US 20070005568A1

(19) United States

(12) Patent Application Publication (10) Pub. No.: US 2007/0005568 A1

Angelo et al. (43) Pub. Date: Jan. 4, 2007

# (54) DETERMINATION OF A DESIRED REPOSITORY

(76) Inventors: **Michael Angelo**, San Francisco, CA (US); **David Braginsky**, Mountain View, CA (US); **Jeremy Ginsberg**, San Francisco, CA (US); **Simon Tong**, Mountain View, CA (US)

Correspondence Address:
**HARRITY SNYDER, LLP**
**11350 Random Hills Road**
**SUITE 600**
**FAIRFAX, VA 22030 (US)**

(21) Appl. No.: **11/169,285**

(22) Filed: **Jun. 29, 2005**

**Publication Classification**

(51) **Int. Cl.**
*G06F 17/30* (2006.01)

(52) **U.S. Cl.** .................................................................. **707/3**

(57) **ABSTRACT**

A system receives a search query from a user and searches a group of repositories, based on the search query, to identify, for each of the repositories, a set of search results. The system also identifies one of the repositories based on a likelihood that the user desires information from the identified repository and presents the set of search results associated with the identified repository.

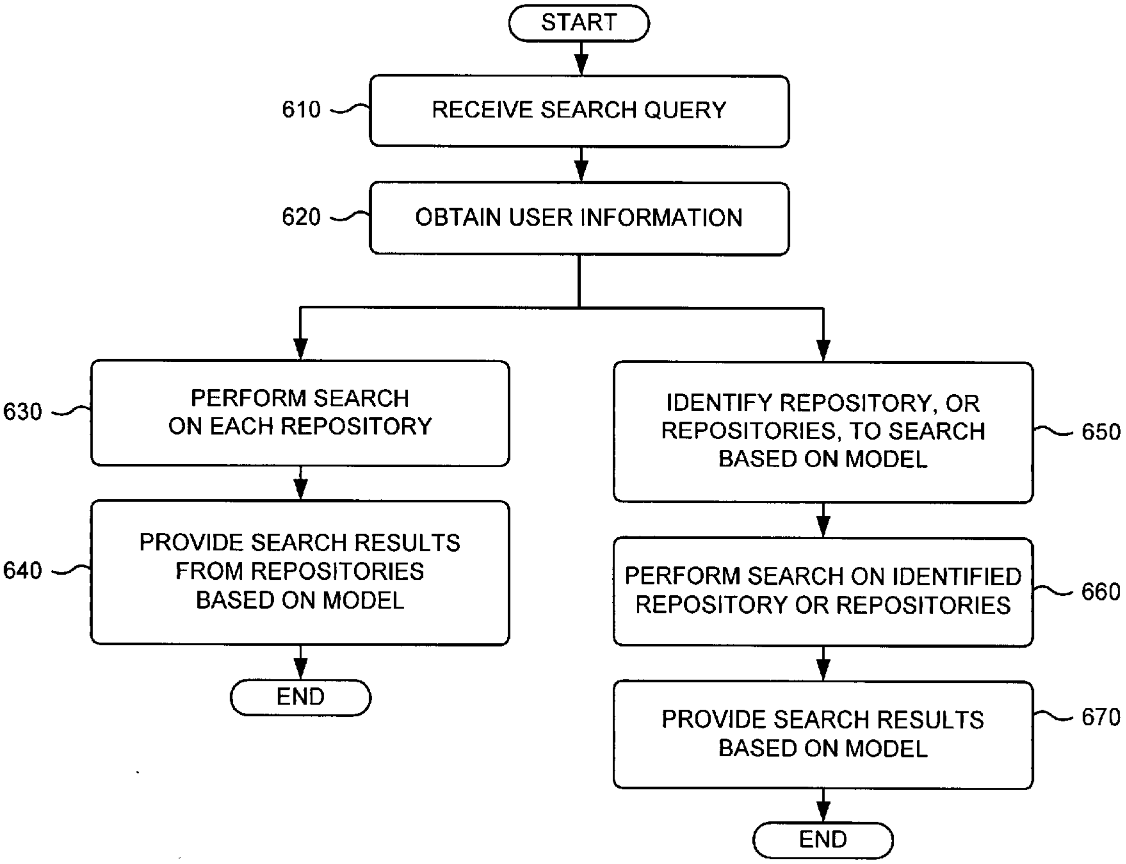
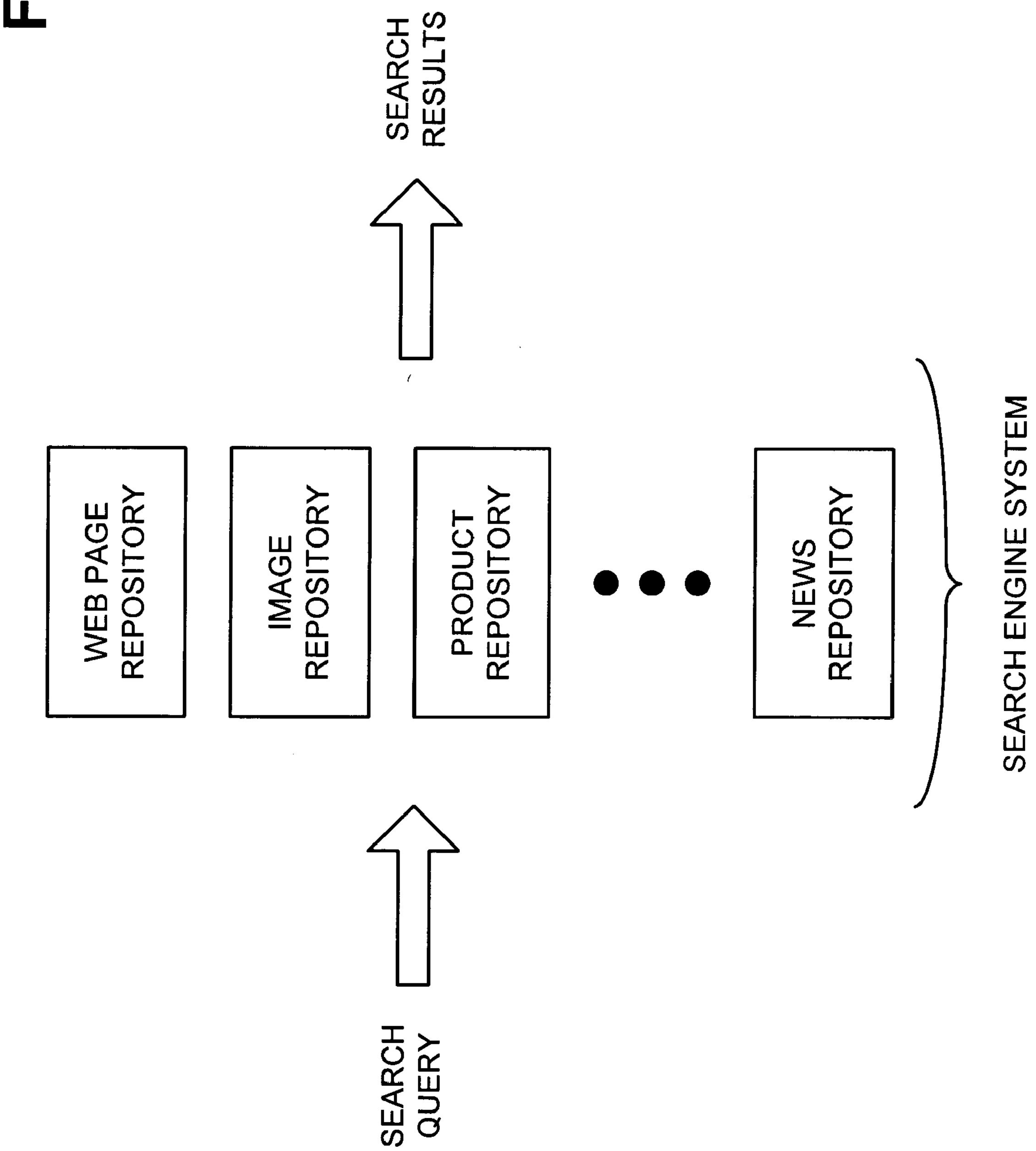

START
610
RECEIVE SEARCH QUERY
620
OBTAIN USER INFORMATION
630
PERFORM SEARCH ON EACH REPOSITORY
640
PROVIDE SEARCH RESULTS FROM REPOSITORIES BASED ON MODEL
END
650
IDENTIFY REPOSITORY, OR REPOSITORIES, TO SEARCH BASED ON MODEL
660
PERFORM SEARCH ON IDENTIFIED REPOSITORY OR REPOSITORIES
670
PROVIDE SEARCH RESULTS BASED ON MODEL
END

FIG. 1


SEARCH QUERY
WEB PAGE REPOSITORY
IMAGE REPOSITORY
PRODUCT REPOSITORY
NEWS REPOSITORY
SEARCH RESULTS
SEARCH ENGINE SYSTEM

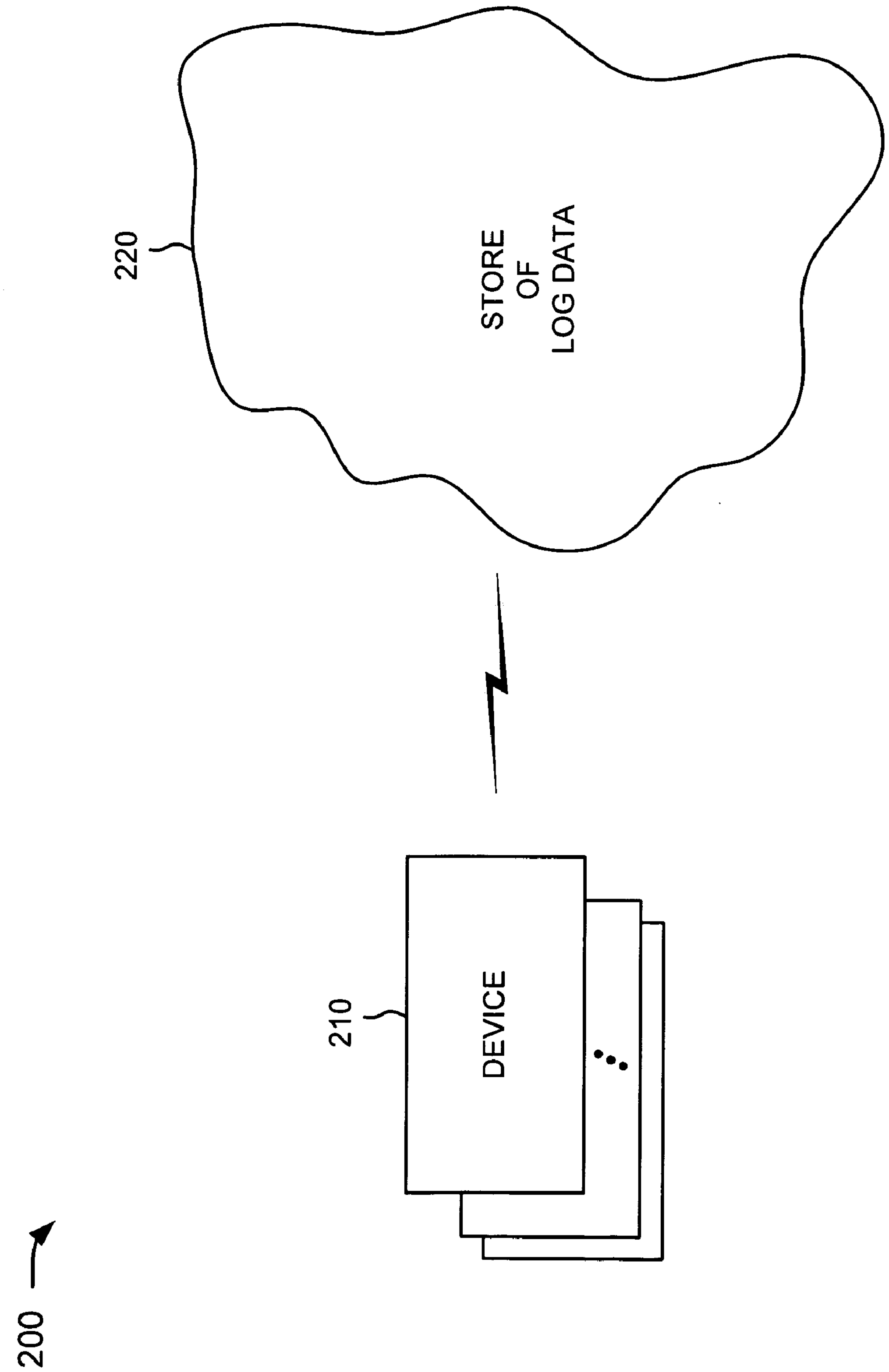


FIG. 2

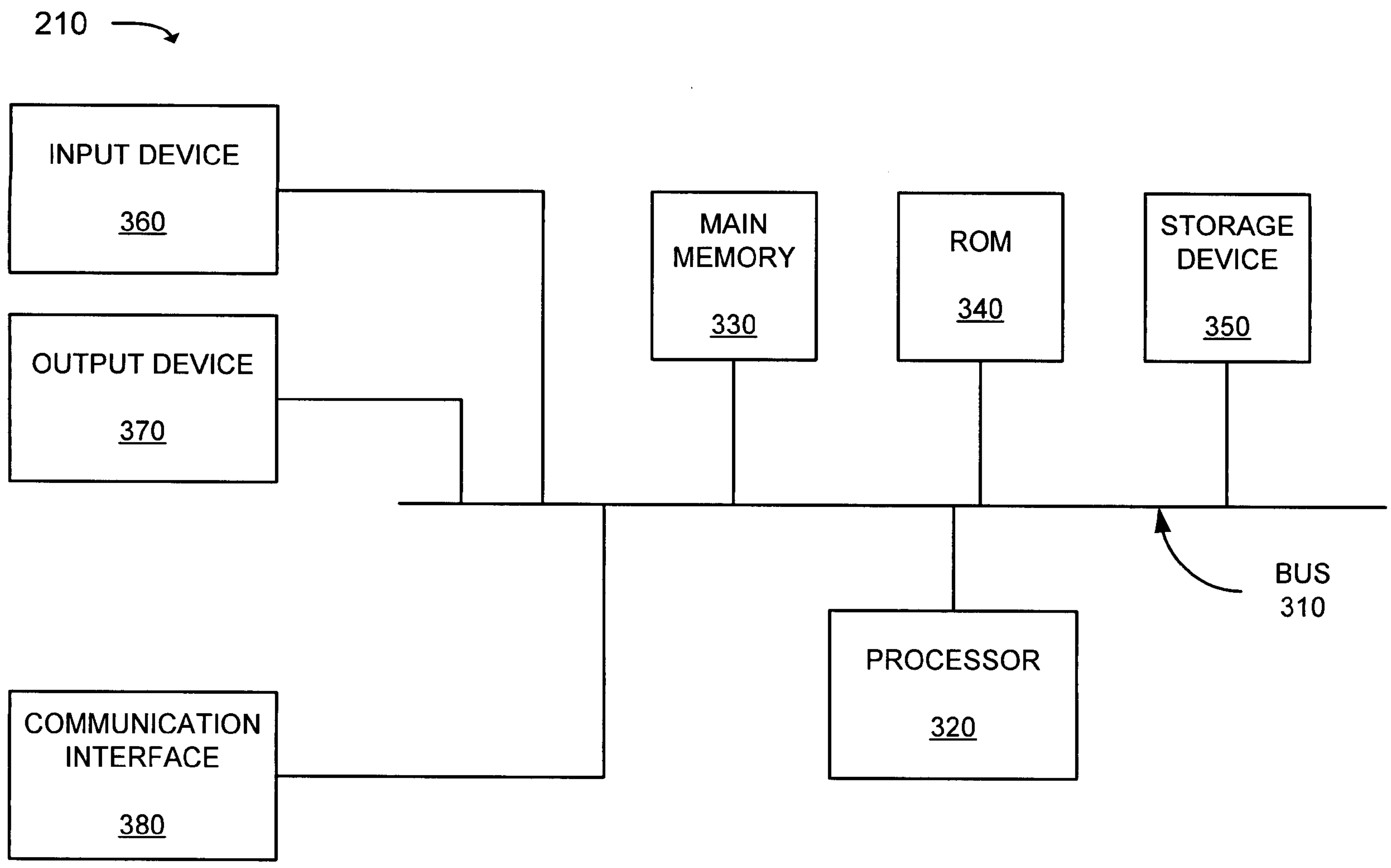
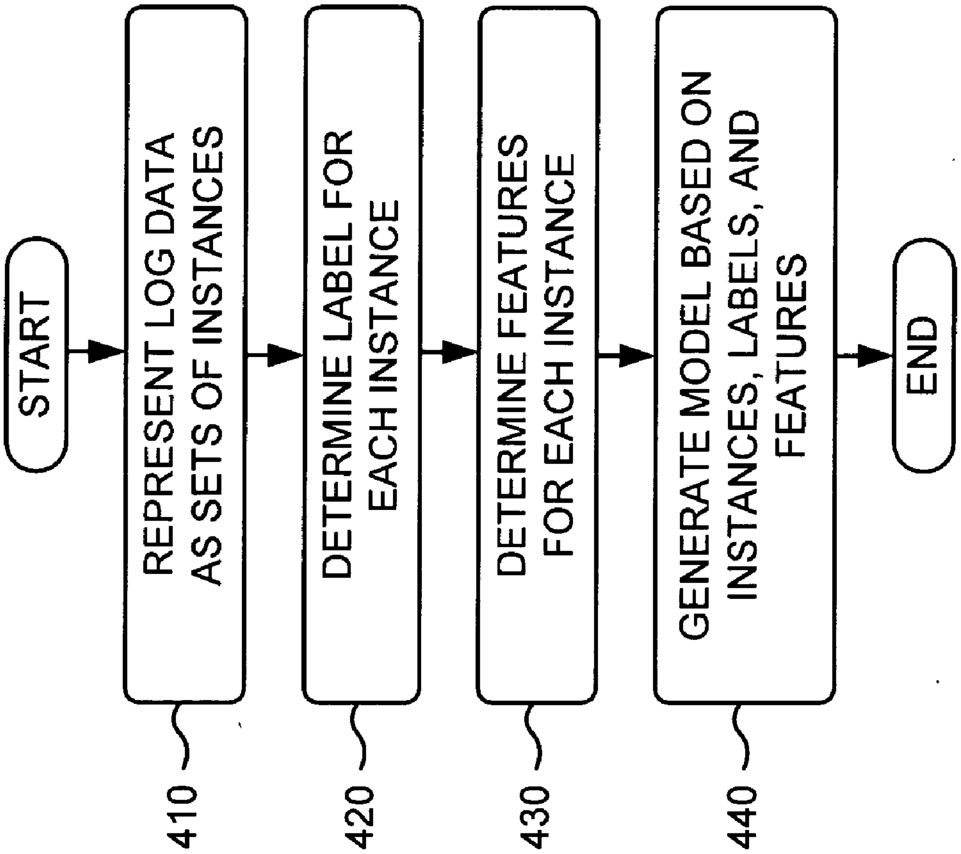

210
INPUT DEVICE
360
OUTPUT DEVICE
370
COMMUNICATION INTERFACE
380
MAIN MEMORY
330
ROM
340
STORAGE DEVICE
350
PROCESSOR
320
BUS
310

FIG. 3

FIG. 4


START
REPRESENT LOG DATA AS SETS OF INSTANCES
410
DETERMINE LABEL FOR EACH INSTANCE
420
DETERMINE FEATURES FOR EACH INSTANCE
430
GENERATE MODEL BASED ON INSTANCES, LABELS, AND FEATURES
440
END

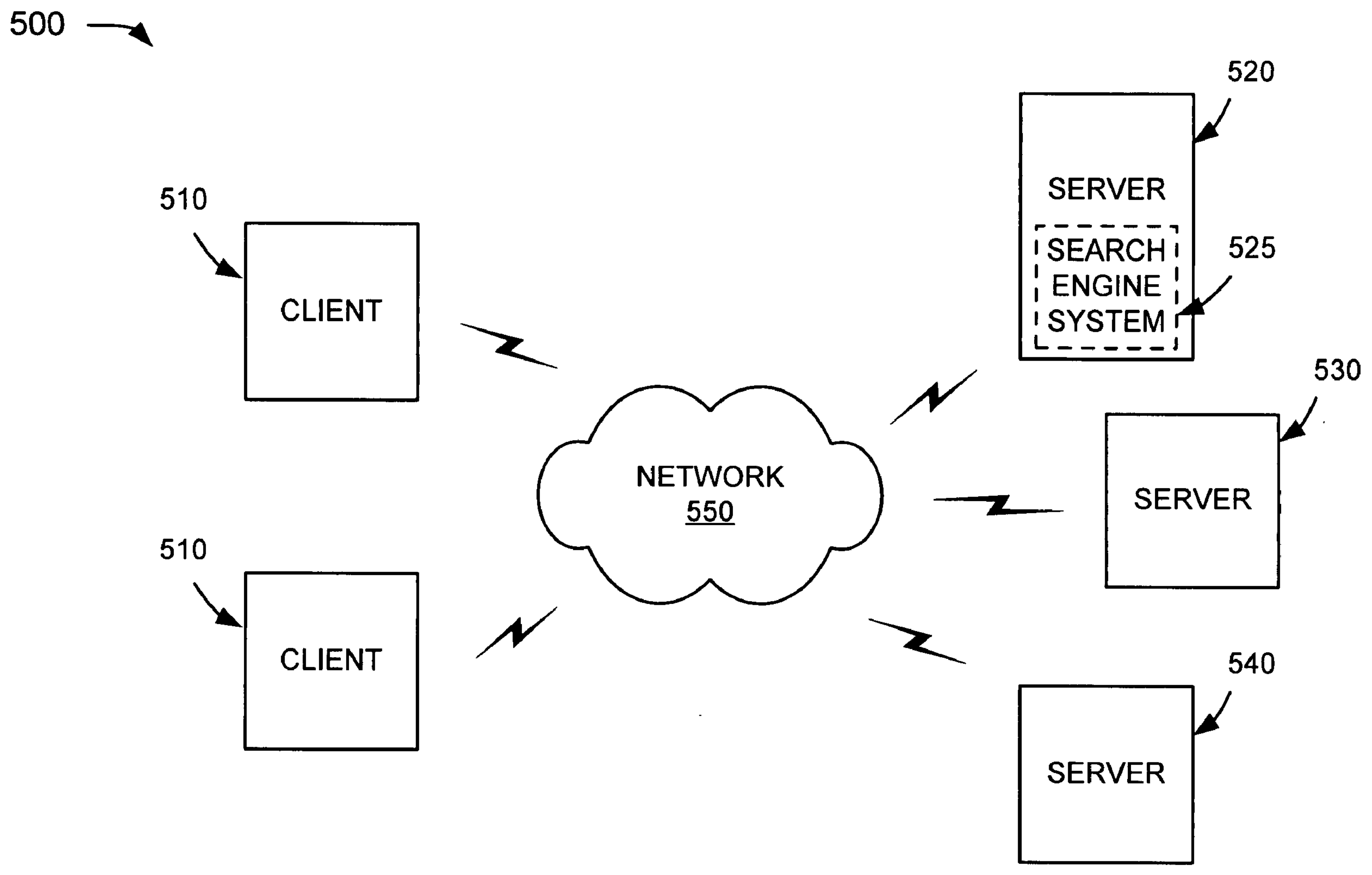
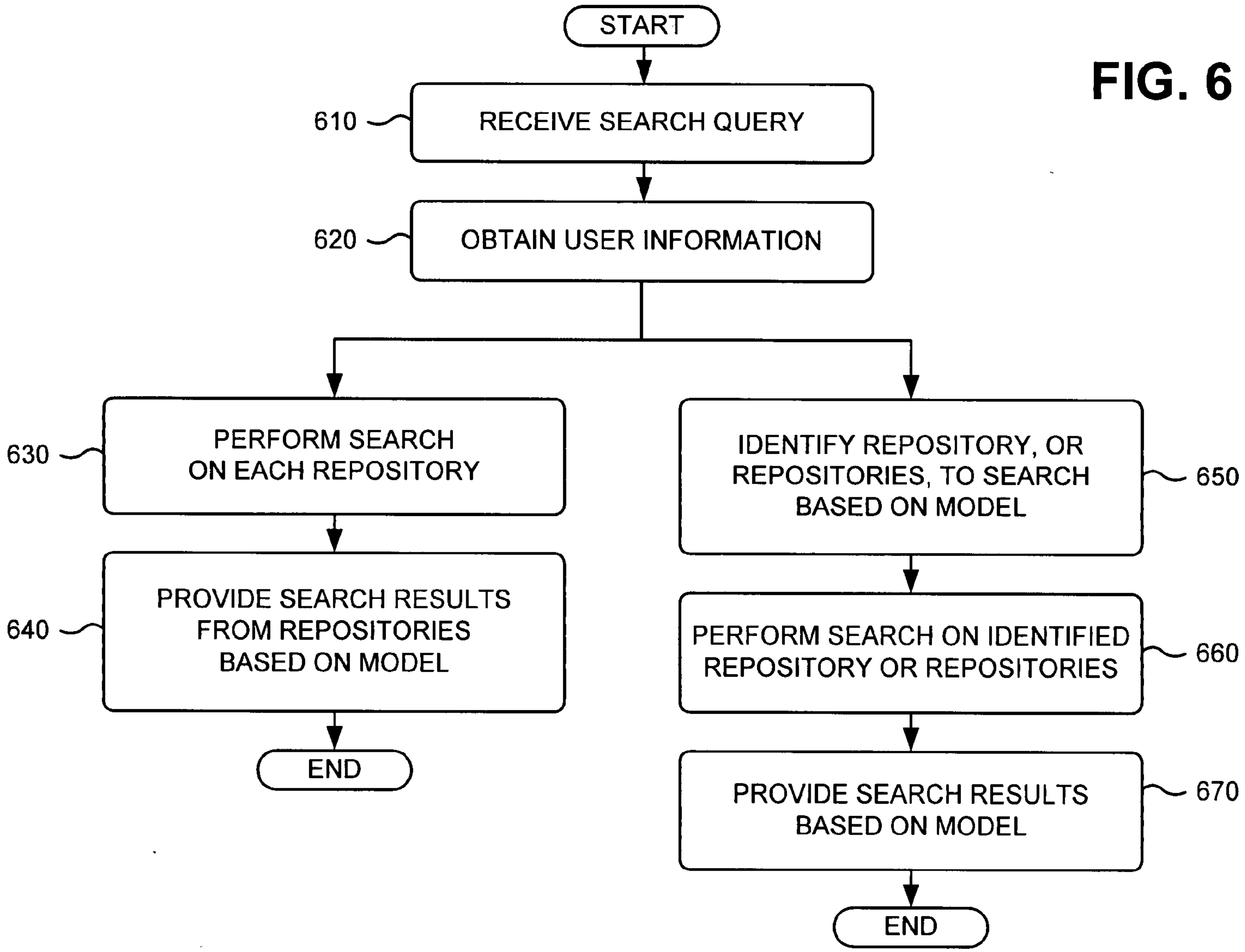

500
510
CLIENT
510
CLIENT
NETWORK
550
520
SERVER
SEARCH
ENGINE
SYSTEM
525
530
SERVER
540
SERVER

FIG. 5

FIG. 6


START
610
RECEIVE SEARCH QUERY
620
OBTAIN USER INFORMATION
630
PERFORM SEARCH ON EACH REPOSITORY
640
PROVIDE SEARCH RESULTS FROM REPOSITORIES BASED ON MODEL
END
IDENTIFY REPOSITORY, OR REPOSITORIES, TO SEARCH BASED ON MODEL
650
PERFORM SEARCH ON IDENTIFIED REPOSITORY OR REPOSITORIES
660
PROVIDE SEARCH RESULTS BASED ON MODEL
670
END

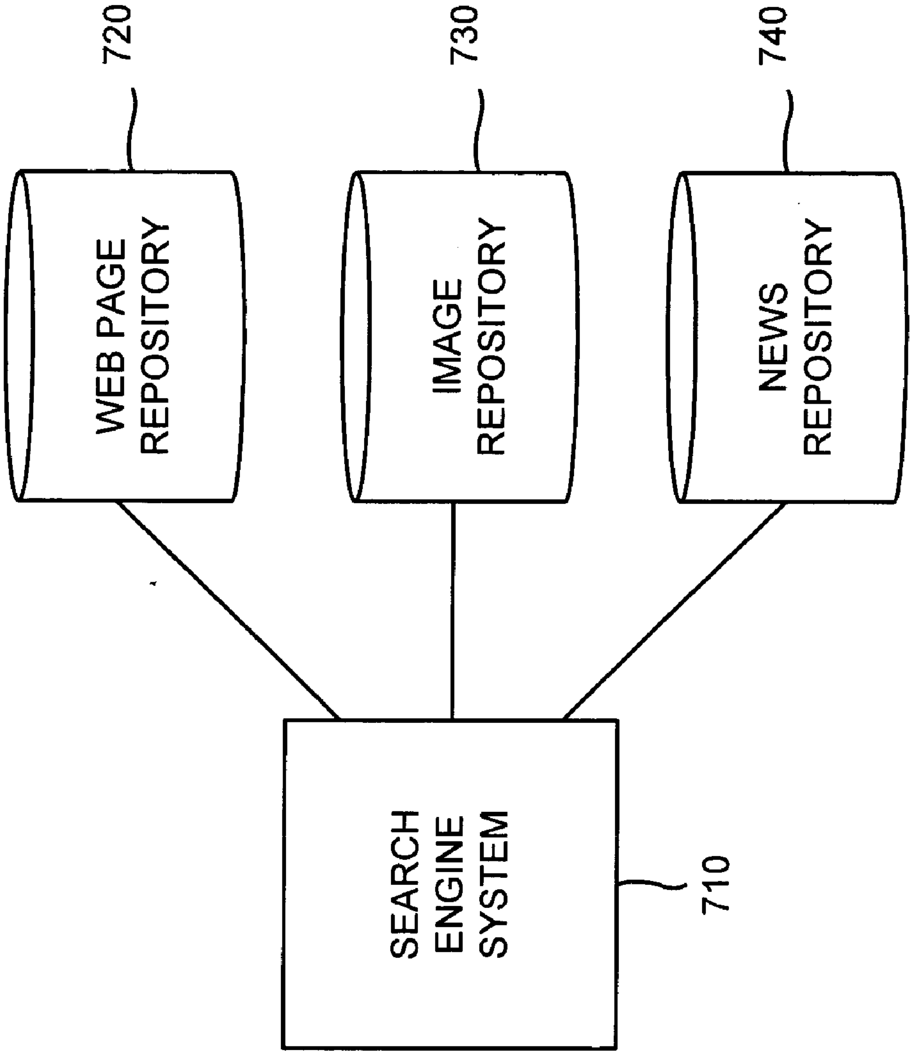
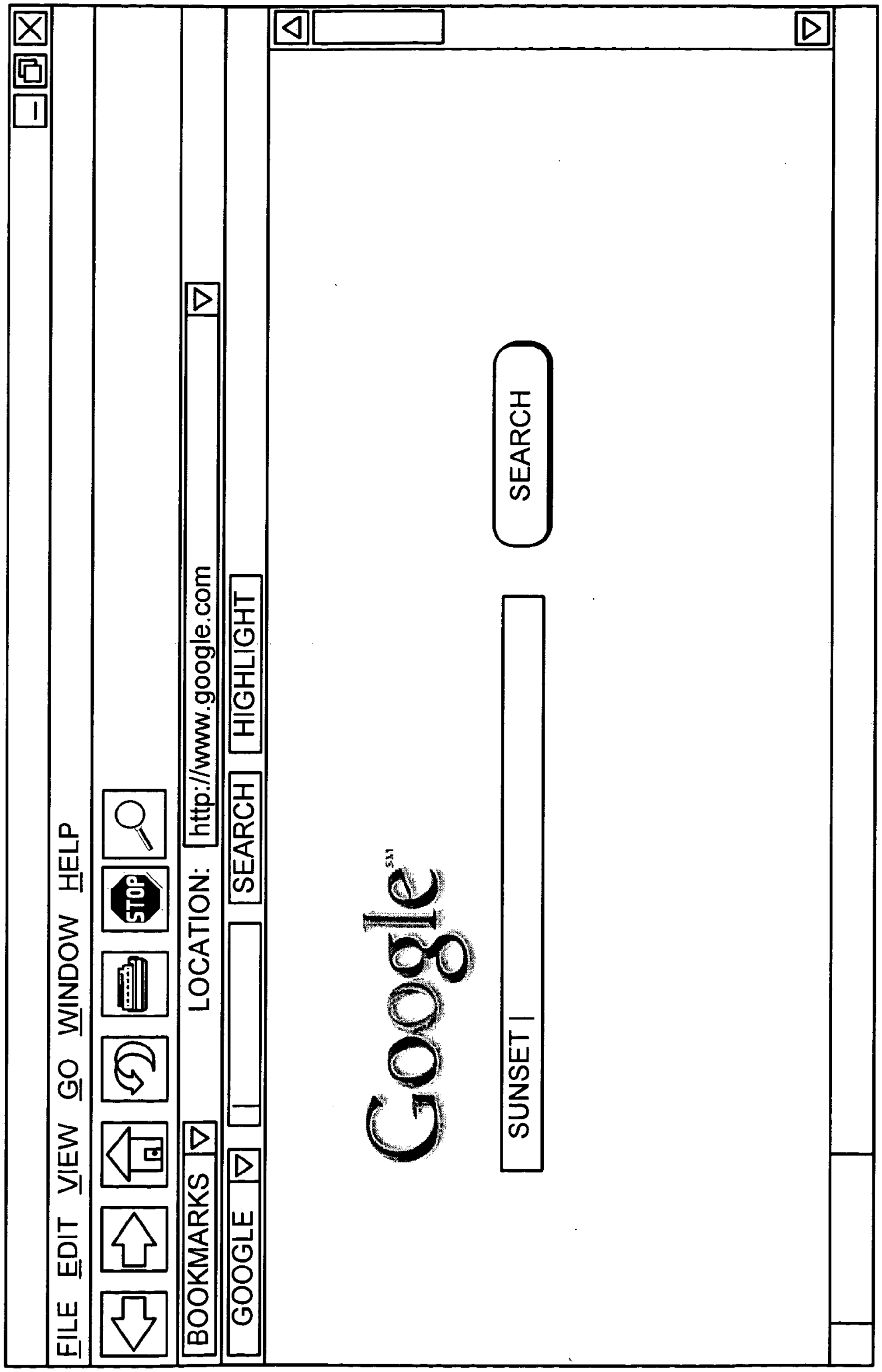

720
WEB PAGE REPOSITORY
730
IMAGE REPOSITORY
740
NEWS REPOSITORY
SEARCH ENGINE SYSTEM
710

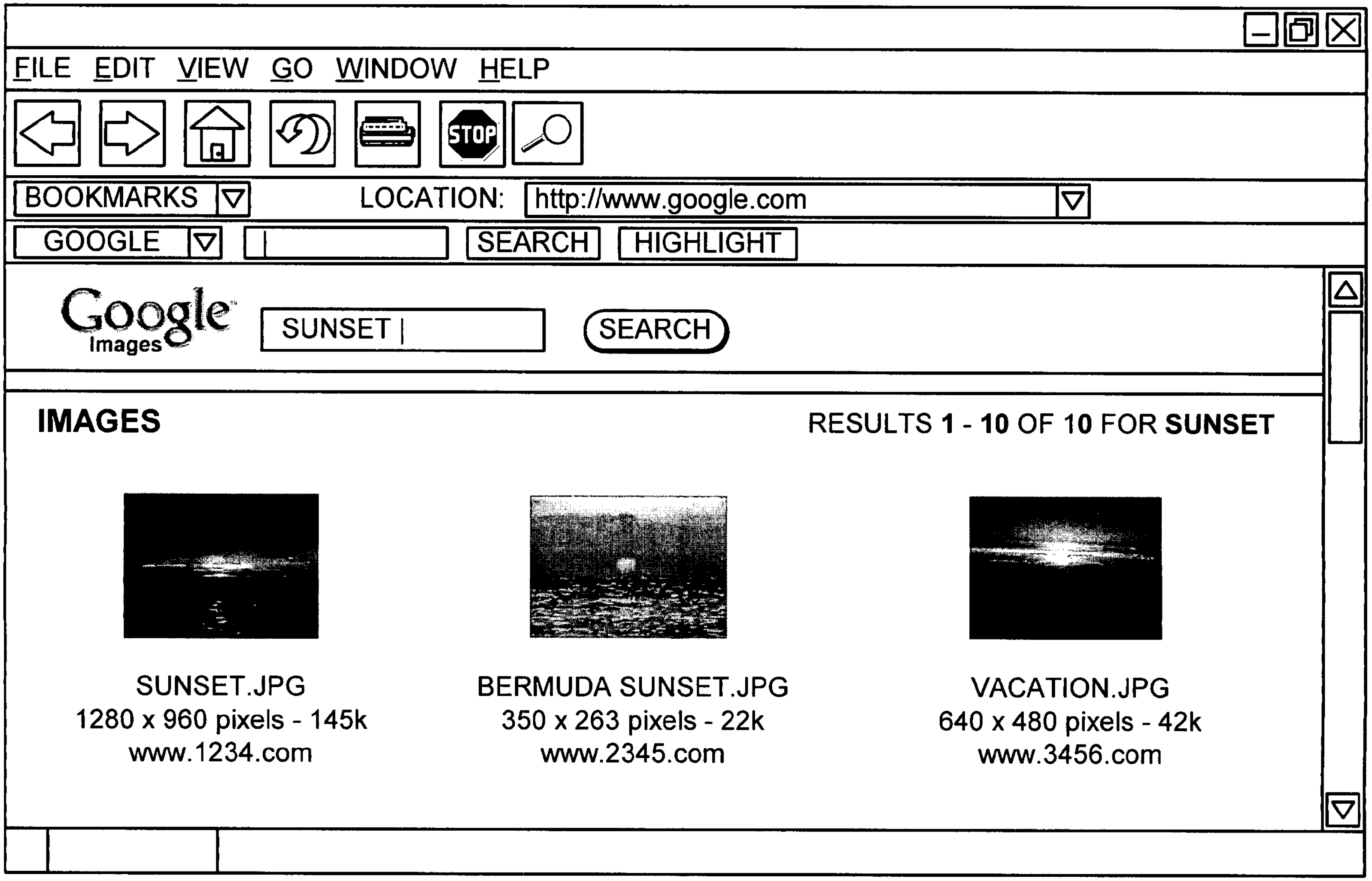
FIG. 7



LOCATION: http://www.google.com

SUNSET |

FIG. 8



LOCATION: http://www.google.com

SUNSET |

IMAGES

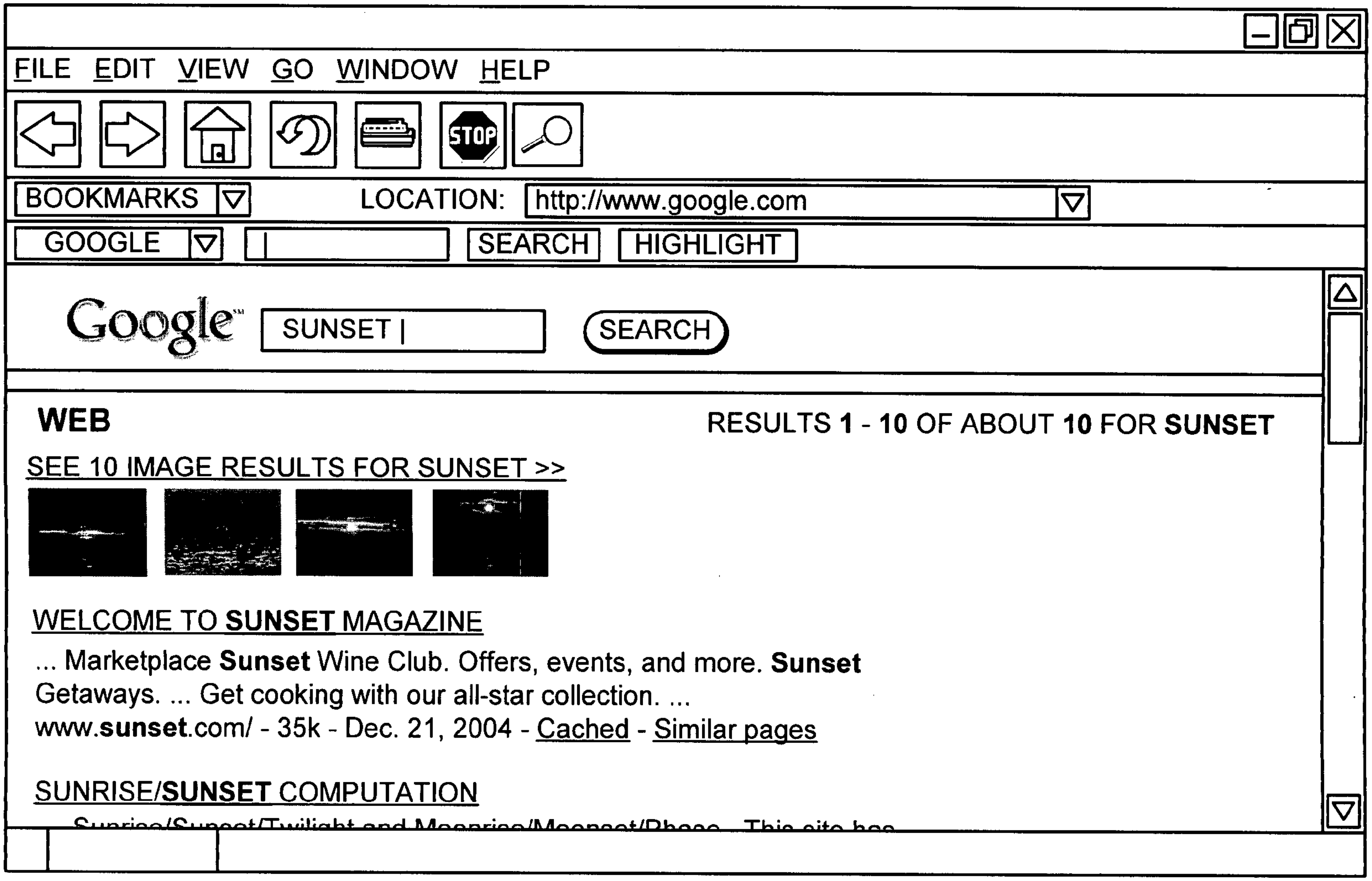
RESULTS **1 - 10** OF **10** FOR **SUNSET**

SUNSET.JPG
1280 x 960 pixels - 145k
www.1234.com

BERMUDA SUNSET.JPG
350 x 263 pixels - 22k
www.2345.com

VACATION.JPG
640 x 480 pixels - 42k
www.3456.com

FIG. 9



http://www.google.com

SUNSET |

**WEB**

RESULTS **1** - **10** OF ABOUT **10** FOR **SUNSET**

SEE 10 IMAGE RESULTS FOR SUNSET >>

WELCOME TO **SUNSET** MAGAZINE

... Marketplace **Sunset** Wine Club. Offers, events, and more. **Sunset** Getaways. ... Get cooking with our all-star collection. ...
www.**sunset**.com/ - 35k - Dec. 21, 2004 - Cached - Similar pages

SUNRISE/**SUNSET** COMPUTATION

Sunrise/Sunset/Twilight and Moonrise/Moonset/Phase - This site has

FIG. 10

# DETERMINATION OF A DESIRED REPOSITORY

## BACKGROUND OF THE INVENTION

[0001] 1. Field of the Invention

[0002] Implementations described herein relate generally to information retrieval and, more particularly, to the determination of a desired repository for a search.

[0003] 2. Description of Related Art

[0004] The World Wide Web ("web") contains a vast amount of information. Locating a desired portion of the information, however, can be challenging. This problem is compounded because the amount of information on the web and the number of new users inexperienced at web searching are growing rapidly.

[0005] Search engine systems attempt to return hyperlinks to web pages in which a user is interested. Generally, search engine systems base their determination of the user's interest on search terms (called a search query) entered by the user. The goal of a search engine system is to provide links to high quality, relevant search results (e.g., web pages) to the user based on the search query. Typically, the search engine system accomplishes this by matching the terms in the search query to a corpus of pre-stored web pages. Web pages that contain the user's search terms are "hits" and are returned to the user as links.

[0006] Some search engine systems can provide various types of information as the search results. For example, a search engine system might be capable of providing search results relating to web pages, news articles, images, merchant products, usenet pages, yellow page entries, scanned books, and/or other types of information. Typically, a search engine system provides separate interfaces to these different types of information.

[0007] When a user provides a search query to a standard search engine system, the user is typically provided with links to web pages. If the user desires another type of information (e.g., images or news articles), the user typically needs to access a separate interface provided by the search engine system.

## SUMMARY OF THE INVENTION

[0008] According to one aspect, a method may include receiving a search query from a user; searching a group of repositories, based on the search query, to identify, for each of the repositories, a set of search results; identifying one of the repositories based on a likelihood that the user desires information from the identified repository; and presenting the set of search results associated with the identified repository.

[0009] According to another aspect, a system may include a search engine system that may receive a search query from a user and determine a score for each of a group of repositories, where the score for one of the repositories is based on a likelihood that the user desires information from the one repository. The search engine system may also perform a search on one or more of the repositories, based on the search query, to identify, for each of the one or more repositories, a set of search results, and provide one or more of the sets of search results based on the scores.

[0010] According to yet another aspect, a computer-readable medium to store data and computer-executable instructions is provided. The computer-readable medium may include log data associated with a number of searches of repositories based on search queries provided by users. The computer-readable medium may also include instructions for representing the log data as triples of data (u, q, r), where u refers to information regarding a user that provided a search query, q refers to information regarding the search query, and r refers to information regarding a repository from which search results were provided in response to the search query; instructions for determining a label for each of the triples of data (u, q, r), where the label includes information regarding whether the user u desired information from the repository r when the user provided the search query q; and instructions for training a model based on the triples of data (u, q, r) and the associated labels, where the model predicts whether a particular user desires information from a repository when the user provides a particular search query.

[0011] According to a further aspect, a system may include a first repository to store a first type of data, a second repository to store a second type of data, and a search engine system. The search engine system may receive a search query from a user, and determine a likelihood that the user desires information from the first or second repository based on information regarding the user, the search query, and the first or second repository.

[0012] According to another aspect, a system may include a model generation system and a search engine system. The model generation system may generate a model that determines a score associated with a likelihood that a particular user desires information from a repository when the user provides a particular search query. The search engine system may receive a search query from a user, determine a score for each of a plurality of repositories based on the model, and present search results from one or more of the repositories based on the scores.

[0013] According to yet another aspect, a method may include receiving a search query from a user; determining a score for each of a plurality of repositories, the score for one of the repositories being based on a likelihood that the user desires information from the one repository; performing a search on at least one of the repositories, based on the search query and the determined scores, to identify, for each of the at least one of the repositories, a set of search results; and providing one or more of the sets of search results.

[0014] According to a further aspect, a system may include a model generation system to generate first and second models, where at least one factor used to generate the second model is different or absent when generating the first model. The system may also include a search engine system to receive a search query from a user, determine a first score for each of a plurality of repositories based on the first model, perform a search on one or more of the repositories based on the search query and the first scores, determine a second score for each of the one or more of the repositories based on the second model, and present search results from at least one of the one or more of the repositories based on the second scores.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0015]** The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate an embodiment of the invention and, together with the description, explain the invention. In the drawings,

**[0016]** FIG. **1** illustrates a concept consistent with principles of the invention;

**[0017]** FIG. **2** is a diagram of an exemplary model generation system according to an implementation consistent with the principles of the invention;

**[0018]** FIG. **3** is an exemplary diagram of a device of FIG. **2** according to an implementation consistent with the principles of the invention;

**[0019]** FIG. **4** is a flowchart of exemplary processing for generating a model according to an implementation consistent with the principles of the invention;

**[0020]** FIG. **5** is a diagram of an exemplary information retrieval network in which systems and methods consistent with the principles of the invention may be implemented;

**[0021]** FIG. **6** is a flowchart of exemplary processing for providing search results according to an implementation consistent with the principles of the invention; and

**[0022]** FIGS. **7-10** are diagrams of exemplary implementations consistent with the principles of the invention.

## DETAILED DESCRIPTION

**[0023]** The following detailed description of the invention refers to the accompanying drawings. The same reference numbers in different drawings may identify the same or similar elements. Also, the following detailed description does not limit the invention.

### Overview

**[0024]** FIG. **1** illustrates a concept consistent with principles of the invention. A search engine system may maintain different types of information that might be desired by a user. The search engine system may maintain a set of repositories relating to the different types of information. As shown in FIG. **1**, the search engine system may be associated with, for example, repositories relating to web pages, images, products, and news. The web page repository may include information relating to web pages. The image repository may include information relating to images. The product repository may include information relating to merchant products. The news repository may include information relating to news documents. The search engine system may provide separate interfaces for searches directed to specific ones of the repositories.

**[0025]** In the description to follow, the term "document" is to be broadly interpreted to include any machine-readable and machine-storable work product. A document may include, for example, a web page, information relating to a news event, an image file, information relating to a merchant product, information relating to a usenet page, a yellow page entry, a scanned book, a file, a combination of files, one or more files with embedded links to other files, a blog, a web advertisement, an e-mail, etc. Documents often include textual information and may include embedded information (such as meta information, hyperlinks, etc.) and/or embedded instructions (such as Javascript, etc.). A "link," as the term is used herein, is to be broadly interpreted to include any reference to/from a document from/to another document or another part of the same document.

**[0026]** As shown in FIG. **1**, a user may provide a search query to the search engine system. The search engine system may determine which repository or repositories the user likely desires. The search engine may perform a search and present search results that include information from one or more of the repositories based on the determination of which repository or repositories the user likely desires.

**[0027]** For example, if a user provides the term "sunset" as a search query to the search engine system, the search engine system may determine that the user is more interested in images of sunsets rather than web pages relating to sunsets. As a result, the search engine system may present the user with search results from the image repository instead of, or in addition to, search results from other repositories.

**[0028]** Similarly, if a user provides the phrase "iraq war" as a search query to the search engine system, the search engine system may determine that the user is more interested in news documents relating to the Iraq war rather than web pages relating to the Iraq war. As a result, the search engine system may present the user with search results from the news repository instead of, or in addition to, search results from other repositories.

**[0029]** Implementations consistent with the principles of the invention may generate a model that predicts which repository, or repositories, a user is interested in when the user provides a search query, and use this model to provide relevant search results to the user.

### Exemplary Model Generation System

**[0030]** FIG. **2** is an exemplary diagram of a model generation system **200** consistent with the principles of the invention. System **200** may include one or more devices **210** and a store of log data **220**. Store **220** may include one or more logical or physical memory devices that may store a large data set (e.g., millions of instances and hundreds of thousands of features) that may be used, as described in more detail below, to create and train a model. The data may include log data concerning prior searches, such as user information, query information, and repository information, that may be used to create a model that may be used to identify one or more repositories that may be desired by a user. In one implementation, the model may predict whether a user desires information from a particular repository when the user provides a certain query.

**[0031]** The user information may include Internet Protocol (IP) addresses, cookie information, languages, and/or geographical information associated with the users, prior queries provided by the users, and/or the time of day and/or day of the week that the users provided the current or prior queries. The query information may include information relating to the query terms that were provided. The repository information may include information relating to the repository interfaces used for the searches, the documents that were displayed and the repositories from which they were obtained, and/or the documents that were selected (e.g., clicked on). In other exemplary implementations, other types of data may alternatively or additionally be maintained by store **220**.

[0032] Device(s) **210** may include any type of computing device capable of accessing store **220** via any type of connection mechanism. According to one implementation consistent with the principles of the invention, system **200** may include multiple devices **210**. According to another implementation, system **200** may include a single device **210**.

[0033] FIG. **3** is an exemplary diagram of a device **210** according to an implementation consistent with the principles of the invention. Device **210** may include a bus **310**, a processor **320**, a main memory **330**, a read only memory (ROM) **340**, a storage device **350**, an input device **360**, an output device **370**, and a communication interface **380**. Bus **310** may include a path that permits communication among the elements of device **210**.

[0034] Processor **320** may include a processor, microprocessor, or processing logic that may interpret and execute instructions. Main memory **330** may include a random access memory (RAM) or another type of dynamic storage device that may store information and instructions for execution by processor **320**. ROM **340** may include a ROM device or another type of static storage device that may store static information and instructions for use by processor **320**. Storage device **350** may include a magnetic and/or optical recording medium and its corresponding drive.

[0035] Input device **360** may include a mechanism that permits an operator to input information to device **210**, such as a keyboard, a mouse, a pen, voice recognition and/or biometric mechanisms, etc. Output device **370** may include a mechanism that outputs information to the operator, including a display, a printer, a speaker, etc. Communication interface **380** may include any transceiver-like mechanism that enables device **210** to communicate with other devices and/or systems. For example, communication interface **380** may include mechanisms for communicating with another device **210** or store **220**.

[0036] As will be described in detail below, device **210**, consistent with the principles of the invention, may perform certain model generating-related operations. Device **210** may perform these operations in response to processor **320** executing software instructions contained in a computer-readable medium, such as memory **330**. A computer-readable medium may be defined as a physical or logical memory device and/or carrier wave.

[0037] The software instructions may be read into memory **330** from another computer-readable medium, such as data storage device **350**, or from another device via communication interface **380**. The software instructions contained in memory **330** may cause processor **320** to perform processes that will be described later. Alternatively, hardwired circuitry may be used in place of or in combination with software instructions to implement processes consistent with the principles of the invention. Thus, implementations consistent with the principles of the invention are not limited to any specific combination of hardware circuitry and software.

## Exemplary Model Geration Processing

[0038] For purposes of the discussion to follow, the set of data in store **220** (FIG. **2**) may include multiple elements, called instances. It may be possible for store **220** to include millions of instances. Each instance may include a triple of data: (u, q, r), where "u" refers to user information, "q" refers to the query that user u provided, and "r" refers to the repository from which search results were provided in response to query q. Store **220** may also store information regarding whether user u desired information from repository r when user u provided query q, where the user's desire may be measured, for example, by determining whether the user selected a document from the repository. This information will be referred to as the "label" for the instance.

[0039] Several features may be extracted for any given (u, q, r). It may be possible for store **220** to include hundreds of thousands of distinct features. In one implementation, some of these features might include one or more of the following: the country in which user u is located, the language of the country in which user u is located, a cookie identifier associated with user u, the language of query q, each term in query q, the time of day user u provided query q, the documents from repository r that were presented to user u, each of the terms in the documents from repository r that were presented to user u, and/or each of the terms in the titles of the documents from repository r that were presented to the user u. Other features might alternatively or additionally be used.

[0040] In another implementation, some of the features might include one or more of the following in addition to, or instead of, some of the features identified above: the fraction of queries that were provided to the interface for repository r, the fraction of queries that were provided to the interface for repository r versus the interfaces for other repositories, the fraction of queries that contain a term in query q that were provided to the interface for repository r versus the interfaces for other repositories, the overall click rate for queries provided to the interface for repository r, the click rate for queries provided to the interface for repository r for user u, the click rate for queries provided to the interface of repository r for users in the same country as user u, and/or the click rate for query q provided to the interface of repository r.

[0041] In a further implementation, the following two features might also be included: the click rate of query q provided to the interface of repository r for user u, and the fraction of queries q that were provided to the interface of repository r for user u. Instead of determining these features directly, models might be generated to predict these features using conventional techniques and the output of the models may be used as features.

[0042] A model may be created based on this data. In one implementation, the model may be used to predict, given a new (u, q, r), whether user u desires information from repository r if user u provided query q. As will be described in more detail below, the output of the model may be used to determine whether to search a repository, whether to include search results from a repository in a search result document, and/or the manner for presenting search results within the search result document.

[0043] FIG. **4** is a flowchart of exemplary processing for generating a model according to an implementation consistent with the principles of the invention. This processing may be performed by a single device **210** or a combination of multiple devices **210**.

[0044] To facilitate generation of the model, the log data in store **220** may be represented as sets of instances (block

**410**). For example, information may be identified relating to prior searches by users, such as information regarding the users, the queries the users provided, and the repositories from which the search results were obtained and/or selected. This information may be formed into triples (u, q, r), as described above.

[**0045**] A label for each instance may then be determined (block **420**). For example, it may be determined for each triple (u, q, r) whether user u desired information (e.g., selected a document) in repository r when user u provided query q. The labels may be associated with their corresponding instances in store **220**. The features relating to each of the instances may also be determined (block **430**).

[**0046**] A model may then be generated based on the instances, labels, and features (block **440**). For example, a standard machine learning or statistical technique may be used to determine the probability that user u desires information from repository r when user u provides query q:

P(desire | u, q, show_r),

where "show_r" indicates that documents from repository r are provided. Any of several well known techniques may be used to generate the model, such as logic regression, boosted decision trees, random forests, support vector machines, perceptrons, and winnow learners. Instead of generating a probability, the model may output a value that reflects a confidence that user u desires information from repository r when user u provides query q. The output of the model will be generally referred to hereinafter as a "score," which may include a probability output and/or an output value.

[**0047**] As explained below, the output of the model may be used to determine whether to search a repository, whether to include search results from a repository in a search result document, and/or the manner for presenting search results within the search result document.

### Exemplary Information Retrieval Network

[**0048**] FIG. **5** is an exemplary diagram of a network **500** in which systems and methods consistent with the principles of the invention may be implemented. Network **500** may include multiple clients **510** connected to multiple servers **520**-**540** via a network **550**. Two clients **510** and three servers **520**-**540** have been illustrated as connected to network **550** for simplicity. In practice, there may be more or fewer clients and servers. Also, in some instances, a client may perform a function of a server and a server may perform a function of a client.

[**0049**] Clients **510** may include client entities. An entity may be defined as a device, such as a personal computer, a wireless telephone, a personal digital assistant (PDA), a lap top, or another type of computation or communication device, a thread or process running on one of these devices, and/or an object executable by one of these devices. Servers **520**-**540** may include server entities that gather, process, search, and/or maintain documents in a manner consistent with the principles of the invention.

[**0050**] In an implementation consistent with the principles of the invention, server **520** may include a search engine system **525** usable by clients **510**. Search engine system **525** may be associated with a number of repositories of documents (not shown), such as a web page repository, a news repository, an image repository, a products repository, a usenet repository, a yellow pages repository, a scanned books repository, and/or other types of repositories. These repositories may physically reside in one or more memory devices located within server **520** or external to server **520**. Servers **530** and **540** may store or maintain documents that may be associated with one or more of the repositories.

[**0051**] While servers **520**-**540** are shown as separate entities, it may be possible for one or more of servers **520**-**540** to perform one or more of the functions of another one or more of servers **520**-**540**. For example, it may be possible that two or more of servers **520**-**540** are implemented as a single server. It may also be possible for a single one of servers **520**-**540** to be implemented as two or more separate (and possibly distributed) devices.

[**0052**] Network **550** may include a local area network (LAN), a wide area network (WAN), a telephone network, such as the Public Switched Telephone Network (PSTN), an intranet, the Internet, or a combination of networks. Clients **510** and servers **520**-**540** may connect to network **550** via wired, wireless, and/or optical connections.

### Exemplary Process for Providing Search Results

[**0053**] FIG. **6** is a flowchart of exemplary processing for providing search results according to an implementation consistent with the principles of the invention. Processing may begin with the receipt of a search query (block **610**). For example, a user may access a search engine interface using web browser software on a client, such as client **510** (FIG. **5**). The user may provide the search query to the search engine interface.

[**0054**] Information concerning the user may be obtained (block **620**). For example, the user may be identified using, for example, an IP address, cookie information, languages, and/or geographical information associated with the user. Conventional techniques may be used for gathering the user information.

[**0055**] In one implementation, a search may be performed on each of the repositories based on the search query (block **630**). A set of search results may be obtained corresponding to each of the repositories. Any information retrieval technique may be used to identify relevant documents to include in the set of search results.

[**0056**] It may then be determined how the search results will be provided based on the model (block **640**). For example, information relating to the user, the search query the user provided, and each of the repositories may be used as inputs to the model. The model may be applied to each repository and the output of the model ("score") may be used to determine whether to provide search results associated with that repository. It may be determined, for example, that search results from the two repositories with the highest associated score should be provided. Alternatively, it may be determined that search results from a particular one of the repositories should always be provided and search results from another one or more repositories should also be provided if the score associated with the other one or more repositories is greater than the score associated with the particular repository. Alternatively, it may be determined that search results from repositories with associated scores above a certain threshold should be provided, and if none of the scores is above the threshold, then provide search results from the repository with the highest associated score. Yet other rules for determining whether to provide search results associated with a repository may alternatively or additionally be used.

**[0057]** The output of the model may alternatively, or additionally, be used to determine the manner in which the search results from the different repositories are provided. For example, it may be determined that if the score associated with a repository is below some threshold, the search results associated with the repository may be presented toward the bottom of the search result document presented to the user rather than toward the top of the search result document. Alternatively, or additionally, it may be determined that if the score associated with a repository is below some threshold, a link to the search results associated with the repository is presented instead of the search results themselves. Yet other rules for determining the manner for providing search results associated with a repository may alternatively or additionally be used.

**[0058]** The search results may then be arranged within a search result document and presented to the user. Each search result may include, for example, a link to a document from the corresponding repository and possibly a brief description of or excerpt from the document.

**[0059]** In another implementation, the repository, or repositories, to search may be identified based on the model (block **650**). For example, information relating to the user, the search query the user provided, and each of the repositories may be used as inputs to the model. The model may be applied to each repository and the output of the model ("score") may be used to determine which repository to search. It may be determined, for example, that the two repositories with the highest associated score should be searched. Alternatively, it may be determined that a particular one of the repositories should always be searched and another one or more repositories should also be searched if the score associated with the other one or more repositories is greater than the score associated with the particular repository. Alternatively, it may be determined that repositories with associated scores above a certain threshold should be searched, and if none of the scores is above the threshold, then search the repository with the highest associated score. Yet other rules for determining which repository to search may alternatively or additionally be used.

**[0060]** A search may be performed to obtain a set of search results from each of the identified repositories (block **660**). Any conventional information retrieval technique may be used to identify relevant documents to include in the set of search results.

**[0061]** The search results may then be provided based on the model (block **670**). For example, the output of the model may be used to determine the manner in which the search results from different repositories are provided. For example, it may be determined that if the score associated with a repository is below some threshold, the search results associated with the repository may be presented toward the bottom of the search result document presented to the user rather than toward the top of the search result document. Alternatively, or additionally, it may be determined that if the score associated with a repository is below some threshold, a link to the search results associated with the repository is presented instead of the search results themselves. Other rules for determining the manner for providing search results associated with a repository may alternatively or additionally be used.

**[0062]** The search results may then be arranged within a search result document and presented to the user. Each search result may include, for example, a link to a document from the corresponding repository and possibly a brief description of or excerpt from the document.

**[0063]** In another implementation, two or more models may be used. For example, a first model may be used to determine whether to search a repository; a second model may be used to determine whether to include search results from one of the searched repositories in a search result document; and the second model, or possibly a third model, may be used to determine the manner for presenting search results within the search result document. The first, second, and/or third models may be generated based on one or more factors that differ from each other. For example, in one implementation, the output of the first model may be used as an input to the second model and/or the output of the first and/or second model may be used as an input to the third model.

**[0064]** It may be possible to provide information concerning this search as log data to store **220**. For example, the information may be used as training data for training or refining the model.

## EXAMPLE

**[0065]** FIGS. **7-10** are diagrams of exemplary implementations consistent with the principles of the invention. As shown in FIG. **7**, assume that a search engine system **710** has three associated repositories, including web page repository **720**, image repository **730**, and news repository **740**. Web page repository **720** may store information relating to web pages. Image repository **730** may store information relating to images. News repository **740** may store information relating to news documents. Search engine system **710** may receive a search query from a user and provide relevant search results from one or more of repositories **720-740**.

**[0066]** As shown in FIG. **8**, assume that a user accesses an interface associated with search engine system **710**. The interface may be associated with one of the repositories or none of the repositories. As shown in FIG. **8**, assume that the user provides the search query "sunset" to search engine system **710**. In addition to the search query, search engine system **710** may obtain information regarding the user, such as an IP address, cookie information, languages, and/or geographical information associated with the user.

**[0067]** In one implementation, as described above, search engine system **710** may perform a search on each of repositories **720-740** to obtain a set of search results for each of repositories **720-740**. Assume that search engine system **710** identifies **10** web page results from web page repository **720**, **10** image results from image repository **730**, and 10 news document results from news repository **740** as relevant search results for the search query "sunset."

**[0068]** Search engine system **710** may input information relating to the user, the search query the user provided, and each of repositories **720-740** as inputs to the model. The model may be used to determine the probability of the user desiring information from each of repositories **720-740** when the user provides the search query "sunset."

[**0069**] Assume, for example, that the following outputs are generated by the model:

*P*(desire |*u, q,* show_web page repository)=0.45

*P*(desire |*u, q,* show_image repository)=0.91

*P*(desire |*u, q,* show_news repository)=0.23,

[**0070**] where "u" refers to user information corresponding to the user that provided the search query, "q" refers to information corresponding to the search query the user provided (i.e., "sunset"), and "show_x repository" (where x corresponds to "web page,""image," or "news") refers to information corresponding to the identified repository. In this case, the probability of the user desiring information from web page repository **720** when the user provides the search query "sunset" is 45%; the probability of the user desiring information from image repository **730** when the user provides the search query "sunset" is 91%; and the probability of the user desiring information from news repository **740** when the user provides the search query "sunset" is 23%.

[**0071**] Search engine system **710** may then use the output of the model with regard to each of repositories **720-740** to determine whether to provide search results associated with that repository. For example, assume that a rule indicates that search engine system **710** is to provide search results only from the repository with the highest score. In this case, search engine system **710** may form a search result document based on the **10** image results identified from image repository **730** (i.e., the repository with the highest score —0.91), as shown in FIG. **9**.

[**0072**] Alternatively, assume that a rule indicates that search engine system **710** is to always provide search results from web page repository **720** and, if another repository has an associated score higher than the score associated with web page repository **720**, provide search results from that repository (or repositories). In this case, search engine system **710** may determine that it is to provide search results from both web page repository **720** and image repository **730** because the score associated with image repository **730** (0.91) is greater than the score associated with web page repository **720** (0.45). **100701** Search engine system **710** may then form a search result document based on the 10 web page results from web page repository **720** and the **10** image results from image repository **730**, as shown in FIG. **10**. Because the score associated with image repository **730** is higher than the score associated with web page repository **720** (or some degree higher or higher and greater than a threshold), information regarding the **10** image results may be presented in a more prominent location than the 10 web page results within the search result document, as also shown in FIG. **10**. The user might select the link associated with the **10** image results (e.g., "SEE 10 IMAGE RESULTS FOR SUNSET>>") to be presented with additional information regarding the image results, similar to that shown in FIG. **9**.

## Conclusion

[**0073**] Implementations consistent with the principles of the invention may generate a model that may be used to predict which repository, or repositories, a user is likely interested in when the user provides a search query, and use this model to provide relevant search results to the user.

[**0074**] The foregoing description of preferred embodiments of the invention provides illustration and description, but is not intended to be exhaustive or to limit the invention to the precise form disclosed. Modifications and variations are possible in light of the above teachings or may be acquired from practice of the invention.

[**0075**] For example, while series of acts have been described with regard to FIGS. **4** and **6**, the order of the acts may be modified in other implementations consistent with the principles of the invention. Further, non-dependent acts may be performed in parallel.

[**0076**] Also, exemplary user interfaces have been described with respect to FIGS. **8-10**. In other implementations consistent with the principles of the invention, the user interfaces may include more, fewer, or different pieces of information.

[**0077**] The preceding description refers to a user. A "user" is intended to refer to a client, such as a client **510** (FIG. **5**), or an operator of a client.

[**0078**] Further, it has been described that the output of the model ("score") can be used to determine whether to search a repository, whether to include search results from a repository in a search result document, and/or the manner for presenting search results within the search result document. In another implementation, the score may be used as one input, of multiple inputs, to a function that determines whether to search a repository, whether to include search results from a repository in a search result document, and/or the manner for presenting search results within the search result document.

[**0079**] Further, some of the features described above are more computationally expensive to determine than others. For example, features based on the documents in the repositories may require those repositories to be queried and the documents to be fetched. For computational efficiency, an approximate main model may be created based on less computationally expensive (e.g., cheaper) features and this approximate main model may be used to determine which repositories to search. Once the documents from these repositories have been fetched, the full main model may be used to determine from which repositories to provide search results.

[**0080**] Also, it may be possible to use the model according to an "exploration" policy in order to gather information on different repositories. For example, it may be desirable to provide search results relating to a sub-optimal repository (e.g., presenting news documents rather than images). One exploration policy may indicate that documents from a random repository be presented to a small fraction of users. Another exploration policy may indicate that documents from a repository be presented in proportion to the score (e.g., if the score for images is determined to be twice the score for news articles, then images may be presented twice as often as news articles).

[**0081**] It has been described that a model may be generated to identify a repository (or a set of repositories) based on a likelihood that a user desires information from the identified repository. In one implementation, the model may be constructed as a lookup table with a key determined based on one or more features, such as one or more features relating to the query (e.g., the query terms). The output of the lookup table might include a click-through rate (or estimated click-through rate) for each of the repositories. In this case, the likelihood that the user desires information from one of the repositories may be a function of the click-through rate for that repository. For example, it might be determined whether to search a repository, whether to include search results from a repository in a search result document, and/or the manner for presenting search results based on the click-through rates for the repositories.

**[0082]** It will be apparent to one of ordinary skill in the art that aspects of the invention, as described above, may be implemented in many different forms of software, firmware, and hardware in the implementations illustrated in the figures. The actual software code or specialized control hardware used to implement aspects consistent with the principles of the invention is not limiting of the invention. Thus, the operation and behavior of the aspects were described without reference to the specific software code—it being understood that one of ordinary skill in the art would be able to design software and control hardware to implement the aspects based on the description herein.

**[0083]** No element, act, or instruction used in the present application should be construed as critical or essential to the invention unless explicitly described as such. Also, as used herein, the article "a" is intended to include one or more items. Where only one item is intended, the term "one" or similar language is used. Further, the phrase "based on" is intended to mean "based, at least in part, on" unless explicitly stated otherwise.

What is claimed is:

**1**. A method, comprising:

receiving a search query from a user;

searching a plurality of repositories, based on the search query, to identify, for each of the repositories, a set of search results;

identifying one of the repositories based on a likelihood that the user desires information from the identified repository; and

presenting the set of search results associated with the identified repository.

**2**. The method of claim 1, further comprising:

generating a model that determines a score associated with a likelihood that a particular user desires information from a repository when the user provides a particular search query.

**3**. The method of claim 2, wherein identifying one of the repositories includes:

determining a score for each of the repositories based on the model, and selecting one of the repositories based on the scores.

**4**. The method of claim 2, wherein generating a model includes:

storing log data associated with a plurality of prior searches, and using the log data to train the model.

**5**. The method of claim 4, wherein generating a model further includes:

representing the log data as triples of data (u, q, r), where u refers to information regarding a user that provided a search query, q refers to information regarding the search query, and r refers to information regarding a repository from which search results were provided in response to the search query.

**6**. The method of claim 5, wherein the log data includes millions of the triples of data (u, q, r).

**7**. The method of claim 5, wherein generating a model further includes:

determining a label for each of the triples of data (u, q, r), where the label includes information regarding whether the user u desired information from the repository r when the user provided the search query q.

**8**. The method of claim 7, wherein using the log data to train the model includes:

training the model based on the triples of data (u, q, r) and the associated labels.

**9**. The method of claim 1, further comprising:

determining a score for each of the repositories, the score for one of the repositories being associated with a likelihood that the user desires information from the one repository.

**10**. The method of claim 9, wherein identifying one of the repositories includes:

selecting one of the repositories with a highest score.

**11**. The method of claim 9, wherein presenting the set of search results associated with the identified repository includes:

providing the sets of search results associated with two or more of the repositories based on their scores.

**12**. The method of claim 11, wherein providing the sets of search results associated with two or more of the repositories based on their scores includes:

arranging the sets of search results within a search result document based on the scores associated with the two or more repositories, and

presenting the search result document to the user.

**13**. The method of claim 12, wherein arranging the sets of search results within a search result document based on the scores associated with the two or more repositories includes:

placing the set of search results associated with a first one of the two or more repositories in a more prominent location within the search result document than the set of search results associated with a second one of the two or more repositories when the score associated with the first repository is higher than the score associated with the second repository.

**14**. The method of claim 12, wherein arranging the sets of search results within a search result document based on the scores associated with the two or more repositories includes:

providing a link to the set of search results associated with at least one of the two or more repositories within the search result document.

**15**. The method of claim 9, further comprising:

selecting a group of repositories to search based on the scores; and

wherein searching a plurality of repositories includes:

performing a search on the group of repositories.

**16**. A system, comprising:

means for receiving a search query from a user;

means for performing a search on a plurality of repositories, based on the search query, to identify, for each of the repositories, a set of search results;

means for determining a score for each of the repositories, the score for one of the repositories being based on a likelihood that the user desires information from the one repository; and

means for providing one or more of the sets of search results based on the scores.

**17**. The system of claim 16, further comprising:

means for selecting a group of the repositories to search based on the scores.

**18**. A system, comprising:

a search engine system to:

receive a search query from a user,

determine a score for each of a plurality of repositories, the score for one of the repositories being based on a likelihood that the user desires information from the one repository,

perform a search on one or more of the repositories, based on the search query, to identify, for each of the one or more repositories, a set of search results, and

provide one or more of the sets of search results based on the scores.

**19**. The system of claim 18, wherein when performing a search on one or more of the repositories, the search engine system is configured to:

identify a group of the repositories to search based on the scores, and

search the group of repositories to identify, for each repository in the group of repositories, a set of search results.

**20**. The system of claim 18, wherein when performing a search on one or more of the repositories, the search engine system is configured to:

search each of the repositories based on the search query.

**21**. The system of claim 18, further comprising:

a model generation system to generate a model that determines a score associated with a likelihood that a particular user desires information from a repository when the user provides a particular search query.

**22**. The system of claim 21, wherein the model is a lookup table and the score corresponds to a click-through rate associated with a repository when the user provides the particular search query.

**23**. The system of claim 21, wherein when determining a score for each of a plurality of repositories, the search engine system is configured to:

determine a score for each of the repositories based on the model.

**24**. The system of claim 21, wherein when generating a model, the model generation system is configured to:

store log data associated with a plurality of prior searches, and

use the log data to train the model.

**25**. The system of claim 24, wherein when generating a model, the model generation system is further configured to:

represent the log data as triples of data (u, q, r), where u refers to information regarding a user that provided a search query, q refers to information regarding the search query, and r refers to information regarding a repository from which search results were provided in response to the search query.

**26**. The system of claim 25, wherein the log data includes millions of the triples of data (u, q, r).

**27**. The system of claim 25, wherein when generating a model, the model generation system is configured to:

determine a label for each of the triples of data (u, q, r), where the label includes information regarding whether the user u desired information from the repository r when the user provided the search query q.

**28**. The system of claim 27, wherein when generating a model, the model generation system is configured to:

train the model based on the triples of data (u, q, r) and the associated labels.

**29**. The system of claim 18, wherein when providing one or more of the sets of search results, the search engine system is configured to:

select one of the repositories with a highest score, and

present the set of search results associated with the selected repository.

**30**. The system of claim 18, wherein when providing one or more of the sets of search results, the search engine system is configured to:

arrange the one or more sets of search results within a search result document based on the scores associated with the one or more repositories, and

present the search result document to the user.

**31**. The system of claim 30, wherein when arranging the one or more sets of search results within a search result document, the search engine system is configured to:

place the set of search results associated with a first one of the one or more repositories in a more prominent location within the search result document than the set of search results associated with a second one of the one or more repositories when the score associated with the first repository is higher than the score associated with the second repository.

**32**. The system of claim 30, wherein when arranging the one or more sets of search results within a search result document, the search engine system is configured to:

provide a link to the set of search results associated with at least one of the one or more repositories within the search result document.

**33**. A computer-readable medium to store data and computer-executable instructions, comprising:

log data associated with a plurality of searches of repositories based on search queries provided by users;

instructions for representing the log data as triples of data (u, q, r), where u refers to information regarding a user that provided a search query, q refers to information regarding the search query, and r refers to information regarding a repository from which search results were provided in response to the search query;

instructions for determining a label for each of the triples of data (u, q, r), where the label includes information regarding whether the user u desired information from the repository r when the user provided the search query q; and

instructions for training a model based on the triples of data (u, q, r) and the associated labels, where the model predicts whether a particular user desires information from a repository when the user provides a particular search query.

**34**. The computer-readable medium of claim 33, wherein the log data includes millions of the triples of data (u, q, r).

**35**. A system, comprising:

a first repository to store a first type of data;

a second repository to store a second type of data; and

a search engine system to:

receive a search query from a user, and

determine a likelihood that the user desires information from the first or second repository based on information regarding the user, the search query, and the first or second repository.

**36**. A system, comprising:

a model generation system to generate a model that determines a score associated with a likelihood that a particular user desires information from a repository when the user provides a particular search query; and

a search engine system to:

receive a search query from a user,

determine a score for each of a plurality of repositories based on the model, and

present search results from one or more of the repositories based on the scores.

**37**. The system of claim 36, wherein the model is a lookup table and the score corresponds to a click-through rate associated with a repository when the user provides the particular search query.

**38**. A method, comprising:

receiving a search query from a user;

determining a score for each of a plurality of repositories, the score for one of the repositories being based on a likelihood that the user desires information from the one repository;

performing a search on at least one of the repositories, based on the search query and the determined scores, to identify, for each of the at least one of the repositories, a set of search results; and

providing one or more of the sets of search results.

**39**. A system, comprising:

a model generation system to generate first and second models, where at least one factor used to generate the second model is different or absent when generating the first model; and

a search engine system to:

receive a search query from a user,

determine a first score for each of a plurality of repositories based on the first model,

perform a search on one or more of the repositories based on the search query and the first scores,

determine a second score for each of the one or more of the repositories based on the second model, and

present search results from at least one of the one or more of the repositories based on the second scores.

**40**. The system of claim 39, wherein an output of the first model is used as an input to the second model.

* * * * *