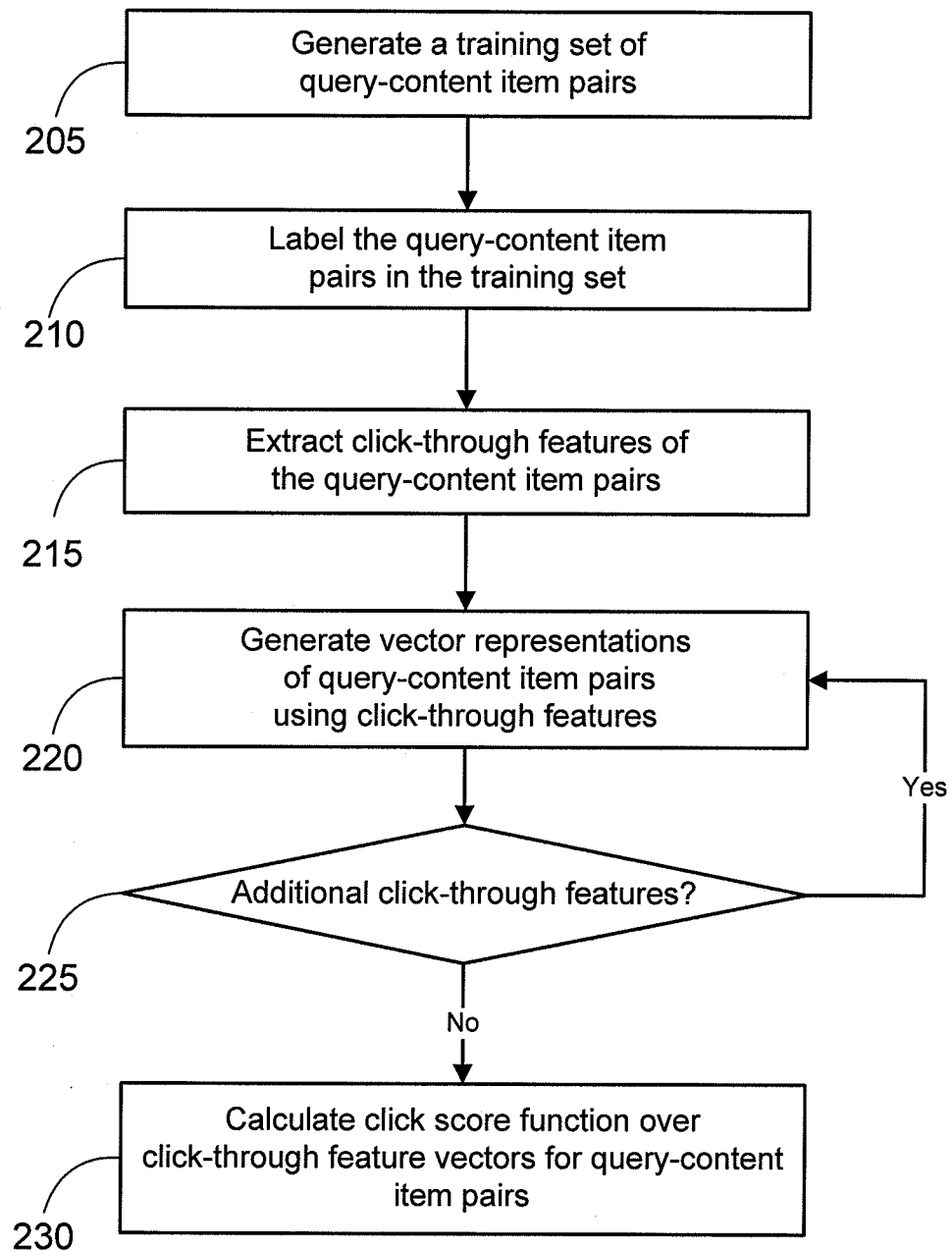
13. The system of claim 10 wherein the click score function is determined by minimizing difference between click score for the given query-content item pair in the training set and the output of the click score function.

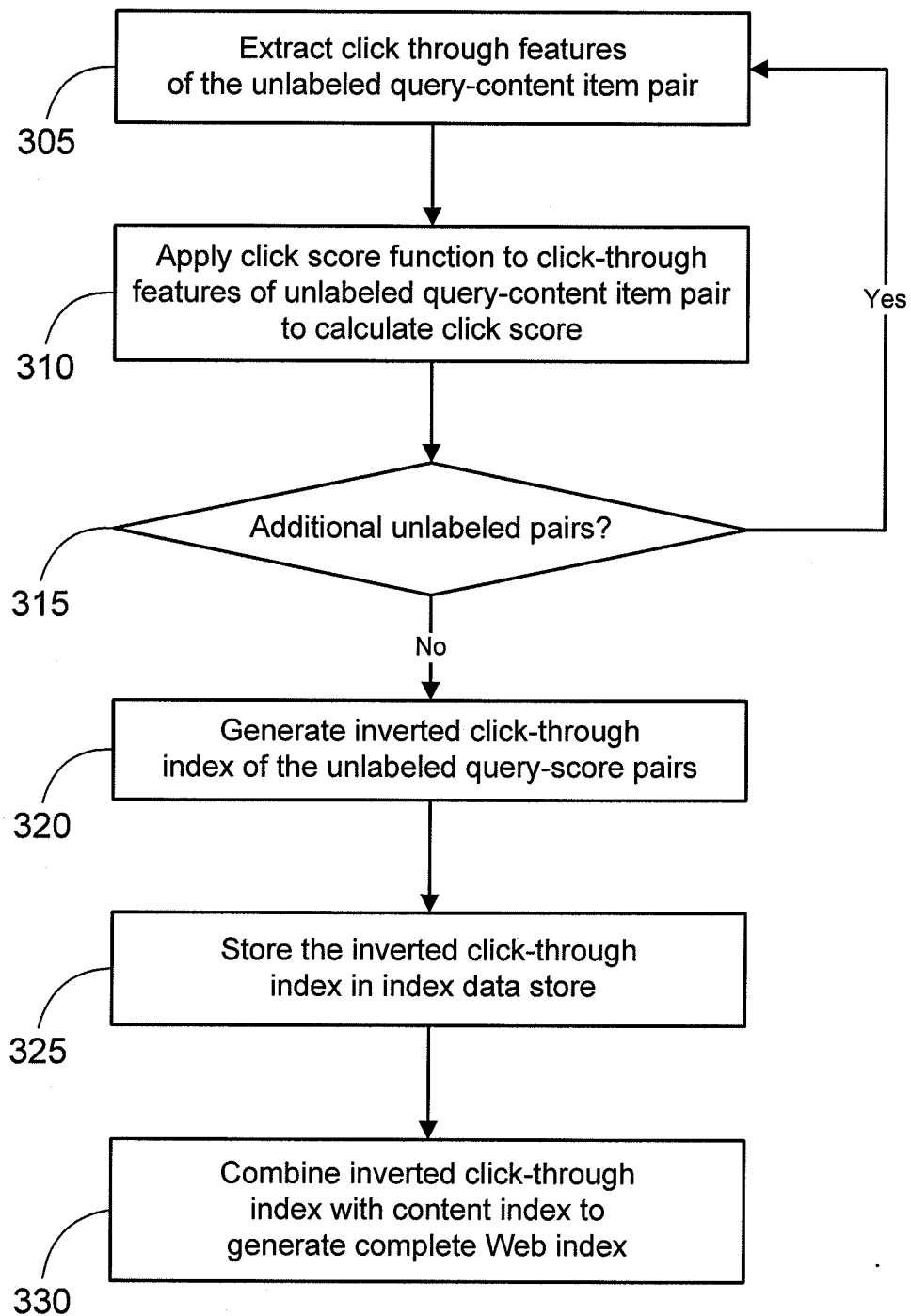
14. The system of claim 10 wherein the click score function is determined using a loss function selected from a set of loss functions including a least mean square loss function, a query pairwise loss function and a loss function that compares a desired ranking order with an output ranking order.

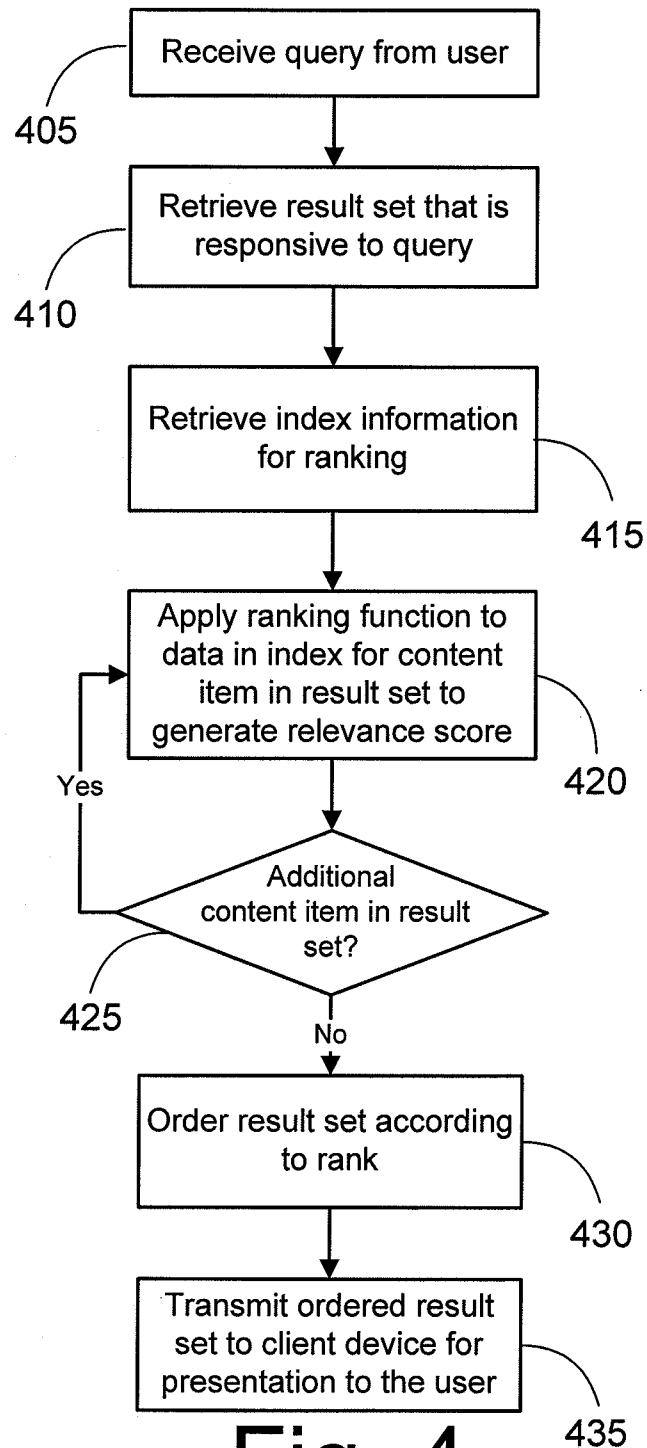
15. The system of claim 10 wherein the index component is operative to store the determined click score function for application to an unlabeled query-content item pair.

16. The system of claim 10 wherein the index component is operative to combine the inverted click-through index with a content index by associating the query-score pairs with the content items in the content index.

**Fig. 1**

**Fig. 2**

**Fig. 3**

**Fig. 4**

PATENT COOPERATION TREATY

PCT

INTERNATIONAL SEARCH REPORT

(PCT Article 18 and Rules 43 and 44)

Applicant's or agent's file reference 600189-500WO	FOR FURTHER ACTION see Form PCT/ISA/220 as well as, where applicable, item 5 below.	
International application No. PCT/US2007/067075	International filing date (<i>day/month/year</i>) 20 APRIL 2007 (20.04.2007)	(Earliest) Priority Date (<i>day/month/year</i>) 28 APRIL 2006 (28.04.2006)
Applicant YAHOO!, Inc. et al		

This International search report has been prepared by this International Searching Authority and is transmitted to the applicant according to Article 18. A copy is being transmitted to the International Bureau.

This international search report consists of a total of 3 sheets.

☐ It is also accompanied by a copy of each prior art document cited in this report.

1. **Basis of the report**

a. With regard to the **language**, the international search was carried out on the basis of:

☒ the international application in the language in which it was filed

☐ a translation of the international application into _____, which is the language of a translation furnished for the purposes of international search (Rules 12.3(a) and 23.1(b))

b. ☐ This international search report has been established taking into account the **rectification of an obvious mistake** authorized by or notified to this Authority under Rule 91 (Rule 43.6bis(a)).

c. ☐ With regard to any **nucleotide and/or amino acid sequence** disclosed in the international application, see Box No. I.

2. ☐ **Certain claims were found unsearchable** (See Box No. II)

3. ☐ **Unity of invention is lacking** (See Box No. III)

4. With regard to the **title**,

☒ the text is approved as submitted by the applicant.

☐ the text has been established by this Authority to read as follows:

5. With regard to the **abstract**,

☒ the text is approved as submitted by the applicant.

☐ the text has been established, according to Rule 38.2, by this Authority as it appears in Box No. IV. The applicant may, within one month from the date of mailing of this international search report, submit comments to this Authority.

6. With regard to the drawings,

a. the figure of the **drawings** to be published with the abstract is Figure No. 1

☐ as suggested by the applicant.

☒ because the applicant failed to suggest a figure.

☐ because this figure better characterizes the invention.

b. ☐ none of the figure is to be published with the abstract.

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)

IPC 8: G09C, G06F1, G06F15, G06K1, G06Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
KOREAN PATENTS AND APPLICATIONS FOR INVENTIONS SINCE 1975

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

PAJ, FPD, USPAT, eKIPASS, IEEE "SEARCH""ENGINE""INDEX""CLICK"

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Kazunari SUGIYAMA, Kenji HATANO, Masatoshi YOSHIKAWA, Shunsuke UEMURA, "Adaptive Web Search Based on User's Implicit Preference", DEWS 2004, MARCH 2004 see the abstract and Introduction, figure 1, "3. Our proposed Method" part	1-16
A	Jaime Teevan, Susan T. Dumais, Eric Horvitz, "Personalizing Search via Automated Analysis of Interests and Activities", SIGIR '05, AUGUST 15-19, 2005 see the abstract, figure 1, "3. PURSUIT OF PERSONALIZATION" part	1-16
A	US 2003/0163454 A1 (BRAIN JOCOBSEN, SOREN ALAIN MORTENSEN) 28 AUGUST 2003 (2003-08-23) see the abstract, figure 3B, paragraphs [0047]-[0050]	1-16



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

27 SEPTEMBER 2007 (27.09.2007)

Date of mailing of the international search report

27 SEPTEMBER 2007 (27.09.2007)

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
920 Dunsan-dong, Seo-gu, Daejeon 302-701,
Republic of Korea

Facsimile No. 82-42-472-7140

Authorized officer

CHEKAL Hyun

Telephone No. 82-42-481-5667



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2007/067075Patent document
cited in search reportPublication
datePatent family
member(s)Publication
date

US20030163454A1

28.08.2003

EP01341099A2

03.09.2003

EP01341099A3

26.05.2004

EP1341099A2

03.09.2003

EP1341099A3

26.05.2004

EP1341099A2

03.09.2003

US2003163454AA

28.08.2003