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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 September 2014 (18.09.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/144905 A1

(51) International Patent Classification:
G06Q 50/10 (2012.01)

(21) International Application Number:
PCT/US2014/029505

(22) International Filing Date:
14 March 2014 (14.03.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

61/800,042	15 March 2013 (15.03.2013)	US
61/800,322	15 March 2013 (15.03.2013)	US
61/800,497	15 March 2013 (15.03.2013)	US

(71) Applicant: **HIGHSPOT, INC.** [US/US]; 2101 4th Avenue, Suite 1830, Seattle, WA 98121 (US).

(72) Inventors: **SHARP, Oliver**; 2101 4th Avenue, Suite 1830, Seattle, WA 98121 (US). **WORTENDYKE, David**; 2101 4th Avenue, Suite 1830, Seattle, WA 98121 (US). **GELLOCK, Scot**; 2101 4th Avenue, Suite 1830, Seattle, WA 98121 (US). **WAHBE, Robert**; 2101 4th Avenue, Suite 1830, Seattle, WA 98121 (US). **VIOLA, Paul**; 2101 4th Avenue, Suite 1830, Seattle, WA 98121 (US).

(74) Agents: **JOHNSON, Anthony, J.** et al.; Perkins Coie LLP, P.O. Box 1247, Seattle, WA 98111-1247 (US).

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

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

Published:

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

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

(54) Title: INTEREST GRAPH-POWERED FEED

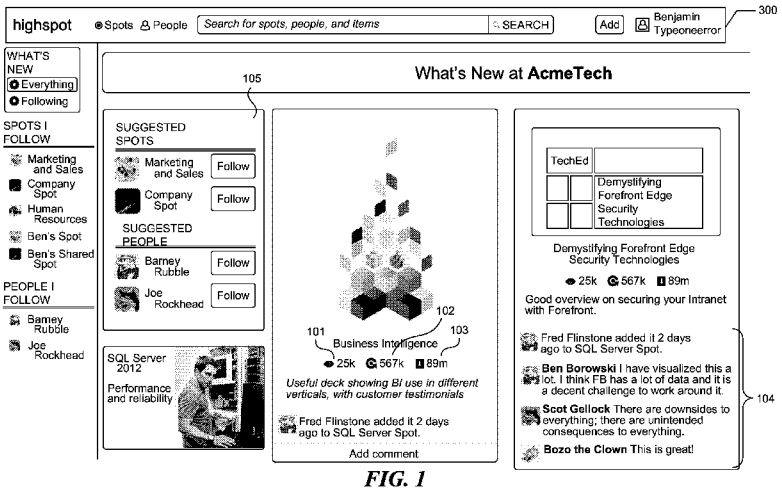


FIG. 1

(57) Abstract: Techniques for organizing information, such as documents, presentations, web sites and web pages, audiovisual media streams, and the like are describe. This disclosed techniques include creating and using an interest graph to support the distribution of information to a user via a feed containing information items likely to be of interest to the user. An interest graph expresses the affinity between people and information - the likelihood that a particular piece of information is of interest to a particular person. The interest graph is based on an understanding of relationships, monitoring of user behavior, and analysis of each piece of information. The interest graph represents many kinds of relationships, including: between users and other users, users and items, and users and collections. The interest graph can be computed using data both from a set of items and from user behavior.

WO 2014/144905 A1

INTEREST GRAPH-POWERED FEED

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Patent Application No. 61/800,042, filed on March 15, 2013, entitled "INTEREST GRAPH-POWERED FEED," U.S. Provisional Patent Application No. 61/800,322, filed on March 15, 2013, entitled "INTEREST GRAPH-POWERED BROWSING", and U.S. Provisional Patent Application No. 61/800,497, filed on March 15, 2013, entitled "INTEREST GRAPH-POWERED SHARING, all of which is herein incorporated by reference in its entirety. This application is related to U.S. Provisional Patent Application No. 61/745,365 filed on December 21, 2012, entitled "INTEREST GRAPH-POWERED SEARCH," U.S. Patent Application No. 14/136,322 filed on December 20, 2013, entitled "INTEREST GRAPH-POWERED SEARCH," U.S. Provisional Patent Application No. 61/914,266, filed on December 10, 2013, entitled "SKIM PREVIEW," U.S. Provisional Patent Application No. 61/953,258, filed on March 14, 2014, entitled "NARROWING INFORMATION SEARCH RESULTS FOR PRESENTATION TO A USER," U.S. Patent Application No. _____ (Attorney Docket No. 0798418002US1), filed on March 14, 2014, entitled "INTEREST GRAPH-POWERED BROWSING," and U.S. Patent Application No. _____ (Attorney Docket No. 0798418006US1), filed on March 14, 2014, entitled "INTEREST GRAPH-POWERED SHARING, all of which are herein incorporated by reference in their entireties.

BACKGROUND

[0002] Currently, internal networks within organizations do not automatically find the information or data (e.g., business information) that employees need to do their jobs and present it for discovery. Intranet services do not gather information about users, search through information available across the company, and find the most relevant documents and other business information. Users must seek out the information they need and are often left unaware of highly relevant information that they could benefit from.

[0003] The need exists for a system that learns about the interests of each employee and identifies the most compelling and relevant information that is accessible to them (from within the company and from outside it).

[0004] Overall, the examples herein of some prior or related systems and their associated limitations are intended to be illustrative and not exclusive. Other limitations of existing or prior systems will become apparent to those of skill in the art upon reading the following Detailed Description.

BRIEF DESCRIPTION OF THE DRAWINGS

[0005] Figure 1 is a display page illustrating a feed of business information for a user.

[0006] Figure 2 is a block diagram illustrating processing of the system for creating an interest graph.

[0007] Figure 3 is a display page illustrating a query and results page.

[0008] Figure 4A is a display page illustrating a search completion feature.

[0009] Figure 4B is a display page illustrating a search completion and number of results feature.

[0010] Figure 5 is a display page illustrating de-duplication of identified search results.

[0011] Figure 6 is a display page illustrating the expansion of duplicated search results.

[0012] Figure 7 is a display page illustrating a collection of information called a "spot."

[0013] Figure 8 is a display page illustrating a sub-collection of information called a "spotlist."

[0014] Figure 9 is a display page illustrating a group of results filtered down via "narrow-by."

[0015] Figure 10 is a display page illustrating an item with a list of related items.

[0016] Figure 11 is a display page illustrating sharing via search.

[0017] Figure 12 is a display page illustrating sharing via browsing.

[0018] Figure 13 is a display page illustrating sharing via discovery.

[0019] Figure 14 is a display page illustrating how item and collection statistics are presented.

[0020] Figure 15 is a display page illustrating how user activity is presented.

[0021] Figure 16 is a block diagram illustrating some of the components that may be incorporated in at least some of the computer systems and other devices on which the facility operates and interacts with.

DETAILED DESCRIPTION

I. FEED

[0022] Various examples of the technology will now be described. The following description provides certain specific details for a thorough understanding and enabling description of these examples. One skilled in the relevant technology will understand, however, that the disclosed technology may be practiced without many of these details. Likewise, one skilled in the relevant technology will understand that the disclosed techniques may include many other features not described in detail herein. Additionally, some well-known structures or functions may not be shown or described in detail below, to avoid unnecessarily obscuring the relevant descriptions of the various examples.

[0023] The terminology used below is to be interpreted in its broadest reasonable manner, even though it is being used in conjunction with a detailed description of certain specific examples of the disclosed technology. Indeed, certain terms may even be emphasized below; however, any terminology intended to be interpreted in any restricted manner will be overtly and specifically defined as such in this Detailed Description section.

[0024] The technology described below creates an *information feed* that contains items like documents and presentations, web sites and pages, and audiovisual media streams. Each item is presented with *user signal* that represents the way that the community inside and outside the organization has been interacting with that information. For example, the feed shows how frequently an item has been viewed and what comments have been made about it. The feed also honors access

rights - within organizations, there are often restrictions as to what data is available to each person, so the feed shows each user information that they have access to.

[0025] This disclosure describes the creation and use of an *interest graph* within a company, and between companies, to drive the presentation of an *information feed*. An interest graph expresses the *affinity* between people and information - the likelihood that a particular piece of information is of interest to a particular person. The information might be a document, a presentation, a video, an image, a web page, a report, or the like. The information might also be a *collection* of items, or a *link* to a collection of items or to a person. The interest graph is based on an understanding of relationships, monitoring of user behavior, and analysis of each piece of information. The interest graph can represent many kinds of relationships, including: between users and other users, users and items, and users and collections. The interest graph can be computed using data both from the set of items and from user behavior. In some examples, there are three steps for computing the interest graph. The first step is to *generate the data;* the system provides mechanisms for the user to quickly browse, share, and organize items of information. By using those features, the users create a large amount of usage data, much of which is currently uncollected and unavailable to existing information management and retrieval software. The next step is to *gather the data*, where the system logs user activities in a set of data structures. The third step is to *compute the interest graph*. By running a series of computations over the information gathered from users, the system computes data structures that are used for a variety of ranking or search operations. The disclosed techniques honor access restrictions that users specify for each item, so that only authorized people will see any item of information.

INFORMATION FEED

[0026] One way that users discover useful and compelling content online is through *discovery*. Discovery is opportunistic - the system learns about that user and what they are interested in, and presents items based on that understanding. For example, the system can track the information that users have viewed in the past, and find items that are similar or that were viewed by other people who looked at the same or similar information as the current user. The information that the

system identifies is presented to the user as a sequence of items, typically in exactly or approximately the order of the time that they were created or updated. This sequence is known as the *feed.*

[0027] The disclosed system creates a feed of information, such as *business information*, based on the interests of the user, which are analyzed by assembling an *interest graph*. Figure 1 shows the feed for a user. Note that the items in the feed are documents, presentations, images, Web pages, and other information that help employees do their work. Also note that associated with each item is a summary of *user signal* - the amount of activity the item has sparked within the community of users who have access to it. In Figure 1, there are three numbers reported for each item – the number of *views* 101, the number of *respots* 102, and the number of *downloads* 103. A "respot" is a decision by one user to republish an item they find of interest into one of their own collections. Similarly, if users have commented on the item, a selection of those comments 104 are also included in the feed. These user signals are shown in the feed to help users assess how useful the community has found the item to be. As is detailed below, in addition to the signal that is shown to other users, many other signals can be used to compute the interest graph.

[0028] In addition to the business content itself, the feed can also be used to recommend users whose activities may be of interest and collections of items that seem relevant. In Figure 1, there is an entry in the feed containing such suggestions 105.

[0029] The feed can be presented to users through a Web experience, as shown in Figure 1, as an online or offline custom application experience on a PC or on a device (such as a phone or tablet), via push mechanisms like email or instant messaging, or in other social networks (such as Twitter, Yammer, Pinterest, or Facebook), and other outlets. A push notification can be triggered automatically when a sufficient number of relevant items are found, or at a user specified interval.

[0030] In some examples, the feed is implemented in the system as follows:

- Determine which items are both accessible to the user and contain words or strings from the search library. This process may be performed using a software package, such as the Lucene software library supported by the Apache Software Foundation. As described below, Lucene computes and

uses an *index* that reports, for every lexeme in the search query, the items that contain that lexeme. This reported set of items is the *candidate set of items.*

- Compute a textual rank for each candidate item (i.e., an item in the candidate set) based on a similarity algorithm. This rank is based on the textual contents of the items and on the search phrase.
- Run a computation (i.e., a subroutine) called **ItemRanker** (described below) on the candidate set to assign an overall rank to each item. The computation relies on the interest graph and the textual rank to compute a rank for each item indicating how interesting it is expected to be to the current user (i.e., the user for whom the feed is generated).
    - Note that the activity counts used by **ItemRanker**, which are stored in the **ItemScore** data structure, can be multiplied by a fractional value each day. This has the effect of causing older actions to be weighted less than more recent ones (producing an exponential effect based on the age in days of the activity). The fractional value determines how quickly the usefulness of older activity attenuates. There are many other algorithms that could be used to calculate the recency of an item, including (but not limited to) using a fixed or adjusted linear scale factor based on the age of the activity, discarding data that is older than a certain window of time, or attenuating the data for a period of time and treating all older activity as equal in weight and hence in importance. Also note that it is possible to use different age algorithms for different computations. For example, the system offers an alternative sort order for items on a spot that biases more for recent activity, by attenuating the weight of older activity more aggressively than it normally does. It is also possible to use different weights in the ranking algorithm for different computations or sort orders.
- Present the resulting items to the user, sorted, for example, in overall relevance rank order (highest value first).

RESULT DEDUPLICATION

[0031] It is very common for the same item to appear many times across the internal networks of an organization. If that item is of strong interest to the user, it could easily appear many times in the feed, "polluting" the feed. To prevent feed pollution, the system identifies when two items are identical (even if they have been separately copied into different collections of information and/or given different filenames) and will only present such items once in the feed.

[0032] In some examples, the system implements de-duplication as follows:

- As the contents of items are indexed, a *hash code* is computed for each item based on its contents (see below for details). This hash code is, with high probability, unique across all the items. When two items have matching hash codes, they are noted in the inverted index as having identical contents.
- As the items are ranked for presentation in the feed using the interest graph, identical items are combined (grouped), though their individual ranks are recorded. When the ranked search results are presented, the system presents the single item with the highest rank.
- If the user requests it, the system expands the collapsed items to show duplicated matches, still ranked in order of likelihood to interest the user based on the interest graph.

OTHER POTENTIAL USES

[0033] There are a number of other ways that the feed can be used and the interest graph powering it can be enhanced in various embodiments of the disclosed technology.

- **Augment the interest graph computation with real-time feedback**. Regularly adjust the algorithms used to compute the interest graph and the weightings based on the effect on user behavior. For example, divide the user community into groups (sometimes called *cohorts*), present the results of different weights or algorithms to the different groups, and adjust the system based on the most successful approach.
- **Track more granular user behavior**. Follow the mouse movements of a user to get additional information about their interests.

- **Voting and reputation**. By providing ways for users to vote on the contributions of others, the voting patterns resemble commenting patterns, and can be added into the data structures that track interactions between users and measure the influence of a user on others in the community. For example, if one user votes for the contribution of another, the system can treat that the same way as making a comment. Each such vote represents an additional amount of influence that can be added to other indications, such as the number of followers they have or the number of times others commented on an item that a user uploaded.
- **Rank subsets of an item**. Based on the sections of video that other users have watched or otherwise interacted with, and the affinity of the current user to the other users, the system can identify and present the sections within the video most likely to be of interest. Similarly with other media types, like audio, or with components of structured documents, like a CAD diagram, the system can identify a section or sections thereof most likely to be of interest to a particular user. Similarly, the system might rank the pages of a document or the slides in a presentation in terms of which are most likely to interest the user based on the extent to which users have viewed or interacted with portions thereof. These ranks could be used to create a digest (summary) of the item that is faster for users to review, such as a "highlights" synopsis of a video, or a summary document.
- **Present derived information**. In addition to indexing the contents of an item, the system can apply a variety of computations that derive new information based on the content, and apply the interest graph to searching that derived data. For example, the system can perform voice recognition of an audio or video stream and search the text that it has computed via the voice recognition. The system may perform a dependency analysis on a CAD diagram and include dependent components in the feed, even though they are not present in the original diagram. Each of the derived items may be treated as a separate content item for ranking and affinity calculation purposes. The system may perform image recognition and search for the names or characteristics of objects and people that have been recognized.

- **Incorporate additional sources of social graph information**. Patterns of email for each user provide an indication of the topics that they are interested in and the people with whom they interact most frequently. The interest graph could be enhanced by performing content and traffic analysis on email and on aliases that have been set up by users to communicate with each other. Another example is the user directory. Many organizations maintain directories that maintain the membership of users in groups and the relationship between groups. A common example is Active Directory. Such directories are used to enforce access permissions, set up email aliases, and a variety of other purposes. There are also HR and payroll systems that capture relationships among users as well. Another source of information is on-premise server systems. For example, by looking at the permission sets for file systems and the access patterns, additional social information would be available that is similar to what the system tracks internally. Another potential source is public social networks outside of the organization (such as Facebook, Twitter, Instagram, and the like). All of these systems have information about the social graph that could be incorporated into the interest graph. For example, if one user "follows" another person in the company on Twitter, that can be treated similarly to following somebody within the system and incorporated into the interest graph. Likewise when two people are "friends" on Facebook, or otherwise connected on a social networking site, this connection can be included in the interest graph. As another example, comments by one person on an item posted by another, which is similar to commenting on an item published within the system, can be incorporated into the interest graph.
- **Use role/title information**. If users are assigned roles or titles, either by self-selection or by the company, those with similar roles and titles are likely to have some interests in common - another signal that can be incorporated into the interest graph.
- **Identify similar items as well as true duplicates**. Often items are related but not identical. For example, a slide deck that contains some of the same slides as another, but reordered or intermixed with others. Or, a document that has been modified, but much of the substance is the same. Similarity

could be used to do more aggressive deduplication, or to identify when content is related, or to identify when users have similar interests because they interact with similar documents as well as identical ones. Another option is to offer users the ability to construct new content from pieces of existing ones. For example, assembling a presentation from slides that come from multiple existing decks. This would reveal the relationships between the assembled and original decks, and would give additional signal on which slides of a presentation are the most valuable.

- **Additional types of content analysis.** There are a variety of ways to analyze content to reveal information that would be useful for the interest graph. For example, performing facial and feature recognition of images. The analysis could be used to find items that are about related topics. Facial recognition could be used to compare content images to user profile pictures, to identify content that is about people in the system. The system could automatically perform language translation on items of particular interest. Or it could create higher resolution previews, or graphical rendering/visualization of data, or create 3D images. The system could automatically perform language translation on items of particular interest or create higher resolution previews, graphical rendering/visualization of data, or 3D images using known techniques.
- **Proactively get ready to deliver items likely to be of interest**. Items likely to be of interest to a user can be downloaded to a device for optimized or offline access, or pre-cached in a content delivery network (CDN) for faster delivery to the browser. A group of items might be presented to the user together for faster viewing.
- **Target advertisements and promotional offers.** The system may target offers to users based on their activities and interests (e.g., the items they are currently browsing). These might be sponsored by the organization they work for, by a partner, or by outside companies. For example, a company might allow vendors to compete for or purchase the right to market services to employees. Similarly, there might be a facility for "sponsoring" particular items and ranking sponsored items higher, highlighting sponsored items visually, indicating the nature of the sponsorship, and so on. Such

sponsorship might be done manually, by an algorithm, by a business rule, by an expert system.

- **Instant search.** The system can present search results incrementally as the user is typing, rather than waiting for them to specify a full query.
- **Semantic search.** Search queries can be semantically analyzed using techniques like latent semantic analysis and a variety of natural language processing algorithms that perform operations, such as relationship extraction, named entity recognition, and the like. Then, the system can do specialized operations appropriate for a particular domain or a particular semantic concept. For example, if the system determined that a search applied to a set of legal cases, it might automatically detect references to legal precedents and search through them as well as through the case itself. In manufacturing, the system could identify that a number was a reference to a part and extend its search to include the supporting information for that particular part.

II. SEARCH/DISCOVERY

SEARCHING

[0034] One of the most common ways that users look for information online is to type a *query* into a *search box*. The system uses the query to identify a candidate set of items, collections, and people that match it, attempt to rank order those candidates based on what is most likely to satisfy that user's request, and present the results. The system uses the interest graph to support search across items within a particular company and between multiple companies. Figure 3 is a display page 300 illustrating a query and results page in accordance with some embodiments of the disclosed technology. In this example, a search and the set of results that have been returned by the system are displayed. In this example, a user has performed a search for "vpn" and the system has identified 31 results, including "Daewoo Engineering and Construction Deploy's Korea's Largest SSL VPN" 310 and "Juniper Networks Enables Secure Remote Access For the Thai Technology Institute" 320. The results are ranked based on how likely they are to interest the

user, as explained below. For each result, the system highlights the elements of the search phrase that matched text associated with the item.

BASIC SEARCH

[0035] In a basic search, the user provides a string, and the system identifies items that the user has access to and that match the string. The items may include information which has been created by another user within the same organization (which will be referred to as an *internal* item) or by a user from elsewhere (an *external* item). The system allows items to be shared within and across organizations, and for their access to be restricted to particular sets of people.

[0036] In some examples, basic search is implemented in the system as follows:

- Determine which items are both accessible to the user and contain words or strings from the search query. This process may be performed using a software package, such as the Lucene software library supported by the Apache Software Foundation . As described below, Lucene computes and uses an *inverted index* that reports, for every *lexeme* in the search query, the items that contain that lexeme. This reported set of items is the *candidate set* of items.
- Compute a *textual rank* for each candidate item (i.e., an item in the candidate set) based on a *similarity* algorithm. This rank is based on the textual contents of the items and on the search phrase.
- Run a computation (i.e., a subroutine) called **ItemRanker** (described below) on the candidate set to assign an overall rank to each item. The computation relies on the interest graph and the textual rank to compute the overall rank.
- Present the resulting items to the user, sorted in overall rank order (highest value first).

SEARCH COMPLETION

[0037] Search completion is a feature that shows possible search queries while the user is typing in the search box (see Figure 4A for an example). Figure 4A is a display page 200 illustrating a search completion feature in accordance with some embodiments of the disclosed technology. In this example, a user has typed "se" in

the search box 410 and the system has shown "secure remote," "secure services gateway," "security gateway," and "security standards" as possible search queries 420. Search completion is a very useful shortcut for users, showing them search queries that are likely to be relevant and that have a good probability of yielding useful results.

[0038] In some examples, search completion is implemented in the system as follows:

- As the user types, the system passes the partial search query to a computation called **CompletionRanker** (described below). **CompletionRanker** returns a ranked set of search completions that are most likely to be of interest to the user, based on the behavior of other users in the system and on the information in the interest graph.
- Present the ranked set of completions to the user in a dropdown.

SEARCH COMPLETION WITH COUNTS AND GROUPINGS

[0039] Additionally the set of the completions can include information about the number of results that would be returned by each query. This information can help the user to formulate a query that will return a satisfactory number of results. Additionally some of the completions can include a speculative scoping of the search to a user specified collection, such as a spot. For example for the query {vision}, one suggestion could be {vision in the Benefits Spot (5 docs)}. Selecting this suggestion will return the 5 documents that are contained in this spot. Similarly the scope can a single user or group of users. For example {sql} could yield {sql by Robert Wahbe (20 docs)}. Selecting this would show the 20 docs uploaded by Robert that contain the term sql. Figure 4B is a display page 450 illustrating a search completion and "number of results" feature in accordance with some embodiments of the disclosed technology. In this example, a user has typed "paul" into search box 460 and the system has shown "paul's or textteaser," "paul feed," "paul viola," and "paul roe" as possible search queries and their corresponding number of results (i.e., 96, 30, 7, and 2) 470. The system has also shown the number of results of "paul" in various collections, such as "Music" {12}, "Systems" {12}, "UX Design" {10}, and "Market Intelligence" {9} 280.

DE-DUPLICATE SEARCH RESULTS

[0040] For business information, it is common that the same item appear many times in many different collections of items. For example, a particularly useful presentation might be placed in an official repository, downloaded and emailed to many people, and then posted to a variety of different collections. The system identifies cases where an item has been duplicated, combines those into a single item in the results presented to the user, and uses the interest graph to choose the one most likely to interest that user. Figure 5 is a display page 500 illustrating de-duplication of identified search results in accordance with some embodiments of the disclosed technology. Figure 3 shows a search that has found at least two duplicated items ("China's Ministry of Railways Secures Information Management System" 510 and "Southeast University Builds a 'University Without Walls'" 520) and combined each set of duplicated items into a single instance.

[0041] Figure 6 is a display page 600 illustrating the expansion of duplicated search results in accordance with some embodiments of the disclosed technology. Figure 4 shows how a single item ("Daewoo Engineering and Construction Deploy's Korea's Largest SSL VPN") can be expanded to show the copies and, for example, each copy's location 610 and 620.

[0042] In some examples, the system implements de-duplication as follows:

- As the contents of items are indexed, a *hash code* is computed for each item based on its contents (see below for details). This hash code is, with high probability, unique across all the items. When two items have matching hash codes, they are noted in the inverted index as having identical contents.
- As the items are ranked using the interest graph, identical items are combined (grouped), though their individual ranks are recorded. When the ranked search results are presented, the system presents the single item with the highest rank (as shown in Figure 5).
- If the user requests it, the system expands the collapsed items to show duplicated matches, still ranked in order of likelihood to interest the user based on the interest graph.

OTHER POTENTIAL USES

[0043] There are a number of other ways that the interest graph can power improved search behavior:

- **Rank subsets of an item**. Based on the sections of video that other users have watched, and the affinity of the current user to the other users, the system can identify the sections within the video most likely to be of interest. Similarly with other media types, like audio, or with components of structured documents, like a CAD diagram, the system can identify a section or sections thereof most likely to be of interest to a particular user. Similarly, the system might rank the pages of a document or the slides in a presentation in terms of which are most likely to interest the user based on the extent to which users have viewed portions thereof.
- **Search for derived information**. In addition to indexing the contents of an item, the system can apply a variety of computations that derive new information based on the content, and apply the interest graph to searching that derived data. For example, the system can perform voice recognition of an audio or video stream and search the text that it has computed via the voice recognition The system may perform a dependency analysis on a CAD diagram and include dependent components in the search operation, even though they are not present in the original diagram. The system may perform image recognition and search for the names or characteristics of objects and people that have been recognized.
- **Instant search**. The system can present search results incrementally as the user is typing, rather than waiting for them to specify a full query.
- **Semantic search**. Search queries can be semantically analyzed using techniques like latent semantic analysis and a variety of natural language processing algorithms that perform operations , such as relationship extraction, named entity recognition, and the like. Then, the system can do specialized operations appropriate for a particular domain or a particular semantic concept. For example, if the system determined that a search applied to a set of legal cases, it might automatically detect references to legal precedents and search through them as well as through the case itself. In manufacturing, the system could identify that a number was a reference to a

part and extend its search to include the supporting information for that particular part.

III. BROWSE

[0044] This disclosure describes the creation and use of an *interest graph* within a company, and between companies, to drive information *browsing*. An interest graph expresses the *affinity* between people and information - the likelihood that a particular piece of information is of interest to a particular person. The information might be a document, a presentation, a video, an image, a web page, a report, or the like. The information might also be a *collection* of items, or a *link* to a collection of items or to a person. The interest graph is based on an understanding of relationships, monitoring of user behavior, and analysis of each piece of information. The interest graph can represent many kinds of relationships, including: between users and other users, users and items, and users and collections. The interest graph can be computed using data both from the set of items and from user behavior. In some examples, there are three steps for computing the interest graph. The first step is to *generate the data;* the system provides mechanisms for the user to quickly browse, share, and organize items of information. By using those features, the users create a large amount of usage data, much of which is currently uncollected and unavailable to existing information management and retrieval software. The next step is to *gather the data*, where the system logs user activities in a set of data structures. The third step is to *compute the interest graph*. By running a series of computations over the information gathered from users, the system computes data structures that are used for a variety of ranking or search operations. The disclosed techniques honor access restrictions that users specify for each item, so that only authorized people will see any item of information.

BROWSING

[0045] One way that users find useful and compelling content online is to *browse* through collections of content. In some examples of the disclosed system, the collections are called *spots,* which can be further organized by placing items of content into *spotlists,* or lists of items. A content item can be placed into any number of spotlists. Spotlists can also be gathered into *folders*. A user can browse content

in many ways, including but limited to: viewing a directory of spots, finding a link to the collection on another spot, having the system suggest a collection, searching, having a link shared with them by another user, and so on. Users can, for example, look at the spot as a whole or look at a sub-collection of the spot by choosing a spotlist or a folder. Figure 7 shows a spot *home page* 700, which is the default view in the system when a user visits a spot in some embodiments. On the left, under the word "Browse", a set of spotlists and folders are presented. For example, the spotlist "Implementor" 705 is a spotlist with 48 items in it and is in a folder called "Audience" 710. If the user clicks on a spotlist (e.g., "Implementor," "Decision Maker," "Proof of Concept"), the system displays the items in that list. Figure 8 shows the result of clicking on the spotlist called "Business Intelligence" 715.

[0046] One of the ways the system helps users refine a particular group of results is with a technique called *narrow-by*. When a particular set of items is being presented, the system computes every spotlist that any item in the set belongs to. For example, an item might belong to a spotlist that relates to its target audience (such as "Implementor" and "Decision Maker" in Figure 7). The item might also belong to a spotlist related to a topic it covers (such as "Business Intelligence"), or a region of the world that it is relevant for, or the like. If a spotlist contains only a subset of the items, it can be offered as a checkbox item that can be used to restrict the set of results.

[0047] Figure 8 is a display page 800 illustrating a sub-collection of information called a "spotlist" in some embodiments. In this example, there are 21 items in the "Business Intelligence" spotlist, but only 10 of them are presentations. In some cases, a spotlist, such as the "Presentation" spotlist 805, is a *smart spotlist,* meaning that its contents are automatically computed by the system rather than specified by the user. Other spotlists, such as the "Post-sale" spotlist 810, are *user spotlists*, where items are assigned to different spotlists by a user. In this example, 12 of the 21 business intelligence items are also in the post-sale spotlist. The user can ask the system to narrow down the results based on one of these other spotlists by clicking the checkbox next to the spotlist name.

[0048] Figure 9 is a display page 900 illustrating a group of results filtered down via "narrow-by." In this example, the user has selected the "post-sale" checkbox 910 – the result set is reduced to only twelve items. The user can further constrain the

results by clicking other checkboxes, or return to the full set by unchecking "post-sale."

[0049] At any given time, there is a current set of results, and these are by default presented to the user in *relevance* order. That order is computed by the interest graph, as described below, can be customized for each user, and is ordered based on what the system knows about that user's interests.

[0050] In some examples, the browsing experience is presented to users through a Web experience, as shown in Figures 7-10. One of ordinary skill in the art will recognize that the browsing experience can also be presented as an online or offline custom application experience on a PC or on a device (such as a phone or tablet).

[0051] In some examples, the ranking of items during browsing is implemented in the system as follows:

- Determine the set of items in the current *result set*. This may be based on the user choosing to look at a collection of items (in the system, this is done by visiting a particular spot), or a subset (by clicking on a spotlist or folder), or performing a search operation. In the example shown in Figure 7, the user can also use the selector 720 in the upper left to see all the items, only the ones that were last added, or all the ones added during the past week.
- Run a computation (i.e., a subroutine) called **ItemRanker** (described below) on the result set to assign an overall rank to each item. The computation uses the interest graph to compute a rank for each item indicating how interesting it is expected to be to the current user (i.e., the user for whom the items are being ranked).
    - Note that the activity counts used by **ItemRanker**, which are stored in the **ItemScore** data structure, can be multiplied by a fractional value each day. This has the effect of causing older actions to be weighted less than more recent ones (producing exponential decay based on the age in days of the activity). The fractional value determines how quickly the usefulness of older activity attenuates. There are many other algorithms that could be used to calculate the recency of an item, including (but not limited to) using a fixed or adjustable linear scale factor based on the age of the activity, discarding data that is older

than a certain window of time, or attenuating the data for a fixed period of time and treating all older activity as equal in weight and hence in importance. Also note that it is possible to use different age algorithms for different computations. For example, the system offers an alternative sort order for items on a spot that biases more for recent activity, by attenuating the weight of older activity more aggressively than it normally does. It is also possible to use different weights in the ranking algorithm for different computations or sort orders.

- Present the resulting items to the user, sorted, for example, in overall relevance rank order (highest value first).

[0052] In some examples, the suggestion of a collection of items that might interest the user is implemented in the system as follows:

1. Determine the set of collections that the user has access to. In the system, the permissions can be stored in a database, indexed by both collection and user.
2. Run a computation (i.e., a subroutine) called **CollectionRanker** (described below) on the result set to assign an overall rank to each item. The computation uses the interest graph to compute a rank for each collection indicating how interesting the collection is expected to be to the current user (i.e., the user for whom the suggested collection of items is generated).
3. Present the collections having an interest value above a threshold, sorted, for example, in overall relevance rank order (highest value first). The threshold is a value that represents a trade-off between *precision* against *recall* that may be established by the system, a user of the system, or an administrator of the system. *Recall* is the likelihood that the system will find something you are interested in, and precision is how well the system weeds out the things that you do not wish to see. A higher threshold improves precision (shows fewer items, that are more likely to be useful) and a lower threshold improves recall (shows more items, and hence is more likely to include a particular one of interest).

RELATED ITEMS

[0053] While browsing for items, users often find an item about a topic of interest and use that item as a "seed" and let the system identify and suggest related items for further exploration. The system supports an interest-graph based model for finding information; whenever the user looks at an item, the system can suggest related items based on the interest graph. The items presented are customized for each user, based on what the system knows about the items and the user. Figure 10 is a display page 1000 illustrating an item with a list of related items and the suggestions 1005 appear in the lower right hand part of the screen.

[0054] In some examples, the system identifies related items as follows:

1. Run a computation (i.e., a subroutine) called **RelatedItemRanker** (described in detail below) on the current item to yield a ranked set of items that have a relationship to the current item. The computation uses the interest graph to compute a rank for each item indicating how interesting the item is expected to be to the current user. The computation uses a variety of inputs to measure the strength of the relationship, such as textual similarity, the fact that the items are of interest to users with similar behavior, the appearance of the items in the same collections of sub-collections, and so forth. Note that in some embodiments the system will only consider related items that the current user has permission to access.
2. Present the resulting items to the user, sorted, for example, in overall rank order (highest value first).

OTHER POTENTIAL USES

[0055] There are a number of other ways that the system can support browsing and the interest graph powering it can be enhanced in various embodiments of the disclosed technology.

- **Augment the interest graph computation with real-time feedback**. Regularly adjust the algorithms used to compute the interest graph and the weightings based on the effect on user behavior. For example, divide the user community into groups (sometimes called *cohorts*), present the results of

different weights or algorithms to the different groups, and adjust the system based on the most successful approach.

- **Track more granular user behavior**. Follow the mouse movements of a user to get additional information about their interests.
- **Voting and reputation**. By providing ways for users to vote on the contributions of others, the voting patterns resemble commenting patterns, and can be added into the data structures that track interactions between users and measure the influence of a user on others in the community. For example, if one user votes for the contribution of another, the system can treat that the same way as making a comment. Each such vote represents an additional amount of influence that can be added to other indications, such as the number of followers they have or the number of times others commented on an item that a user uploaded.
- **Rank subsets of an item**. Based on the sections of video that other users have watched or otherwise interacted with, and the affinity of the current user to the other users, the system can identify and present the sections within the video most likely to be of interest. Similarly with other media types, like audio, or with components of structured documents, like a CAD diagram, the system can identify a section or sections thereof most likely to be of interest to a particular user. Similarly, the system might rank the pages of a document or the slides in a presentation in terms of which are most likely to interest the user based on the extent to which users have viewed or interacted with portions thereof. These ranks could be used to create a digest (summary) of the item that is faster for users to review, such as a "highlights" synopsis of a video, or a summary document.
- **Search for derived information**. In addition to indexing the contents of an item, the system can apply a variety of computations that derive new information based on the content, and apply the interest graph to searching that derived data. For example, the system can perform voice recognition of an audio or video stream and search the text that it has computed via the voice recognition. The system may perform a dependency analysis on a CAD diagram and include dependent components for a user to browse, even though they are not present in the original diagram. Each of the derived items

may be treated as a separate content item for ranking and affinity calculation purposes. The system may perform image recognition and search for the names or characteristics of objects and people that have been recognized.

- **Incorporate additional sources of social graph information**. Patterns of email for each user provide an indication of the topics that they are interested in and the people with whom they interact most frequently. The interest graph could be enhanced by performing content and traffic analysis on email and on aliases that have been set up by users to communicate with each other. Another example is the user directory. Many organizations maintain directories that maintain the membership of users in groups and the relationship between groups. A common example is Active Directory. Such directories are used to enforce access permissions, set up email aliases, and a variety of other purposes. There are also HR and payroll systems that capture relationships among users as well. Another source of information is on-premise server systems. For example, by looking at the permission sets for file systems and the access patterns, additional social information would be available that is similar to what the system tracks internally. Another potential source is public social networks outside of the organization (such as Facebook, Twitter, Instagram, and the like). All of these systems have information about the social graph that could be incorporated into the interest graph. For example, if one user "follows" another person in the company on Twitter, that can be treated similarly to following somebody within the system and incorporated into the interest graph. Likewise when two people are "friends" on Facebook, or otherwise connected on a social networking site, this connection can be included in the interest graph. As another example, comments by one person on an item posted by another, which is similar to commenting on an item published within the system, can be incorporated into the interest graph.
- **Use role/title information**. If users are assigned roles or titles, either by self-selection or by the company, those with similar roles and titles are likely to have some interests in common - another signal that can be incorporated into the interest graph.

- **Identify similar items as well as true duplicates**. Often items are related but not identical. For example, a slide deck that contains some of the same slides as another, but reordered or intermixed with others. Or, a document that has been modified, but much of the substance is the same. Similarity could be used to do more aggressive deduplication, or to identify when content is related, or to identify when users have similar interests because they interact with similar documents as well as identical ones. Another option is to offer users the ability to construct new content from pieces of existing ones. For example, assembling a presentation from slides that come from multiple existing decks. This would reveal the relationships between the assembled and original decks, and would give additional signal on which slides of a presentation are the most valuable.
- **Additional types of content analysis.** There are a variety of ways to analyze content to reveal information that would be useful for the interest graph. For example, performing facial and feature recognition of images. The analysis could be used to find items that are about related topics. Facial recognition could be used to compare content images to user profile pictures, to identify content that is about people in the system. The system could automatically perform language translation on items of particular interest. Or it could create higher resolution previews, or graphical rendering/visualization of data, or create 3D images. The system could automatically perform language translation on items of particular interest or create higher resolution previews, graphical rendering/visualization of data, or 3D images using known techniques.
- **Proactively get ready to deliver items likely to be of interest**. Items likely to be of interest to a user can be downloaded to a device for optimized or offline access, or pre-cached in a content delivery network (CDN) for faster delivery to the browser. A group of items might be presented to the user together for faster viewing.
- **Target advertisements and promotional offers.** The system may target offers to users based on their activities and interests (e.g., the items they are currently browsing). These might be sponsored by the organization they work for, by a partner, or by outside companies. For example, a company might

allow vendors to compete for or purchase the right to market services to employees. Similarly, there might be a facility for "sponsoring" particular items and ranking sponsored items higher, highlighting sponsored items visually, indicating the nature of the sponsorship, and so on. Such sponsorship might be done manually, by an algorithm, by a business rule, by an expert system.

- **Instant search.** The system can present search results incrementally as the user is typing, rather than waiting for them to specify a full query.
- **Semantic search.** Search queries can be semantically analyzed using techniques like latent semantic analysis and a variety of natural language processing algorithms that perform operations, such as relationship extraction, named entity recognition, and the like. Then, the system can do specialized operations appropriate for a particular domain or a particular semantic concept. For example, if the system determined that a search applied to a set of legal cases, it might automatically detect references to legal precedents and search through them as well as through the case itself. In manufacturing, the system could identify that a number was a reference to a part and extend its search to include the supporting information for that particular part.

IV. SHARE

[0056] This disclosure describes the creation and use of an *interest graph* within a company, and between companies, to support *sharing* information (e.g., business information) via search, browsing, and discovery, and *measuring consumption, engagement, and/or influence* based on that information. A piece of information is "consumed" when a user views it, listens to it, or otherwise interacts with it. "Engagement" measures user activity against the item - sharing it, adding it to another item collection, commenting on it, and so forth. The amount of "influence" of a user can be measured in a variety of ways. For example, one approach is to count the number of "followers" a person has – the other users who have asked to be notified when the user performs actions like commenting on a document communicating a piece or item of information. An interest graph expresses the *affinity* between people and information - the likelihood that a particular piece of

information is of interest to a particular person. The information might be a document, a presentation, a video, an image, a web page, a report, or the like. The information might also be a *collection* of items, or a *link* to a collection of items or to a person. The interest graph is based on an understanding of relationships, monitoring of user behavior, and analysis of each piece of information. The interest graph can represent many kinds of relationships, including: between users and other users, users and items, and users and collections. The interest graph can be computed using data both from the set of items and from user behavior. In some examples, there are three steps for computing the interest graph. The first step is to *generate the data;* the system provides mechanisms for the user to quickly browse, share, and organize items of information. By using those features, the users create a large amount of usage data, much of which is currently uncollected and unavailable to existing information management and retrieval software. The next step is to *gather the data*, where the system logs user activities in a set of data structures. The third step is to *compute the interest graph*. By running a series of computations over the information gathered from users, the system computes data structures that are used for a variety of ranking and search operations. The disclosed techniques honor access restrictions that users specify for each item, so that only authorized people will see any item of information.

SHARE VIA SEARCH

[0057] Search is a common means by which users find items that have been shared with them. Figure 11 is a display page 1100 illustrating how the disclosed system allows a user to type in a search query and retrieve the most relevant items in accordance with an embodiment of the present technology. The system first identifies those items that match the query, and then *ranks* them based on how relevant the item is for that query and to that user. In this example, the user has searched for the query string 1110 "vpn", and the system has returned items that contain some reference to that phrase (e.g., an exact match or a partial match), ranked by relevance. Where possible, the system will include a highlighted set of text showing where in the description or contents of that item the query string appeared.

[0058] In some examples, ranking of items for a search query is implemented in the system as follows:

1. Determine which items are both accessible to the user and contain words or strings from the search query. This process may be performed using a software package, such as the Lucene software library supported by the Apache Software Foundation. As described below, Lucene computes and uses an *inverted index* that reports, for every *lexeme* in the search query, the items that contain that lexeme. This reported set of items is the *candidate set* of items.
2. Compute a *textual rank* for each candidate item (i.e., an item in the candidate set) based on a *similarity* algorithm. This rank is based on the textual contents of the items and on the search phrase.
3. Run a computation (i.e., a subroutine) called **ItemRanker** (described below) on the candidate set to assign an overall rank to each item. The computation relies on the interest graph and the textual rank to compute the overall rank.
4. Present the resulting items to the user, sorted in overall rank order (highest value first).

SHARE VIA BROWSING

[0059] In order to find useful and compelling content online, the system allows users to browse through organized collections of content. In some examples of the disclosed system, the collections are called *spots,* which can be further organized by placing items of content into *spotlists,* or lists of items. An item can be placed into any number of spotlists. Spotlists can also be gathered into *folders*. A user can browse content in many ways, including but limited to: viewing a directory of spots, finding a link to the collection on another spot, having the system suggest a collection, searching, having a link shared with them by another user, and so on. Users can, for example, look at the spot as a whole or look at a sub-collection of the spot by choosing a spotlist or a folder. Figure 12 shows a spot *home page* 1200, which is the default view in the system when a user visits a spot in some embodiments. On the left, under the word "Browse", a set of spotlists and folders are presented. For example, the spotlist "Implementor" 1205 is a spotlist with 48 items in it and is in a folder called "Audience" 1210. If the user clicks on a spotlist

(e.g., "Implementor," "Decision Maker," "Proof of Concept"), the system displays the items in that list.

[0060] At any given time, there is a current set of results, and these are by default presented to the user in *relevance* order. That order is computed by the interest graph, as described below, can be customized for each user, and is ordered based on what the system knows about that user's interests.

[0061] In some examples, the browsing experience is presented to users through a Web experience, as shown in Figure 12. One of ordinary skill in the art will recognize that the browsing experience can also be presented as an online or offline custom application experience on a PC or on a device (such as a phone or tablet).

[0062] In some examples, the ranking of items during browsing is implemented in the system as follows:

- Determine the set of items in the current *result set*. This may be based on the user choosing to look at a collection of items (in the system, this is done by visiting a particular spot), or a subset of items (by clicking on a spotlist or folder), or performing a search operation. In the example shown in Figure 12, the user can also use the selector 1220 ("All Items") in the upper left of the screen to see all the items, only the ones that were last added, or all the ones added during the past week.
- Run a computation (i.e., a subroutine) called **ItemRanker** (described below) on the result set to assign an overall rank to each item. The computation uses the interest graph to compute a rank for each item indicating how interesting it is expected to be to the current user (i.e., the user for whom the items are being ranked).
    - Note that the activity counts used by **ItemRanker**, which are stored in the **ItemScore** data structure, can be multiplied by a fractional value each day. This has the effect of causing older actions to be weighted less than more recent ones (producing exponential effect based on the age in days of the activity). The fractional value determines how quickly the usefulness of older activity attenuates. There are many other algorithms that could be used to calculate the recency of an item, including (but not limited to) using a fixed or adjustable linear scale factor based on the age of the activity, discarding data that is older

than a certain window of time, or attenuating the data for a period of time and treating all older activity as equal in weight and hence in importance. Also note that it is possible to use different age algorithms for different computations. For example, the system offers an alternative sort order for items on a spot that biases more for recent activity, by attenuating the weight of older activity more aggressively than it normally does. It is also possible to use different weights in the ranking algorithm for different computations or sort orders.

- Present the resulting items to the user, sorted in overall relevance rank order (highest value first).

SHARE VIA DISCOVERY

[0063] Another way that users find information that has been shared with them is via *discovery*. The system can automatically suggest items to a user that the user is likely to find interesting, based on what the system knows about the user. Figure 13 is a display page 1300 illustrating an *item feed,* a sequence of items, and how users have interacted with those items in accordance with an embodiment of the disclosed technology. For example, the item 1310 has been viewed 25,000 times (the number 1315 next to the icon shaped like an eye), has inspired one comment, and so forth. Also note in the upper left area of the feed, the system suggests people and spots (collections of items) that may be of interest 1320.

[0064] In some examples, the system determines what to put in the feed as follows:

1. Determine which items are accessible to the user. This process may be performed using a software package, such as the Lucene software library supported by the Apache Software Foundation. As described below, Lucene computes and uses an *index* that references items in the system and enumerates a set of people who have access to those items. The set of items accessible to the user is the *candidate set.*
2. Run a computation (i.e., a subroutine) called **ItemRanker** (described below) on the candidate set to assign an overall rank to each item. The computation

relies on the interest graph to compute a rank for each item indicating how interesting it is expected to be to the current user.

3. Present the resulting items to the user, sorted in overall rank order (highest value first).

MEASURING CONSUMPTION, ENGAGEMENT, AND INFLUENCE

[0065] In some embodiments, the disclosed system allows a publisher to measure the degree to which shared items have been consumed, how engaged members of the community are around the information, and how much influence community members based on information sharing. Figure 14 is a display page 1400 illustrating statistics for a collection in accordance with an embodiment of the disclosed technology. The summary header 1401 indicates how much activity has been generated by this collection. For example, items from this collection have been viewed 1001 times in the last 7 days and 275,000 times over the lifetime of this collection, or spot. Furthermore, this spot has been shared with 21 users. Rows below the summary header show items in the collection and how much activity each has garnered from the group of users with permission to access it. For example, row 1402 shows that the item labeled "There and Back Again…A Hobbit's Tale" has been viewed 789 times and that 35 actions have been performed with respect to this item, including 15 comments, 3 likes, and 10 downloads.

[0066] Figure 15 is a display page 1500 illustrating statistics for an individual in accordance with an embodiment of the disclosed technology. The summary header 1501 shows a basic measure of the community influence and engagement that the corresponding user has (e.g., how many followers the user has and how many other users the user follows). Rows below the summary header show the activities they have performed on various items (which they have commented on, etc.), such as likes, respots, and comments. In addition to the basic statistics, the system may also compute complex measures of influence.

[0067] In some examples, the system measures these statistics as follows:

1. Using the **ItemScore** data structure (described in detail below), report activities performed by a particular user on any item. These are captured by the values in the row U = current user ID (e.g., a row in the data structure

corresponding to the current user). The system allows a publisher to sort users based on their activity, revealing the most and least active members.

2. Report on activities performed by any user on a particular item by looking at the values in the column I = item ID (e.g., a column in the data structure corresponding to the item having the item ID). This allows the publisher to determine which items are the most actively consumed and which are the least. The system can sort based on total number of activities or any particular type of activity (e.g., number of views, number of downloads, etc.).
3. Sort the list of users based on the **UserInfluence** data structure, described below. This will identify the most and least followed and connected users in the community.

OTHER POTENTIAL USES

[0068] There are a number of other ways that the system can support sharing and the interest graph powering it can be enhanced in various embodiments of the disclosed technology.

- **Auto-categorize content**. Based on content analysis and behavioral information, the system could suggest which items have affinity for one another (e.g., items for which interactions are often correlated, items that are regularly viewed together or in sequence, and so on) and hence would make sense to categorize together. The system could also automatically categorize content without user involvement.
- **Identify characteristics of high and low value content**. The system could do feature analysis to discover what popular items (or unpopular items) have in common and report those commonalities to a user to help the user determine what to share. For example, type of item (presentation vs. document vs. video), length of item, language complexity, role of author, level of categorization, language, or other measurable characteristics. The analysis can be segmented by audience – users in a certain location, or with a certain title, or belonging to a certain group.
- **A/B testing**. The system could allow publishers to set up sharing in different ways and measure effectiveness. There might be multiple versions of an item, or multiple ways to organize items. By presenting these different options

to different subsets of the users (one subset would be shown the "A" version, and other would see "B"), the publisher could determine which option led to the greatest level of engagement by users.

- **Versioning and expiration.** By detecting when items have identical or similar content (e.g., two presentations with multiple slides in common, or a document that shares most of its content with another), the system can detect that one item has been updated or has expired and hence the related ones may also be out of date or may need to be updated as well. For identical items, it could update all copies of the content automatically.
- **Recommended sharing.** Based on the content of an item or a collection of items, and the relationships among users, the system could recommend people to include in permission lists for collections. It could also recommend whom to share individual items with.
- **Identify missing content needed by the community**. By looking at the most common search queries that have yielded few to no results, or yielded results that have had few to no views, the system can identify unsatisfied needs of the community. A lack of search results often indicates that the needed material is missing or poorly organized. Where users are not clicking through and consuming the results, the items available may not be satisfying user needs.
- **Augment the interest graph computation with real-time feedback**. Regularly adjust the algorithms used to compute the interest graph and the weightings based on the effect on user behavior. For example, divide the user community into groups (sometimes called *cohorts*), present the results of different weights or algorithms to the different groups, and adjust the system based on the most successful approach.
- **Track more granular user behavior**. Follow the mouse movements of a user to get additional information about their interests.
- **Voting and reputation**. By providing ways for users to vote on the contributions of others, the voting patterns resemble commenting patterns, and can be added into the data structures that track interactions between users and measure the influence of a user on others in the community. For example, if one user votes for the contribution of another, the system can treat

that the same way as making a comment. Each such vote represents an additional amount of influence that can be added to other indications, such as the number of followers they have or the number of times others commented on an item that a user uploaded.

- **Rank subsets of an item**. Based on the sections of video that other users have watched or otherwise interacted with, and the affinity of the current user to the other users, the system can identify and present the sections within the video most likely to be of interest. Similarly with other media types, like audio, or with components of structured documents, like a CAD diagram, the system can identify a section or sections thereof most likely to be of interest to a particular user. Similarly, the system might rank the pages of a document or the slides in a presentation in terms of which are most likely to interest the user based on the extent to which users have viewed or interacted with portions thereof. These ranks could be used to create a digest (summary) of the item that is faster for users to review, such as a "highlights" synopsis of a video, or a summary document.
- **Search for derived information**. In addition to indexing the contents of an item, the system can apply a variety of computations that derive new information based on the content, and apply the interest graph to searching that derived data. For example, the system can perform voice recognition of an audio or video stream and search the text that it has computed via the voice recognition. The system may perform a dependency analysis on a CAD diagram and include dependent components in the feed, even though they are not present in the original diagram. Each of the derived items may be treated as a separate content item for ranking and affinity calculation purposes. The system may perform image recognition and search for the names or characteristics of objects and people that have been recognized.
- **Incorporate additional sources of social graph information**. Patterns of email for each user provide an indication of the topics that they are interested in and the people with whom they interact most frequently. The interest graph could be enhanced by performing content and traffic analysis on email and on aliases that have been set up by users to communicate with each other. Another example is the user directory. Many organizations maintain

directories that maintain the membership of users in groups and the relationship between groups. A common example is Active Directory. Such directories are used to enforce access permissions, set up email aliases, and a variety of other purposes. There are also HR and payroll systems that capture relationships among users as well. Another source of information is on-premise server systems. For example, by looking at the permission sets for file systems and the access patterns, additional social information would be available that is similar to what the system tracks internally. Another potential source is public social networks outside of the organization (such as Facebook, Twitter, Instagram, and the like). All of these systems have information about the social graph that could be incorporated into the interest graph. For example, if one user "follows" another person in the company on Twitter, that can be treated similarly to following somebody within the system and incorporated into the interest graph. Likewise when two people are "friends" on Facebook, or otherwise connected on a social networking site, this connection can be included in the interest graph. As another example, comments by one person on an item posted by another, which is similar to commenting on an item published within the system, can be incorporated into the interest graph.

- **Use role/title information**. If users are assigned roles or titles, either by self-selection or by the company, those with similar roles and titles are likely to have some interests in common - another signal that can be incorporated into the interest graph.
- **Identify similar items as well as true duplicates**. Often items are related but not identical. For example, a slide deck that contains some of the same slides as another, but reordered or intermixed with others. Or, a document that has been modified, but much of the substance is the same. Similarity could be used to do more aggressive deduplication, or to identify when content is related, or to identify when users have similar interests because they interact with similar documents as well as identical ones. Another option is to offer users the ability to construct new content from pieces of existing ones. For example, assembling a presentation from slides that come from multiple existing decks. This would reveal the relationships between the

assembled and original decks, and would give additional signal on which slides of a presentation are the most valuable.

- **Additional types of content analysis.** There are a variety of ways to analyze content to reveal information that would be useful for the interest graph. For example, performing facial and feature recognition of images. The analysis could be used to find items that are about related topics. Facial recognition could be used to compare content images to user profile pictures, to identify content that is about people in the system. The system could automatically perform language translation on items of particular interest. Or it could create higher resolution previews, or graphical rendering/visualization of data, or create 3D images. The system could automatically perform language translation on items of particular interest or create higher resolution previews, graphical rendering/visualization of data, or 3D images using known techniques.
- **Proactively get ready to deliver items likely to be of interest**. Items likely to be of interest to a user can be downloaded to a device for optimized or offline access, or pre-cached in a content delivery network (CDN) for faster delivery to the browser. A group of items might be presented to the user together for faster viewing.
- **Target advertisements and promotional offers.** The system may target offers to users based on their activities and interests (e.g., the items they are currently browsing). These might be sponsored by the organization they work for, by a partner, or by outside companies. For example, a company might allow vendors to compete for or purchase the right to market services to employees. Similarly, there might be a facility for "sponsoring" particular items and ranking sponsored items higher, highlighting sponsored items visually, indicating the nature of the sponsorship, and so on. Such sponsorship might be done manually, by an algorithm, by a business rule, by an expert system.
- **Instant search.** The system can present search results incrementally as the user is typing, rather than waiting for them to specify a full query.
- **Semantic search.** Search queries can be semantically analyzed using techniques like latent semantic analysis and a variety of natural language

processing algorithms that perform operations, such as relationship extraction, named entity recognition, and the like. Then, the system can do specialized operations appropriate for a particular domain or a particular semantic concept. For example, if the system determined that a search applied to a set of legal cases, it might automatically detect references to legal precedents and search through them as well as through the case itself. In manufacturing, the system could identify that a number was a reference to a part and extend its search to include the supporting information for that particular part.

CREATING THE INTEREST GRAPH

[0069] The choice and ordering of information items relies on the interest graph. Figure 2 is a block diagram illustrating process of the system for creating an interest graph in accordance with some embodiments of the disclosed technology. In some examples, the process of building the interest graph includes generating the data, gathering the data, and computing the interest graph.

STEP 1: GENERATING THE DATA

[0070] In some examples, an interest graph is computed from a number of different data sources and benefits greatly from having additional data to analyze. Machine learning research and practice consistently shows that accuracy improves as the *number* of data sources and the *amount* of data increases. This is referred to as *user signal*.

[0071] Therefore, step 1 is *generating* the data, which means encouraging users to engage in activities that generate signal. Historically, activities that provide the most useful data have been overly complex inside of companies, and hence have not occurred as often as they otherwise might.

[0072] For example, *sharing files with others* in a rich online experience (like a web site that offers a structured view, supports search, and enables browsing) has been cumbersome to set up. As a result, people often settle for simple sharing solutions, such as relying on email attachments or on keeping their files in a shared disk drive. The disclosed system provides a simple and easy-to-use sharing solution

that encourages users to interact more heavily with each other's information and hence to generate more signal.

[0073] *Browsing files* on a web site generally involves downloading them to the local computer and viewing them in a program like Microsoft Word or PowerPoint, which is quite slow. Accordingly, users are discouraged from browsing as many items as they might otherwise do. The disclosed system provides a much faster way to browse (called "*skim" preview)*, which offers very fast viewing of items and collections of items. Skim allows users to explore information online without requiring them to download anything or launch any applications on their machine, encouraging far more browsing. Skim preview works by tracking the way that the user slides their mouse across the item's thumbnail. Based on how far the mouse has moved horizontally across the thumbnail, a preview of that part of the item is shown. For example, if the user is running the mouse over the thumbnail for a presentation, as the mouse moves left to right, each slide of the presentation is shown in succession. By sliding the mouse back and forth, at any desired speed, the user can quickly view all the slides. Similarly, for a document, the thumbnails show each page of the document. There is an equivalent browsing experience for each type of information supported by the system. In seconds, the user can see every part of the item - it is much faster than the traditional method of downloading the file to a client application.

[0074] Another example is *organizing information*. The traditional approach is to use a directory structure, which provides a limited way to establish a taxonomy and to associate related files. Another approach is to use metadata tagging, where items are assigned a set of properties. These systems have been deployed extensively within companies and are generally felt to be rigid and awkward - most users resist them and the vast majority of information is never put into them. The disclosed system offers lists and folders that support dragging and dropping items into multiple places, a model that is familiar to users from other domains like organizing music into playlists. The system offers three levels of hierarchy: (1) *spots*, which are collections of items that can be found via a directory or search, (2) *folders,* which exist within a spot and optionally allow users to group a set of lists together, and (3) *lists*, which are simple groups of items. An item can be in zero, one, or many different lists. Users can place individual items into lists or can drag a

group into a list. This is a much simpler structuring model than is traditionally used by systems like enterprise content managers. Each user can create their own hierarchy, if they wish, and can take an item from one spot and put it into another one (using an operation called *respot*). So users might create a spot called "Widget Marketing", which contains the marketing material for widgets. Within that spot, they might have a folder called "vertical markets" containing lists, such as "manufacturing", "media", etc. They might have another folder called "sales stage" with lists, such as "pre-sale", "proof-of-concept", "post-sale." Any piece of information can be put into any number of lists, allowing for a flexible browsing experience based on spots, folders, and lists.

[0075] The first step towards creating an effective interest graph is to provide an information management environment that makes it much easier and faster for users to engage in useful data-generating activities and generate user signal to be analyzed.

STEP 2: GATHERING THE DATA

[0076] The next step is to *gather* the data. Producing an accurate interest graph relies on detailed analysis of data from a variety of sources. Table 1, at the bottom of this section, lists and defines input data structures used by the system.

USER BEHAVIOR

[0077] A source of data is the way that users interact with each piece of information. The system tracks actions that a user performs on any item (share, download, copy from one collection to another, recommend, comment, etc.) and monitors how much time they spend looking at each part of a document, presentation, video, training program, or the like.

[0078] Traditional content systems invoke other programs when users wish to view the contents of a document – for example, such an environment might download a presentation and invoke Microsoft PowerPoint to let the user read it. What users do inside of a program like PowerPoint is usually opaque to the content manager. And, most such editing programs (e.g., word processors or presentation programs) do not track and report which parts of the file users spend time on, and

how much time. Therefore user engagement with each piece of information does not generate any signal that can be analyzed.

[0079] The disclosed system presents high resolution previews and views of various document types that are available online and, in some embodiments, can be quickly browsed using skim preview – which can be accomplished in the web browser, so that no additional software download is required, and no software applications need to be installed or invoked on the user's machine other than the web browser. The system monitors views and previews, tracking how often they happen and how long the user spends looking at any part of the item.

[0080] The actions that users have taken on items and their viewing behavior are captured in the **ItemScore**, **CollectionScore**, and **RecentActivity** data structures. In addition, the system creates a feedback loop – whenever it presents items that might be of interest to the user, the click-through behavior is tracked in **ClickThroughs**.

[0081] Item Analysis

[0082] The system extracts data by analyzing each item of information:

- In some examples, the system uses an information retrieval library, such as the Lucene software package supported by the Apache Software Foundation, to parse text, apply Porter stemming analysis, create an inverted index, and compute a similarity score for a query string against the index. The index tracks the number of times each word appears and also records collections of words that appear together, to support searching for phrases. Each word in the index is *stemmed,* meaning that it is divided into its component parts. This allows, for example, a search for the word "run" to match a document that contains "running." Note that one of ordinary skill in the art will recognize that there are a variety of other algorithms for stemming (e.g., suffix-stripping and lemmatization), assembly of the index (e.g., a suffix tree or n-gram tree), and scoring a query (e.g., compression distance, Dice's coefficient) that would also serve. This information is stored in **InvertedIndex**.
- For each item, the system computes a content vector that expresses how many times any particular word appeared in it. The result is stored in **ContentVectors.**

- Each piece of metadata is extracted - documents created within Microsoft Office, for example, have a section that captures tags like the author, date, description, and so forth. A similar model exists for images in JPEG format and for many other file types. The resulting <field name, value> pairs are added to **InvertedIndex**.
- For each piece of information, the system computes a large *hash* function of the contents of the document (using, for example, the SHA-256 algorithm, although there are a variety of cryptographic hash functions with low collision rates that would also serve). The hash is, with high probability, unique for each piece of content in the system and allows the system to quickly recognize when the same item has been added to the system multiple times, by the same or by different users. The hashes are stored in **ItemHashCodes**.

SOCIAL GRAPH

[0083] Another valuable clue to user interest is the set of people to whom they are connected. The system computes the *social graph,* which captures the connections between people. Such connections can take many different forms; for example:

- They may both belong to the same group of users.
- They may both have similar access permissions to a collection of items. The strength of this indicator is inversely proportional to the number of other people who have similar permissions. In other words, if only two people have access to a body of documents, that is a much stronger indicator of mutual interest than if two people have access to information that is also available to hundreds or thousands of other people.
- A user A may choose to *follow* another user B, which means that user A will be notified when user B performs certain kinds of actions. This creates an *asymmetrical* connection – user A is likely to be interested in something that user B cares about, but it is weaker evidence that user B will share interests with user A.
- A user may own a collection of information and grant access to another.
- A user may invite another user to join the service - accepting that invitation represents a stronger connection than simply receiving it.

- A user may have created a *link* to another user.

[0084] The system examines the social graph, distilling it into **UserConnectedness**.

INFORMATION GRAPH

[0085] The system has a variety of ways that information can be categorized – it provides a hierarchy of collections and any piece of information can be in any number of those collections. One collection may have a *link* to another. As a result, there is also an *information graph* capturing the relationships between items of information. The system stores that graph in the **ItemConnectedness** data structure. Different types of collections imply different levels of relationship between the items.

[0086] Similarly, the system aggregates these individual relationships between items into a measure of connectedness between collections, stored in **CollectionConnectedness**.

QUERIES

[0087] The system offers search, both within a collection and across many of them. There is valuable information in the phrases that users search on, and their subsequent decisions whether or not to click through on the results presented. The system keeps track of queries that have been performed in **QueryCount**, the ones that are most popular (e.g., top 10, top 20%, top 15 in the past 24 hours) in **PopularQueries**, and the subsequent click-through decisions by users in **ClickThroughs**.

Table 1: Input Data Structures

ItemScore – total activity applied to an item by each user ItemScore is an array [U, I] of tuples, where U is the number of users in the system and I is the number of items. Each tuple = <*weightedsum*, <$action_1$, $action_2$, .., $action_n$>, *views*, <$preview_1$, $preview_2$, …, $preview_m$>> The tuple contains a count of each allowed type of *action* for an item (e.g., "downloaded"), a count of the number of times it is viewed, and a count of the amount of time each part of it (e.g., a page of a document) was previewed. The

tuple also contains a weighted sum of these counts; weights are adjusted depending on the relative importance of each of the counts.
CollectionScore – total activity applied to a collection of items by each user CollectionScore is an array [U, C] of *element,* where U is the number of users in the system and C is the number of collections. Each element is the same tuple as for ItemScore.
RecentActivity – a log of recent activities each user has done with every item RecentActivity is an array [U, I] of tuples, where U is the number of users and I is the number of items. Each tuple = <<$action_1$, $timestamp_1$>, <$action_2$, $timestamp_2$>, … <$action_n$, $timestamp_n$>> The tuple is the set of recent actions performed by the user on the item, each with a time stamp.
ClickThroughs – a log of the result when each item was presented to each user Clickthroughs is an array [U, I] of tuples, where U is the number of users and I is the number of items. Each tuple = <<*context*, $position_1$, $click_number_1$>, …> The tuple contains the set of times this item was presented to this user. The system records the context (e.g., "search query"), the position of the item in the presented list (e.g., "the item was the third result"), and which of the selected items from that result set it was (e.g., "the item was selected second" or "the item was never selected").
ContentVectors – a representation of the content of every document. In some examples, the system uses the Mahout software package developed by the Apache Software Foundation to create a normalized vector space model (VSM) representation for every item, using term-frequency inverse document frequency (TF-IDF) weighting to compute the values in each vector. Collocation-based n-gram analysis with log-likelihood ratio test improves the accuracy of the weighting. There are other algorithms for vectorizing content that would also serve. ContentVectors is an array [I, T] of *values*, where I is the number of items and T is the number of n-gram terms that appear in any of those items. The value is a weighted count of the number of times that term appears in that item.
InvertedIndex – an index of a set of documents In some examples, the disclosed system uses the Lucene indexing package to create an inverted index from a set of documents. This index contains every lexeme that appears in any item. For each lexeme, Lucene enumerates the set of documents that contain the lexeme. Each document is also annotated to reflect the set of individuals who are allowed to access it, and the Lucene search contains a mask to choose those items that are visible to the user.
ItemHashCodes – a pointer to the items corresponding to any hash code present in the system

The system computes a cryptographic hash value of the contents of every item. In some examples, the system uses SHA-256, but there are a variety of other algorithms that similarly compute a value for any item that has a very low probability of colliding with the value for any other. ItemHashCodes is an array [H] of item lists, where H is the number of unique hash values present across all items. List contains the set of items that correspond to that hash value.
UserConnectedness - the degree to which each user is connected to every other user in the social graph UserConnectedness is an array [U, U] of tuples, where U is the number of users. Each tuple = <*weightedsum, <<strength$_1$, type$_1$>, <strength$_2$, type$_2$>, …>>* The tuple enumerates the strength and type of each connection between this pair of users (from X -> Y, if the tuple is element [X, Y] of the array). The type might be "appear in the same access control list" or "X invited Y to join the community and that invitation was accepted." The strength can be the same for every connection of a particular type or it can be weighted (e.g., "the value is one divided by the number of people on the shared access control list"). The system computes a weighted sum across the connections, factoring in their strengths.
ItemConnectedness - the degree to which every item is connected in the information graph to every other item. ItemConnectedness is an array [I, I] of tuples, where I is the number of items. The tuple has the same form as the one for UserConnectedness.
CollectionConnectedness – the degree to which each collection of information is connected to every other collection. CollectionConnectedness is an array [C, C] of tuples, where C is the number of collections. The tuple has the same form as the one for UserConnectedness.
QueryCount – the queries that have been executed QueryCount is an array [Q, U] of tuples, where Q is the number of queries that have been executed on the system and U is the number of users. Each tuple = <*querystring, count, <<clickeditem$_1$, click$_1$>, <clickeditem$_2$, click$_2$>, …>* The tuple expresses the number of times that user U has executed query Q. *querystring* is the text of the query, *count* is the number of times the query was executed, and the next value is the set of results from those queries. Each item in the set is a pair - the item that was clicked, and its position in the clickstream of user choices (e.g., "first item clicked", "second item clicked", etc.).

STEP 3: COMPUTING THE INTEREST GRAPH

[0088] In some examples, the system computes the interest graph by taking the raw user signal (captured in the input data structures described in the previous section) and processing that data through a series of *intermediate computations*.

[0089] Each of the intermediate computations is called "Compute <X>", where <X> is the name of the output that it generates. For example, "**Compute UserUserAffinity"** produces the **UserUserAffinity** data structure. The system runs these intermediate computations at periodic intervals and the outputs are updated over time as additional user data is gathered. Table 2 enumerates the intermediate data structures that are produced by these algorithms.

Table 2: Intermediate Data Structures

UserInfluence – measure of how much social influence each user has on others UserInfluence [U] is an array of real numbers representing the influence of each of the U users in the system.
<X><Y>Affinity – a measurement of the affinity for every X to every Y These are a family of data structures that represent affinity – the likelihood of a user to be interested in another user, an item, or a collection, or the likelihood that an interest in one item implies an interest in another. In each case, affinity can be represented as a real number from 0 to 1 on a logarithmic scale, where 1 represents extremely strong predicted affinity and 0 represents none. Note that an alternative model is to make zero represent "no information," negative numbers represent negative affinity (the belief that an item of *not* likely to be of interest), and positive numbers represent positive affinity. • **UserUserAffinity** is an array [U, U] with affinity from every user to every other user • **UserItemAffinity** is an array [U, I] with affinity from every user to every item • **UserCollectionAffinity** is an array [U, C] with affinity from every user to every collection • **ItemItemAffinity** is an array [I, I] with affinity from every item to every other item
ItemClusters – divides the items in the system into clusters whose content is related ItemClusters is an array [I] of tuples, where I is the number of items. Each tuple = <<*cluster*$_1$, *membershipweight*$_1$>, <*cluster*$_2$, *membershipweight*$_2$>, …> The tuple enumerates the clusters that the item is in and the weight of the item's

membership to each cluster. In some examples, the system uses a non-uniform weight (so called "fuzzy clustering"), though it is also possible to make membership boolean.

[0090] When the system displays a set of values to the user, it invokes one of the *ranking computations*. In some examples, the names of these ranking computations takes the form "<Y> Ranker", depending on what kind of values they are ranking, where <Y> represents the kind of values being ranked (e.g., **RelatedItemRanker** ranks related items). Ranking computations are given an argument and then compute a set of ranked results based on that argument and on a set of other inputs.

[0091] Figure 2 is a block diagram illustrating processing of the system in some examples. Figure 2 shows the processing steps of the system and how the data flows through the system. Each named arrow represents an input data structure capturing raw user signal. Each rounded rectangle represents a computation. For example, "**Compute ItemClusters**" 210 is an intermediate computation with one input, the **ContentVectors** data structure. Its output (**ItemClusters**) is fed into the "**Compute ItemItemAffinity**" 215 computation, along with two other inputs – the **ItemConnectedness** and the **ItemScore** data structures.

[0092] The system uses the ranking computations to produce output that users can see. For example, suppose the user is looking at an item, and the system wants to display a set of related items next to it. The goal is to identify the items that are most likely to interest the user. For example, if a salesperson is looking at a presentation about a particular product, they might also be interested in a price sheet for the product, white papers on how to use that product most effectively, presentations and documents about related products that work with it, etc.

[0093] The system uses the ranking computation called **RelatedItemRanker 220** to identify and rank related items. When the user pulls up a particular item on a web site, the system hands that item to **RelatedItemRanker**, which returns the ranked set of items (in a **RankedItems** data structure) that it has identified as being most likely to be of interest to the user. The computation relies on one input data structure – the popularity of items (**ItemScore**) and the results from two intermediate computations - the likelihood that the current user would be interested in any

particular item (**UserItemAffinity**), and the degree of similarity between any two items (**ItemItemAffinity**).

[0094] The following data structures are used to hold groups of different types.

Table 3: Group Data Structures

<value>Set – a set of <values> This family of data structures holds an unordered set of items of type <value>. • **ItemSet** is an array [I] of items, **PeopleSet** is an array [P] of people, and **CollectionSet** is an array [C] of collections
Ranked<value> - a set of <values>, with an associated ranking This family of data structures holds a set of items of type <value> with an associated *rank* that represents an ordering. Note that ranks are real numbers, allowing the structure to both establish an ordering and to measure the "distance" between two items in terms of their rank. • **RankedItems** is an array [I] of ranked items, **RankedPeople** is an array [P] of ranked people, **RankedCollections** is an array [C] of collections, **RankedQueries** is an array [Q] of ranked queries, and **RankedActivities** is an array [A] of ranked activities

INTERMEDIATE COMPUTATIONS

[0095] These computations operate on input data structures and on the results produced by other intermediate computations. In each case, they produce a data structure as output with the results.

[0096] These functions or algorithms compute the degree of *affinity* between pairs of things. "Affinity" means the likelihood that interest in one of those items means interest in the other. Note that affinity is not symmetrical; a salesperson who is looking at a particular product description might be highly likely to look at the price sheet containing that product (among hundreds of others), but somebody looking at the price sheet is much less likely to care about any particular product's description.

COMPUTE ITEMCLUSTERS

[0097] This algorithm operates on **ContentVectors**, applying a clustering algorithm to compute **ItemClusters** that represent groups of items that have related

textual content. In some examples, the system uses the Mahout software package to perform this computation, applying canopy generation to identify cluster centroids, then using k-means clustering based on the cosine of the Euclidean distance between documents as a similarity metric. One of ordinary skill in the art will recognize that other clustering algorithms can be used.

COMPUTE ITEMITEMAFFINITY

[0098] This algorithm computes the degree of affinity between pairs of items in the system.

[0099] The inputs are **ItemConnectedness** (the degree to which the items are "close" in the information graph), **ItemScore** (the amount of interactions users have had with items), and **ItemClusters** (the degree to which the contents of items are related). Here is the algorithm:

```
Compute_ItemItemAffinity(ItemConnectedness, ItemScore, ItemClusters)
{
   FrequentGroups = AssociationRuleAnalysis(ItemScore)
   For every pair of items (I, J)
      ItemItemAffinity[I, J] = A * ItemConnectedness [I, J] +
                               B * ItemScore [*, J].weightedsum +
                               C * number of appearances of I & J in
FrequentGroups
}
```

[00100] AssociationRuleAnalysis determines which pairs of items are frequently viewed together. In some examples, the system uses the algorithm known as Apriori to determine these pairs. One of ordinary skill in the art will recognize that there are a variety of similar algorithms that could also be used. The weighting parameters A, B, and C allow the system to balance the importance of items being placed in related collections, the popularity of particular items with users, and the degree to which other users have viewed both items.

COMPUTE USERUSERAFFINITY 235

[00101] This algorithm computes the degree of affinity between pairs of users – the likelihood that each user is interested in what the other one does. The inputs are **ItemScore** (which captures how users have interacted with items) and

UserConnectedness (the degree to which they are connected in the social graph). The algorithm is:

```
Compute_UserUserAffinity(ItemScore, UserConnectedness)
{
  UserBehaviorSimilarity = PearsonCorrelation(ItemScore)
  For every pair of users (I, J)
    UserUserAffinity[I, J] = A * UserBehaviorSimilarity [I, J] +
                             B * tanh(UserConnectedness [I, J])
}
```

[00102] The system uses, for example, the Mahout software to compute the Pearson correlation of behavior across the weighted sum of item scores. The user connectedness value is normalized into the range 0-1 using hyperbolic tangent. Then the values are weighted, to reflect the relative importance of behavior vs. the social graph. The weighting parameters A and B allow the system to balance the importance of these values. Note that one of ordinary skill in the art will recognize that numerous other algorithms can be used to compute behavioral similarity (e.g., Euclidean distance or the Tanimoto Coefficient) and normalization (e.g., the logistic function or Z-scores).

COMPUTE USERITEMAFFINITY 245

[00103] This algorithm computes the degree of affinity between every user and every item in the system. The inputs are **UserUserAffinity** (from above), **ItemScore**, and **ItemConnectedness**. The algorithm is:

```
Compute_UserItemAffinity(UserUserAffinity, ItemScore, ItemConnectedness)
{
  For every item I, for every user U {
    ActivitySum = UserInterest = 0
    For every user U2
      ActivitySum += UserUserAffinity[U, U2] * ItemScore[I, U2].weightedsum
    For every item I2
      UserInterest += ItemScore[I2, U] * tanh(ItemConnectedness [I, I2])
    UserItemAffinity[U,I] = A * ActivitySum + B * UserInterest
  }
}
```

[00104] The system computes the sum of the activity that other users have performed on the item (weighted by affinity to those users) and the sum of item activities that the current user has performed (weighted by the affinity of the current

item to those other items). Those two values are combined in a weighted sum, based on the relative importance of behavior vs. item connectivity. In some examples, connectedness is normalized using hyperbolic tangent, but one of ordinary skill in the art will recognize that other algorithms could be used.

COMPUTE USERCOLLECTIONAFFINITY 255

[00105] This algorithm computes the degree of affinity between every user and every collection, where a collection is a grouping of items. Note that collections can overlap, can be organized into a hierarchy, or can be disjoint – the model works in any of those cases. The inputs are **UserUserAffinity** (from above), **CollectionConnectedness** (the degree to which collections are connected), **ItemHashCodes** (the hash values of every item), and **CollectionScore** (the activities user have performed on each collection). The algorithm is:

```
Compute_UserCollectionAffinity(UserUserAffinity, CollectionConnectedness,
                                                ItemHashCodes, CollectionScore)
{
  For every collection C, for every collection C2 {
    For every item I in C, for every item I2 in C2
      if (ItemHashCode[I] = ItemHashCode[I2])
        CollectionSimilarity [C, C2] += SharedItemWeight
    }
  For every collection C, for every user U {
    ActivitySum = UserInterest = 0
    For every user U2
      ActivitySum += UserUserAffinity[U, U2] * CollectionScore[C,
U2].weightedsum
    For every collection C2
      UserInterest += CollectionScore[C2, U] *
                          (tanh(CollectionConnectedness [C, C2]) +
                           CollectionSimilarity [C, C2])
    UserCollectionAffinity[U,C] = A * ActivitySum + B * UserInterest
  }
}
```

[00106] The system computes the frequency with which the same item appears in every pair of collections, using a constant weight. The system then computes the sum of the activity other users have performed on the collection (weighted by the affinity to those users) and the sum of collection activities that the current user has performed (weighted by the affinity of the current collection to those collections

based on both behavior and similarity of content). Note that connectedness is normalized using hyperbolic tangent, but other algorithms could be used. These values are then combined in a weighted sum, where the weights reflect the relative importance of user behavioral similarity vs. structural relationships and similarity of content.

COMPUTE USERQUERYAFFINITY 265

[00107] This algorithm computes the degree of affinity between every user and every query that has been executed on the system. The inputs are **UserUserAffinity** (from above) and **QueryCount** (a summary of the queries that have been executed by each user). The algorithm is:

```
Compute_UserQueryAffinity(UserUserAffinity, QueryCount)
{
  For every query Q, for every user U {
    ActivitySum = 0
    For every user U2
      ActivitySum += UserUserAffinity [U, U2] * QueryCount[Q, U2].count
    UserQueryAffinity[Q, U] = A * ActivitySum
  }
}
```

[00108] The system computes the sum of the number of times other users have executed this particular query, weighted by the affinity with that other user. The result is then multiplied by a weight to compute affinity for this user and the query.

COMPUTE USERINFLUENCE

[00109] This algorithm computes the amount of influence that each User has within the community of users on the system. Its inputs are **UserConnectedness** (the degree of connectivity in the social graph), and **ItemScore**. The algorithm is:

```
Compute_UserInfluence(UserConnectedness, ItemScore)
{
  For every user U, for every user U2
    UserInfluence[U] += A * UserConnectedness.weightedsum[U, U2]

  For every user U, for every item I that user U was responsible for creating
    For every user U2
      UserInfluence[U] += B * ItemScore[I, U2].weightedsum
}
```

[00110] The system computes a weighted sum of how connected other users are to a particular user, and for how much activity has been generated by the items that the particular user created.

RANKING COMPUTATIONS

[00111] The *ranking computations* produce ranked lists of items; a typical use for ranking computations is to produce lists that are displayed to users in various contexts. For example, **ItemRanker** is used in deciding which items to display to users as the result of a search query. **ItemRanker** takes candidate items that might match the query, and orders them appropriately.

[00112] Each ranking computation is invoked on an input. Using that input and data structures that are passed to it (per the workflow in Figure 2), the computation produces a ranked set as the output.

RELATED ITEM RANKER 220

[00113] This algorithm is invoked on an item and also gets **ItemScore**, **ItemItemAffinity**, and **UserItemAffinity**. The algorithm is:

```
RelatedItemRanker(Item, ItemScore, ItemItemAffinity, UserItemAffinity)
{
  For each item I
    Score = 0
    For each user U
      Score += ItemScore[I, U].weightedsum
    RankedItems[I].rank = (A * Score) * (1 + ItemItemAffinity[Item, I]) +
                                  (B * UserItemAffinity[CurrentUser, I])
}
```

[00114] The system finds the items most related to Item by computing a weighted sum. The factors are the total amount of user activity against other items, weighted by the affinity of those other items to this one, and the current user's affinity to the item.

ACTIVITY RANKER 240

[00115] When this algorithm is invoked, it is optionally given an item and also gets **RecentActivity** (the set of activities that have recently been performed on the system, such as the set of activities performed during the last year, month, week,

day, hour, or portion thereof), **UserUserAffinity,** and **UserItemAffinity.** If an item is provided, it returns the set of activities that have been performed on that item, ranked in terms of how likely they are to interest the current user. If no item is provided, it returns the list of activities on any item in the system, ranked in terms of how likely they are to interest the current user. The algorithm is:

```
ActivityRanker(optional: Item, RecentActivity, UserUserAffinity, UserItemAffinity)
{
  if Item was provided
    RankedActivities = set of activities in RecentActivity performed on Item
  else
    RankedActivities = RecentActivity
  For each activity A in RankedActivities
    RankedActivities[A].rank = B * ActivityValue(A) *
                               (C * (1 + UserUserAffinity[CurrentUser, A.user])) *
                               (D * (1 + UserItemAffinity[CurrentUser, A.item]))
}
```

[00116] The system chooses a candidate set of activities. For each activity in the candidate set of activities, the system computes a ranking using a weighted product of the intrinsic interest for that type of activity, the affinity of the current user with the user who performed the activity, and the affinity of the current user for the item on which the activity was performed.

ITEM RANKER

[00117] This algorithm is invoked on a set of items, which is either unranked (an **ItemSet**) or already ranked with a preliminary ranking (a **RankedItems** set) and also gets **ItemScore**, **ItemHashCodes**, and **UserItemAffinity**. The algorithm is:

```
ItemRanker(InputSet, ItemScore, ItemHashCodes, UserItemAffinity,
UserUserAffinity)
{
  Remove duplicate items from InputSet (using ItemHashCodes)
  For every item I in InputSet
    For every user U
      Score += A * ItemScore[I, U] * (1 + UserUserAffinity[CurrentUser,U])
    RankedItems[I].rank = (B * Score) * (C * (1 + UserItemAffinity[I,
CurrentUser]))
    If InputSet is ranked
      RankedItems[I].rank *= D * InputSet[I].rank
}
```

[00118] The system computes the sum of user actions against each item in the set, weighted by the affinity of the current user to the other users and then computes the weighted product of that sum, the affinity of the user to the item, and the existing rank of each item (if it was provided). The weights reflect the relative importance of user behavior directly against the items vs. the predictability of user interest vs. the effectiveness of the original input ranking. The output is a ranking for each unique item in the set.

COLLECTION RANKER 260

[00119] This algorithm is invoked on a set of collections, which is either unranked (a **CollectionSet**) or ranked (a **RankedCollections** set) and also gets **CollectionScore**. The algorithm is:

```
CollectionRanker(InputSet, CollectionScore, UserUserAffinity,
UserCollectionAffinity)
{
  For every collection C in InputSet {
    Score = 0
    For every user U
      Score += A * CollectionScore[C, U] * (1 + UserUserAffinity[CurrentUser, U])
    RankedCollections[C].rank = (B * Score) *
                                (D * (1 + UserCollectionAffinity[I, CurrentUser))
    if InputSet is ranked
      RankedCollections[C].rank *= E * InputSet[C].rank
  }
}
```

[00120] The system computes the sum of user actions against each collection, weighted by the affinity of the current user to the other users and then computes the weighted product of that sum, the affinity of the user to the collection, and the existing rank of each collection (if it was provided). The weights reflect the relative importance of user behavior directly against the collections vs. the predictability of user interest vs. the effectiveness of the original collection ranking. The output is a ranking for each collection in the input set.

PEOPLE RANKER 230

[00121] This algorithm is invoked on a set of people, which is either unranked (a **PeopleSet**) or ranked (a **RankedPeople** set) and also gets **UserUserAffinity** and **UserInfluence**. The algorithm is:

```
PeopleRanker(InputSet, UserUserAffinity, UserInfluence)
{
  For every user U in InputSet
    RankedPeople[U].rank = (A * UserInfluence[U]) *
                           (B * (1 + UserUserAffinity[CurrentUser, U]))
    If InputSet is ranked
      RankedPeople[U].rank *= C * InputSet[U].rank
}
```

[00122] For each of the users being ranked, the system computes the weighted product of their influence on other users, the affinity of the current user to the other users, and the existing rank of that user (if it was provided). The weights reflect the relative importance of influence, affinity, and the effectiveness of the original ranking. The output is a ranking for each user in the input set.

QUERY COMPLETION RANKER 270

[00123] This algorithm is invoked on a partial query string, and computes the set of completions for it (suggested full queries the user might have in mind) and also gets **QueryCount**, **UserQueryAffinity**, and the **InvertedIndex**. This algorithm returns up to COMPLETION_MAX ranked query completions. COMPLETION_MAX may be defined by a user or an administrator of the system. The algorithm is:

```
QueryCompletionRanker(QueryPrefix, QueryCount, InvertedIndex)
{
  RankedQueries = set of queries in QueryCount that begin with QueryPrefix
                  rank for query Q = (A * QueryPrefix[Q].count) +
                                     (B * (1 + UserQueryAffinity[Q,
CurrentUser]))
  if (number of queries in RankedQueries < COMPLETION_MAX) {
    QueryLexemes = set of lexemes in InvertedIndex that begin with QueryPrefix
    Sort QueryLexemes by the number of times the lexeme appears in the index
    Copy from QueryLexemes into RankedQueries until you reach
      COMPLETION_MAX or have copied them all. Assign each the rank
      A * (count of appearances of lexeme in index)
  }
}
```

[00124] The system computes query completions from the set of queries that have already been executed and from textual analysis of the inverted index. In some cases, the system biases towards the former, but fills out the potential query list from the latter as needed to reach the desired number of completions. The rank for previously executed queries is a weighted sum of the number of times the query has been executed and the affinity of the current user to each query. The rank for matching lexemes is the count of that lexeme's appearances, weighted accordingly. The output is a ranked set of query completions.

[00125] Figure 16 is a block diagram illustrating some of the components that may be incorporated in at least some of the computer systems and other devices on which the system operates and interacts with in some examples. In various examples, these computer systems and other devices 1600 can include server computer systems, desktop computer systems, laptop computer systems, netbooks, tablets, mobile phones, personal digital assistants, televisions, cameras, automobile computers, electronic media players, and/or the like. In various examples, the computer systems and devices include one or more of each of the following: a central processing unit ("CPU") 1601 configured to execute computer programs; a computer memory 1602 configured to store programs and data while they are being used, including a multithreaded program being tested, a debugger, the facility, an operating system including a kernel, and device drivers; a persistent storage device 1603, such as a hard drive or flash drive configured to persistently store programs and data; a computer-readable storage media drive 1604, such as a floppy, flash, CD-ROM, or DVD drive, configured to read programs and data stored on a computer-readable storage medium, such as a floppy disk, flash memory device, a CD-ROM, a DVD; and a network connection 1605 configured to connect the computer system to other computer systems to send and/or receive data, such as via the Internet, a local area network, a wide area network, a point-to-point dial-up connection, a cell phone network, or another network and its networking hardware in various examples including routers, switches, and various types of transmitters, receivers, or computer-readable transmission media. While computer systems configured as described above may be used to support the operation of the facility, those skilled in the art will readily appreciate that the facility may be implemented using devices of various types and configurations, and having various components.

Elements of the facility may be described in the general context of computer-executable instructions, such as program modules, executed by one or more computers or other devices. Generally, program modules include routines, programs, objects, components, data structures, and/or the like configured to perform particular tasks or implement particular abstract data types and may be encrypted. Moreover, the functionality of the program modules may be combined or distributed as desired in various examples. Moreover, display pages may be implemented in any of various ways, such as in C++ or as web pages in XML (Extensible Markup Language), HTML (HyperText Markup Language), JavaScript, AJAX (Asynchronous JavaScript and XML) techniques or any other scripts or methods of creating displayable data, such as the Wireless Access Protocol ("WAP").

[00126] The following discussion provides a brief, general description of a suitable computing environment in which the invention can be implemented. Although not required, aspects of the invention are described in the general context of computer-executable instructions, such as routines executed by a general-purpose data processing device, e.g., a server computer, wireless device or personal computer. Those skilled in the relevant art will appreciate that aspects of the invention can be practiced with other communications, data processing, or computer system configurations, including: Internet appliances, hand-held devices (including personal digital assistants (PDAs)), wearable computers, all manner of cellular or mobile phones (including Voice over IP (VoIP) phones), dumb terminals, media players, gaming devices, multi-processor systems, microprocessor-based or programmable consumer electronics, set-top boxes, network PCs, mini-computers, mainframe computers, and the like. Indeed, the terms "computer," "server," "host," "host system," and the like are generally used interchangeably herein, and refer to any of the above devices and systems, as well as any data processor.

[00127] Aspects of the invention can be embodied in a special purpose computer or data processor that is specifically programmed, configured, or constructed to perform one or more of the computer-executable instructions explained in detail herein. While aspects of the invention, such as certain functions, are described as being performed exclusively on a single device, the invention can also be practiced in distributed environments where functions or modules are shared among disparate

processing devices, which are linked through a communications network, such as a Local Area Network (LAN), Wide Area Network (WAN), or the Internet. In a distributed computing environment, program modules may be located in both local and remote memory storage devices.

[00128] Aspects of the invention may be stored or distributed on tangible computer-readable media, including magnetically or optically readable computer discs, hard-wired or preprogrammed chips (e.g., EEPROM semiconductor chips), nanotechnology memory, biological memory, or other data storage media. Alternatively, computer implemented instructions, data structures, screen displays, and other data under aspects of the invention may be distributed over the Internet or over other networks (including wireless networks), on a propagated signal on a propagation medium (e.g., an electromagnetic wave(s), a sound wave, etc.) over a period of time, or they may be provided on any analog or digital network (packet switched, circuit switched, or other scheme).

[00129] Unless the context clearly requires otherwise, throughout the description and the claims, the words "comprise," "comprising," and the like are to be construed in an inclusive sense, as opposed to an exclusive or exhaustive sense; that is to say, in the sense of "including, but not limited to." As used herein, the terms "connected," "coupled," or any variant thereof means any connection or coupling, either direct or indirect, between two or more elements; the coupling or connection between the elements can be physical, logical, or a combination thereof. Additionally, the words "herein," "above," "below," and words of similar import, when used in this application, refer to this application as a whole and not to any particular portions of this application. Where the context permits, words in the above Detailed Description using the singular or plural number may also include the plural or singular number respectively. The word "or," in reference to a list of two or more items, covers all of the following interpretations of the word: any of the items in the list, all of the items in the list, and any combination of the items in the list.

[00130] The above Detailed Description of examples of the invention is not intended to be exhaustive or to limit the invention to the precise form disclosed above. While specific examples for the invention are described above for illustrative purposes, various equivalent modifications are possible within the scope of the invention, as those skilled in the relevant art will recognize. For example, while

processes or blocks are presented in a given order, alternative implementations may perform routines having steps, or employ systems having blocks, in a different order, and some processes or blocks may be deleted, moved, added, subdivided, combined, and/or modified to provide alternative or subcombinations. Each of these processes or blocks may be implemented in a variety of different ways. Also, while processes or blocks are at times shown as being performed in series, these processes or blocks may instead be performed or implemented in parallel, or may be performed at different times. Further any specific numbers noted herein are only examples: alternative implementations may employ differing values or ranges.

[00131] The teachings of the invention provided herein can be applied to other systems, not necessarily the system described above. The elements and acts of the various examples described above can be combined to provide further implementations of the invention. Some alternative implementations of the invention may include not only additional elements to those implementations noted above, but also may include fewer elements.

[00132] Any patents and applications and other references noted above, including any that may be listed in accompanying filing papers, are incorporated herein by reference. Aspects of the invention can be modified, if necessary, to employ the systems, functions, and concepts of the various references described above to provide yet further implementations of the invention.

[00133] These and other changes can be made to the invention in light of the above Detailed Description. While the above description describes certain examples of the invention, and describes the best mode contemplated, no matter how detailed the above appears in text, the invention can be practiced in many ways. Details of the system may vary considerably in its specific implementation, while still being encompassed by the invention disclosed herein. As noted above, particular terminology used when describing certain features or aspects of the invention should not be taken to imply that the terminology is being redefined herein to be restricted to any specific characteristics, features, or aspects of the invention with which that terminology is associated. In general, the terms used in the following claims should not be construed to limit the invention to the specific examples disclosed in the specification, unless the above Detailed Description section explicitly defines such terms. Accordingly, the actual scope of the invention encompasses not only the

disclosed examples, but also all equivalent ways of practicing or implementing the invention under the claims. In some cases, various steps in the algorithms discussed herein may be added, altered, or removed without departing from the disclosed subject matter. Those skilled in the art will appreciate that features described above may be altered in a variety of ways. For example, the order of the logic may be rearranged, sublogic may be performed in parallel, illustrated logic may be omitted, other logic may be included, etc.

[00134] To reduce the number of claims, certain aspects of the invention are presented below in certain claim forms, but the applicant contemplates the various aspects of the invention in any number of claim forms. For example, while only one aspect of the invention is recited as a means-plus-function claim under 35 U.S.C. § 112(f), other aspects may likewise be embodied as a means-plus-function claim, or in other forms, such as being embodied in a computer-readable medium. (Any claims intended to be treated under 35 U.S.C. § 112(f) will begin with the words "means for", but use of the term "for" in any other context is not intended to invoke treatment under 35 U.S.C. § 112(f).) Accordingly, the applicant reserves the right to pursue additional claims after filing this application to pursue such additional claim forms, in either this application or in a continuing application.

CLAIMS

We claim:

1. A computer-implemented method of identifying information of interest within an organization, wherein the organization includes a structured body of users with associated roles within the organization and who have access to information items within the organization, the method comprising:

determining use data that characterizes relationships among the information items with respect to users within the organization,

wherein the information items include user data and collections of information items;

generating interest data indicating affinity among the information items based on the determined use data;

receiving a request for a feed or set of information items from a particular user; and

responding to the request by providing one or more items based on the generated interest data.

2. The computer-implemented method of claim 1,

wherein the information items within the organization include profiles of the users, a document, or a portion of a document, and

wherein the organization is a business enterprise or a legal entity.

3. The computer-implemented method of claim 1,

wherein a relationship between a first user and a first information item corresponds to an activity performed by the first user on the first information item,

wherein the activity is querying, browsing, opening, viewing, editing, critiquing, bookmarking, liking, sharing, downloading, collecting, or curating the information item, and

wherein determining the use data includes tracking the activity.

4. The computer-implemented method of claim 3, wherein tracking the activity is performed using web browser software without downloading additional software or documents.

5. The computer-implemented method of claim 1, wherein a relationship between two users corresponds to:

an organizational relationship between the two users with respect to the roles of the two users with the organization,

an activity performed by the two users together within the organization, or

a pair of relationships respectively between the two users and a first information item.

6. The computer-implemented method of claim 1, wherein responding to the request comprises:

combining multiple, similar results into one result;

ordering multiple results by –

a recency of an information item,

a popularity of an information item,

a popularity of a collection of information items,

an authority of the particular user, or

the interest data; and

presenting the multiple results based on the ordering.

7. The computer-implemented method of claim 6, wherein the request is generated automatically, and responding to the request comprises:

generating a message to be sent by email or other messaging system.

8. The computer-implemented method of claim 7, further comprising:

generating automated requests for feeds based on a user-specified frequency.

9 The computer-implemented method of claim 7, further comprising:

generating automated requests for feeds based on:

the number of items provided based on the generated interest data,
the popularity of the relevant information items, or
the popularity of the users who own or performed activities on the relevant information items.

10. The computer-implemented method of claim 6, further comprising:
determining the recency of a first information item based on the number of activities performed on the first information item in the last hour, day, week, or other time period.

11. The computer-implemented method of claim 6, further comprising:
determining the popularity of a first information item based on
a number of activities performed on the first information item,
an authority of users who performed activities on the first information item, or
a popularity of collections of information items to which the first information item belongs.

12. The computer-implemented method of claim 6, further comprising:
determining the authority of the particular user based on a number of activities performed on information items created by the particular user.

13. The computer-implemented method of claim 6, further comprising:
determining the popularity of a first collection of information items based on a number of activities performed on the information items in the first collection of information items.

14. The computer-implemented method of claim 1, further comprising:
determining an affinity between a first user and a second user based on:
a relationship between the first user and the second user, and
an interest indicated by the first user with respect to the second user.

15. The computer-implemented method of claim 1, further comprising:

determining an affinity between a first user and a first information item based on:

an affinity between the first user and a second user and a relationship between the second user and the first information item, or

an affinity between the first user and a collection to which the first information item belongs.

16. The computer-implemented method of claim 1, further comprising:

determining an affinity between a user and a collection of information items based on a relationship between the user and the collection.

17. The computer-implemented method of claim 1, further comprising determining additional use data characterizing relationships among users and information items across the organization and at least one other, independent organization.

18. A system to identify information of interest within an organization, wherein the organization includes a group of users on a private network and sharing an internet domain, the system comprising:

means for gathering use data that characterizes relationships among information items within the organization,

wherein the information items include user data and collections of information items;

means for computing interest data indicating affinity among the information items based on the determined use data to generate interest graph data structures,

wherein each interest graph data structure expresses the affinity between at least one user and one information item, and

wherein the affinity represents a likelihood that the one information item is of interest to the at least one user;

means for receiving a request for data regarding the information items; and

means for responding to the request by providing one or more results based on the computed interest data.

19. The system of claim 18 wherein the user data includes user profile data specifying access control privileges of at least one user and wherein the means for responding provides only results for which the user has access control privileges.

20. A computer-readable storage medium storing instructions that, if executed by a computing system having a processor, cause the computing system to perform a method comprising:

determining use data that characterizes relationships among information items with respect to users within an organization,

wherein the information items include user data and collections of information items;

generating interest data indicating affinity among the information items based on the determined use data;

receiving a request for a feed or set of information items from a particular user; and

responding to the request by providing one or more items based on the generated interest data.

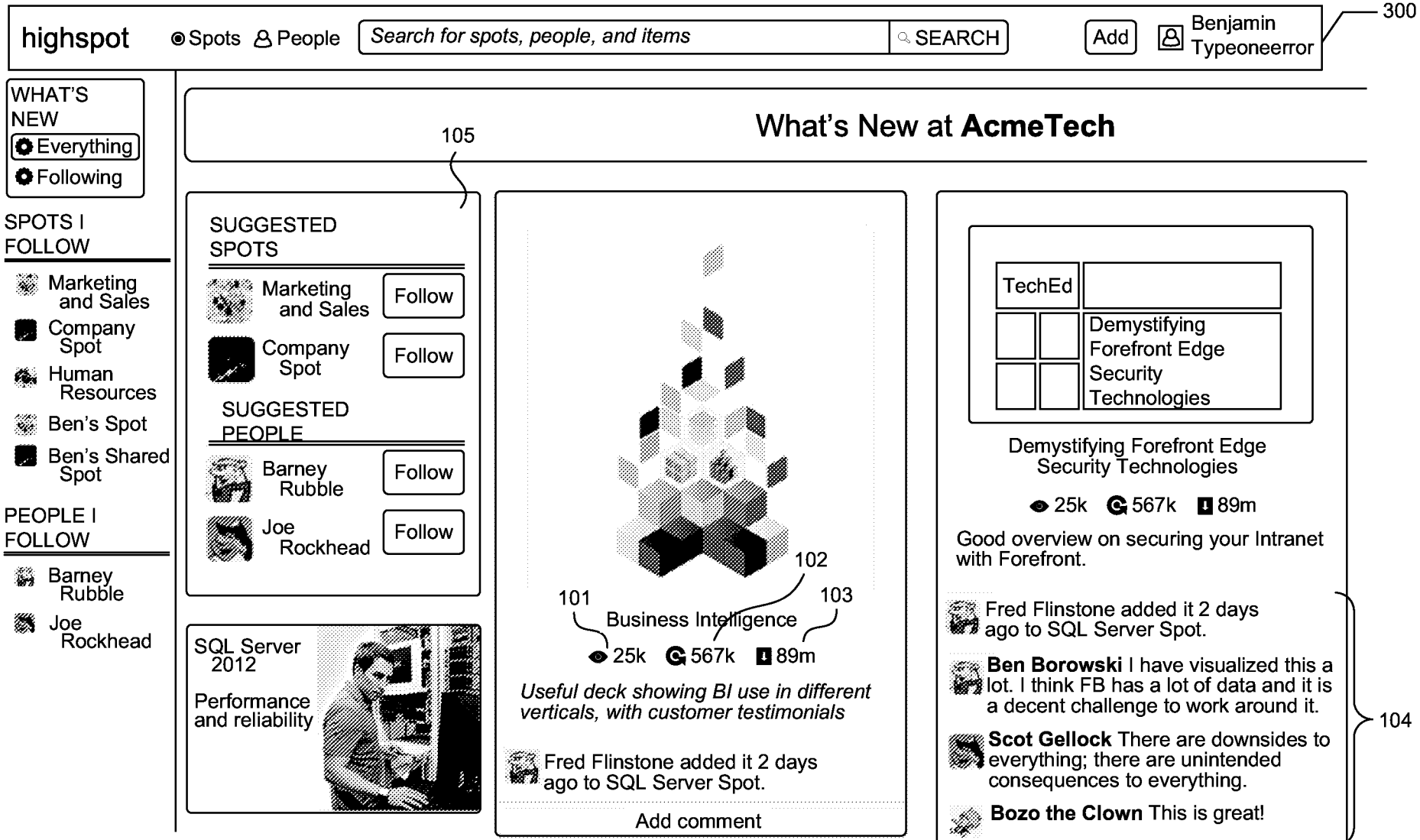


FIG. 1

SUBSTITUTE SHEET (RULE 26)

2/17

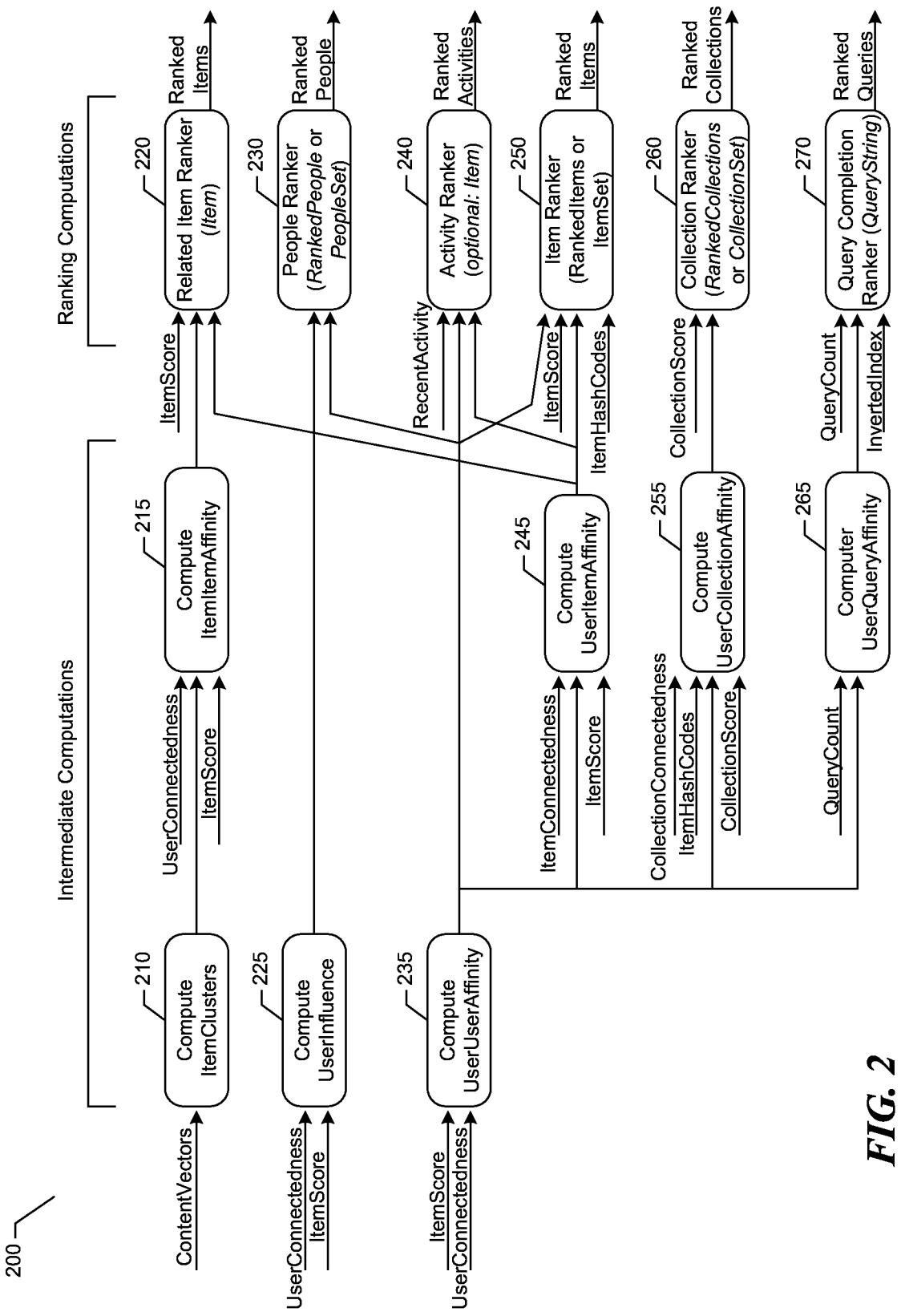

200
Intermediate Computations
Ranking Computations
ContentVectors
UserConnectedness
ItemScore
ItemScore
UserConnectedness
210
Compute ItemClusters
215
Compute ItemItemAffinity
220
Related Item Ranker (Item)
Ranked Items
225
Compute UserInfluence
230
People Ranker (RankedPeople or PeopleSet)
Ranked People
235
Compute UserUserAffinity
240
Activity Ranker (optional: Item)
Ranked Activities
245
Compute UserItemAffinity
250
Item Ranker (RankedItems or ItemSet)
Ranked Items
255
Compute UserCollectionAffinity
260
Collection Ranker (RankedCollections or CollectionSet)
Ranked Collections
265
Computer UserQueryAffinity
270
Query Completion Ranker (QueryString)
Ranked Queries
ItemScore
UserConnectedness
RecentActivity
ItemScore
ItemHashCodes
ItemConnectedness
ItemScore
CollectionScore
CollectionConnectedness
ItemHashCodes
CollectionScore
QueryCount
QueryCount
InvertedIndex

FIG. 2

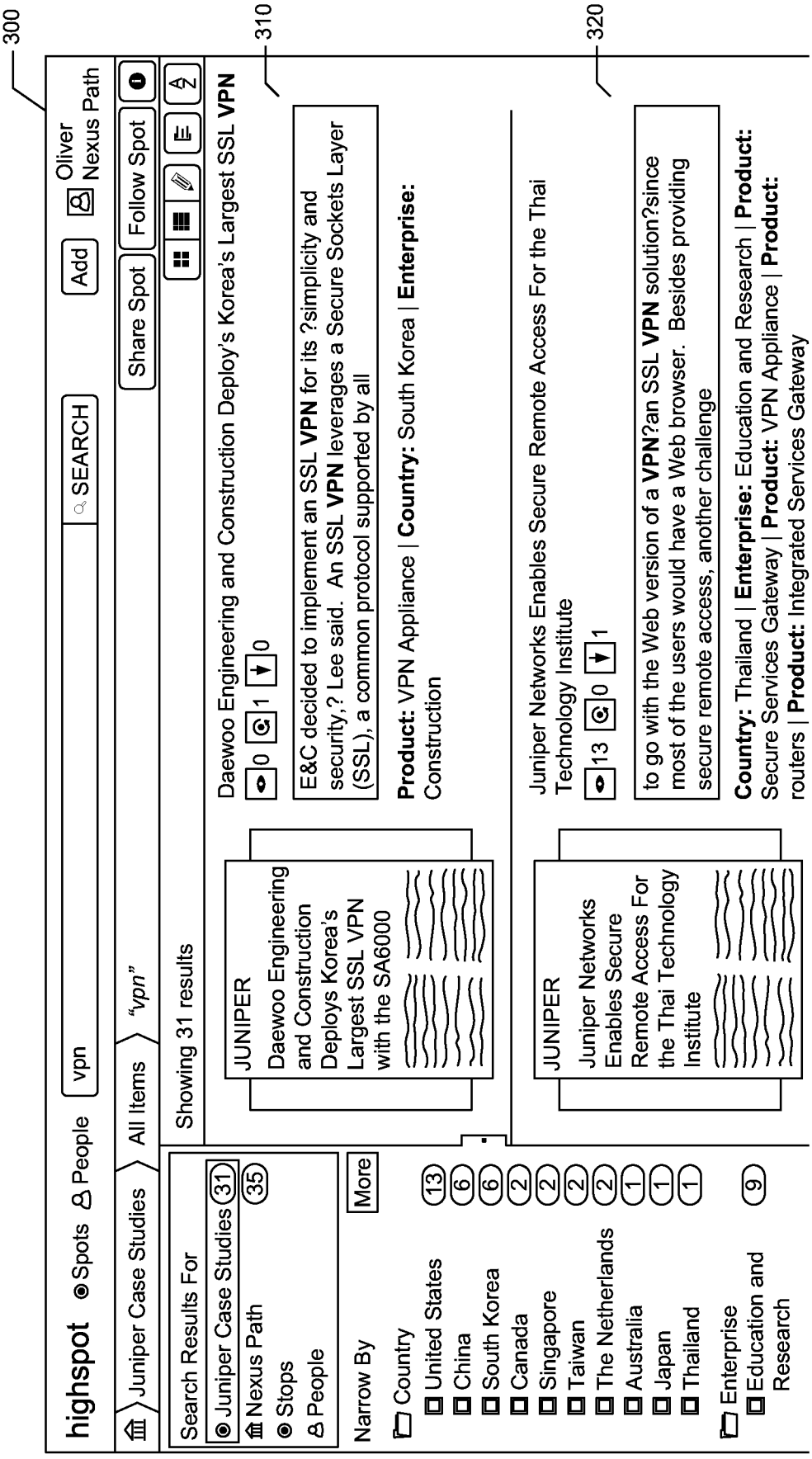

300
highspot
Spots
People
vpn
SEARCH
Add
Oliver
Nexus Path
Juniper Case Studies
All Items
"vpn"
Share Spot
Follow Spot
Search Results For
Juniper Case Studies 31
Nexus Path 35
Stops
People
Narrow By
More
Country
United States 13
China 6
South Korea 6
Canada 2
Singapore 2
Taiwan 2
The Netherlands 2
Australia 1
Japan 1
Thailand 1
Enterprise
Education and Research 9
Showing 31 results
310
JUNIPER
Daewoo Engineering and Construction Deploys Korea's Largest SSL VPN with the SA6000
Daewoo Engineering and Construction Deploy's Korea's Largest SSL VPN
0 1 0
E&C decided to implement an SSL VPN for its ?simplicity and security,? Lee said. An SSL VPN leverages a Secure Sockets Layer (SSL), a common protocol supported by all
Product: VPN Appliance | Country: South Korea | Enterprise: Construction
320
JUNIPER
Juniper Networks Enables Secure Remote Access For the Thai Technology Institute
Juniper Networks Enables Secure Remote Access For the Thai Technology Institute
13 0 1
to go with the Web version of a VPN?an SSL VPN solution?since most of the users would have a Web browser. Besides providing secure remote access, another challenge
Country: Thailand | Enterprise: Education and Research | Product: Secure Services Gateway | Product: VPN Appliance | Product: routers | Product: Integrated Services Gateway

FIG. 3

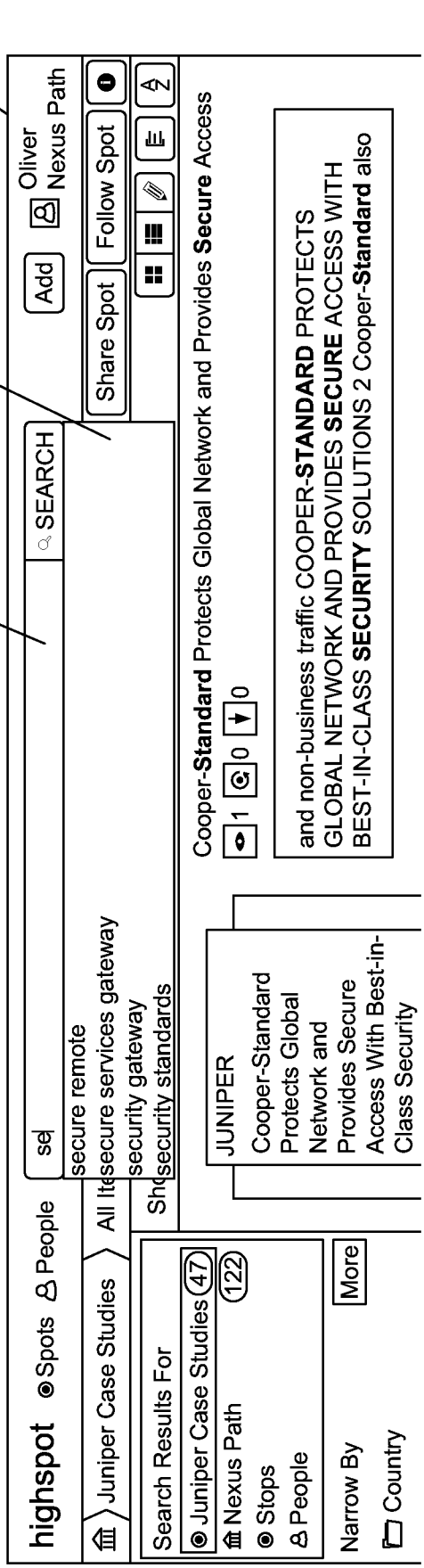

400
410
420
highspot
Spots
People
sel
SEARCH
Add
Oliver
Nexus Path
Share Spot
Follow Spot
Juniper Case Studies
All Its
secure remote
secure services gateway
security gateway
Sh
security standards
Search Results For
Juniper Case Studies 47
Nexus Path 122
Stops
People
Narrow By
More
Country
JUNIPER
Cooper-Standard
Protects Global
Network and
Provides Secure
Access With Best-in-
Class Security
Cooper-Standard Protects Global Network and Provides Secure Access
1
0
0
and non-business traffic COOPER-STANDARD PROTECTS
GLOBAL NETWORK AND PROVIDES SECURE ACCESS WITH
BEST-IN-CLASS SECURITY SOLUTIONS 2 Cooper-Standard also

FIG. 4A

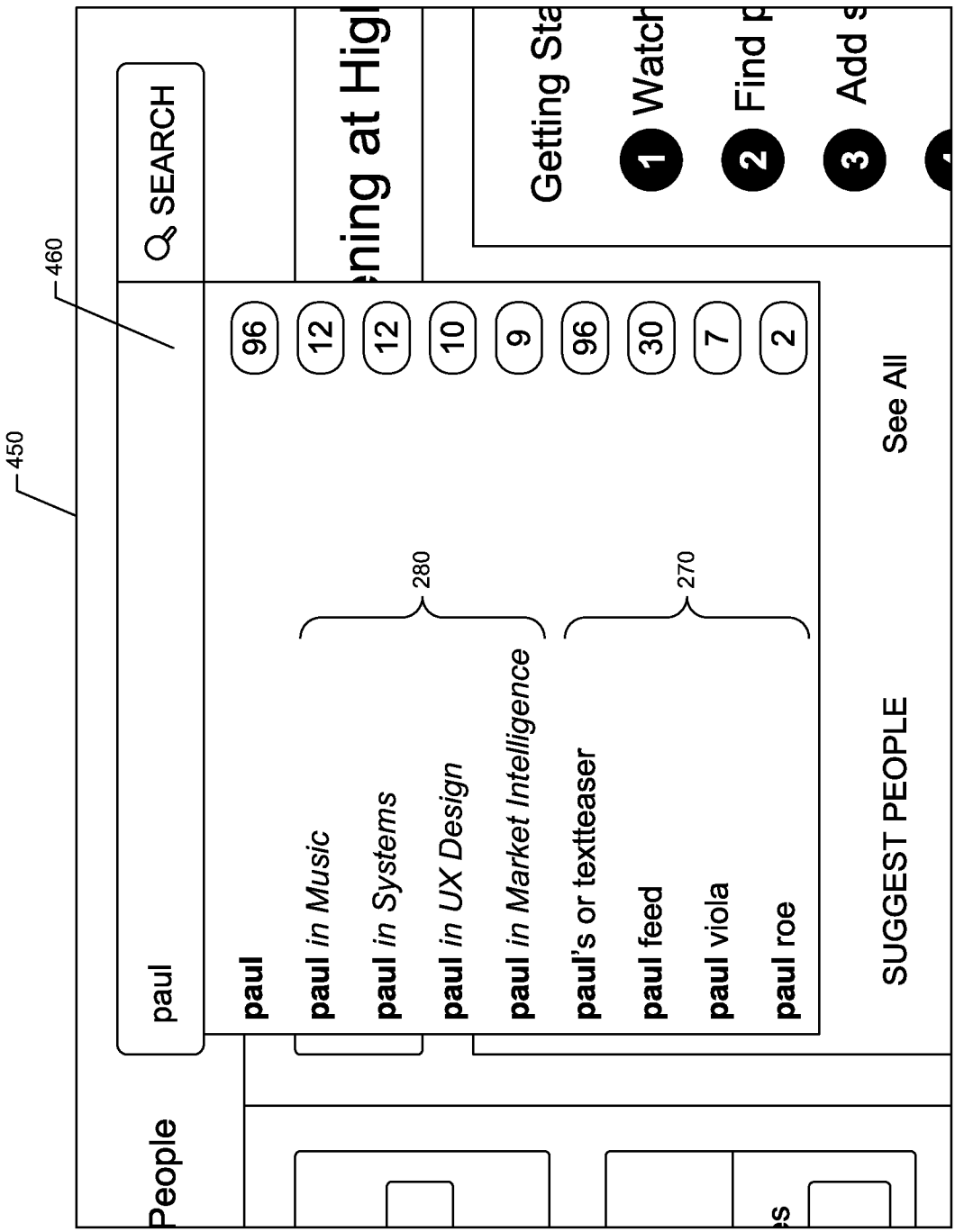

450
460
People
paul
SEARCH
paul
paul in Music
paul in Systems
paul in UX Design
paul in Market Intelligence
paul's or textteaser
paul feed
paul viola
paul roe
96
12
12
10
9
96
30
7
2
280
270
SUGGEST PEOPLE
See All
Getting Sta
1 Watch
2 Find p
3 Add s

FIG. 4B

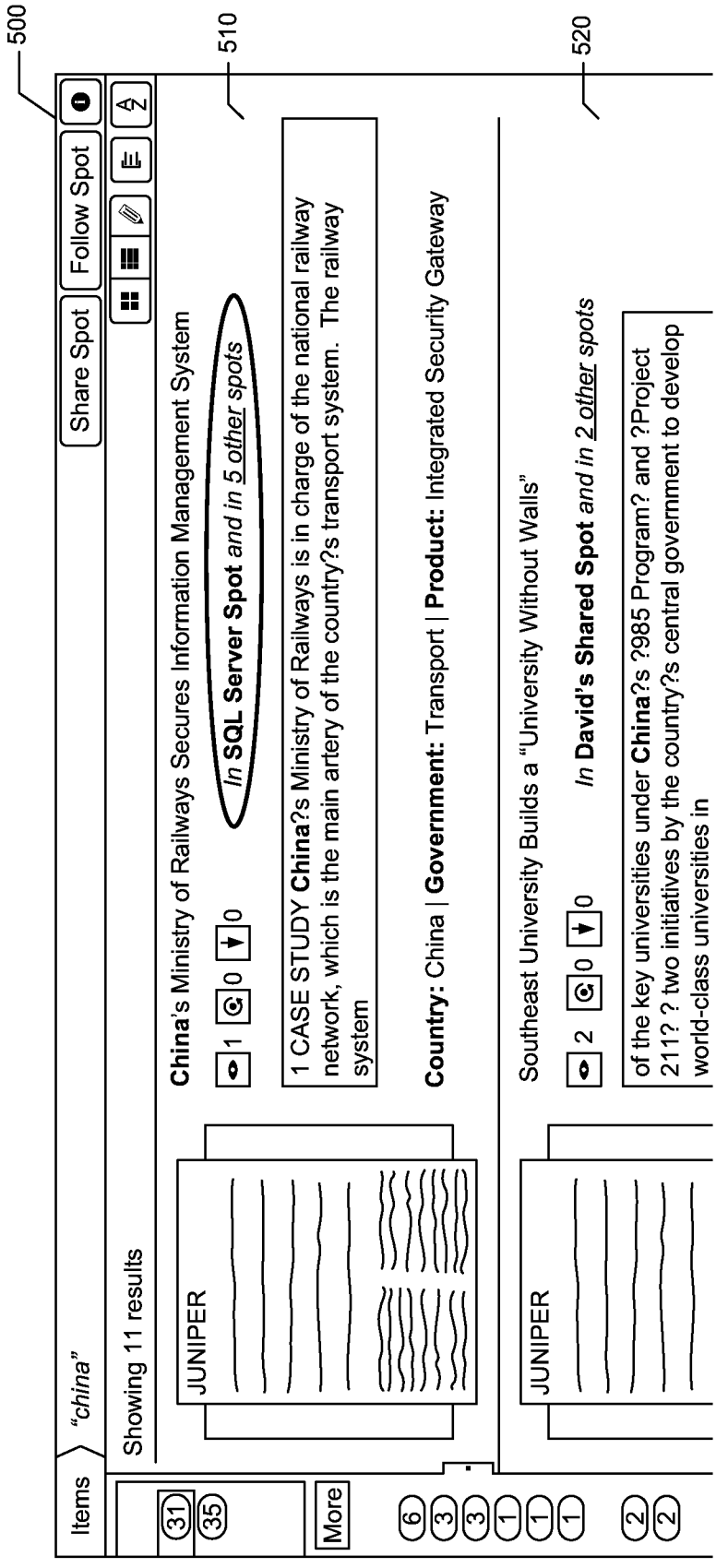

500
510
520
Items
"china"
Share Spot
Follow Spot
Showing 11 results
JUNIPER
China's Ministry of Railways Secures Information Management System
1 0 0
In SQL Server Spot and in 5 other spots
1 CASE STUDY China?s Ministry of Railways is in charge of the national railway network, which is the main artery of the country?s transport system. The railway system
Country: China | Government: Transport | Product: Integrated Security Gateway
JUNIPER
Southeast University Builds a "University Without Walls"
2 0 0
In David's Shared Spot and in 2 other spots
of the key universities under China?s ?985 Program? and ?Project 211? ? two initiatives by the country?s central government to develop world-class universities in
More
31
35
6
3
3
1
1
1
2
2

FIG. 5

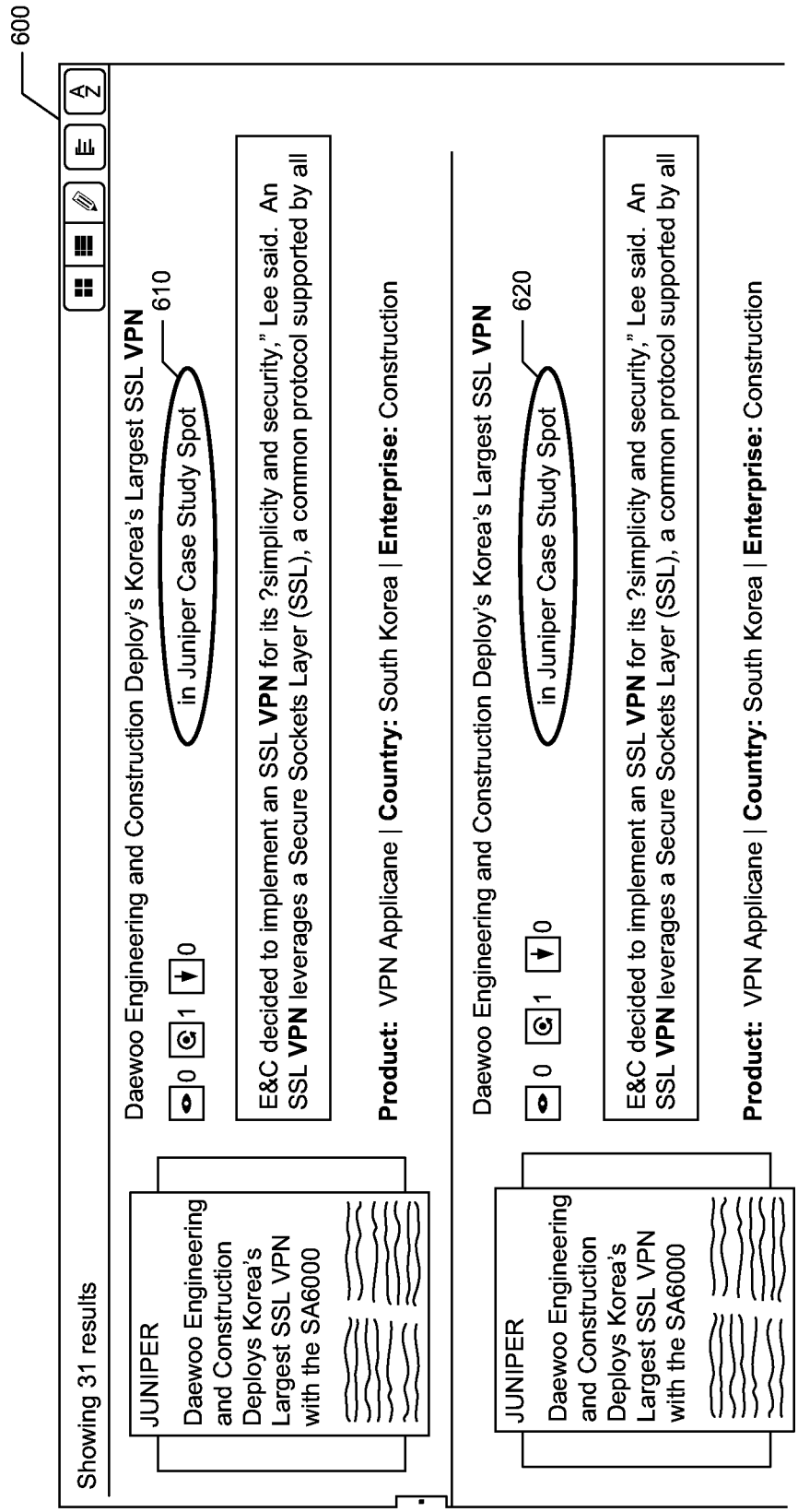

600
Showing 31 results
JUNIPER
Daewoo Engineering and Construction Deploys Korea's Largest SSL VPN with the SA6000
Daewoo Engineering and Construction Deploy's Korea's Largest SSL VPN
0 1 0
610
in Juniper Case Study Spot
E&C decided to implement an SSL VPN for its ?simplicity and security," Lee said. An SSL VPN leverages a Secure Sockets Layer (SSL), a common protocol supported by all
Product: VPN Applicane | Country: South Korea | Enterprise: Construction
JUNIPER
Daewoo Engineering and Construction Deploys Korea's Largest SSL VPN with the SA6000
Daewoo Engineering and Construction Deploy's Korea's Largest SSL VPN
0 1 0
620
in Juniper Case Study Spot
E&C decided to implement an SSL VPN for its ?simplicity and security," Lee said. An SSL VPN leverages a Secure Sockets Layer (SSL), a common protocol supported by all
Product: VPN Applicane | Country: South Korea | Enterprise: Construction

FIG. 6

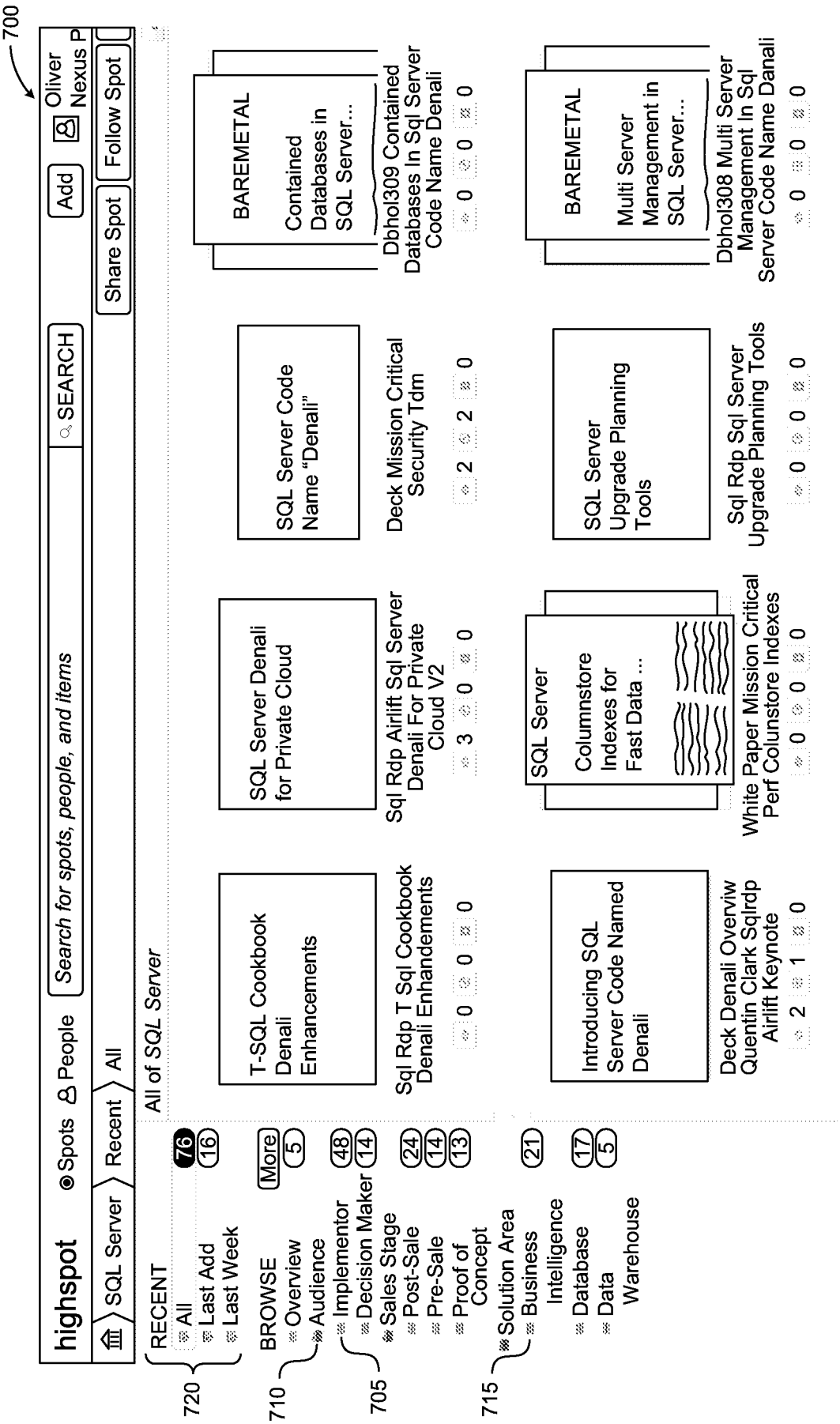

700
highspot
Spots
People
Search for spots, people, and items
SEARCH
Add
Oliver Nexus P
SQL Server
Recent
All
Share Spot
Follow Spot
All of SQL Server
RECENT
All 76
Last Add 16
Last Week
BROWSE
Overview More 5
Audience
Implementor 48
Decision Maker 14
Sales Stage
Post-Sale 24
Pre-Sale 14
Proof of Concept 13
Solution Area
Business Intelligence 21
Database 17
Data Warehouse 5
720
710
705
715
T-SQL Cookbook Denali Enhancements
SQL Server Denali for Private Cloud
SQL Server Code Name "Denali"
BAREMETAL
Contained Databases in SQL Server...
Dbhol309 Contained Databases In Sql Server Code Name Denali
Sql Rdp Airlift Sql Server Denali For Private Cloud V2
Deck Mission Critical Security Tdm
Sql Rdp T Sql Cookbook Denali Enhandements
SQL Server Columnstore Indexes for Fast Data ...
SQL Server Upgrade Planning Tools
BAREMETAL
Multi Server Management in SQL Server...
Dbhol308 Multi Server Management In Sql Server Code Name Danali
Introducing SQL Server Code Named Denali
White Paper Mission Critical Perf Colunstore Indexes
Sql Rdp Sql Server Upgrade Planning Tools
Deck Denali Overviw Quentin Clark Sqlrdp Airlift Keynote

FIG. 7

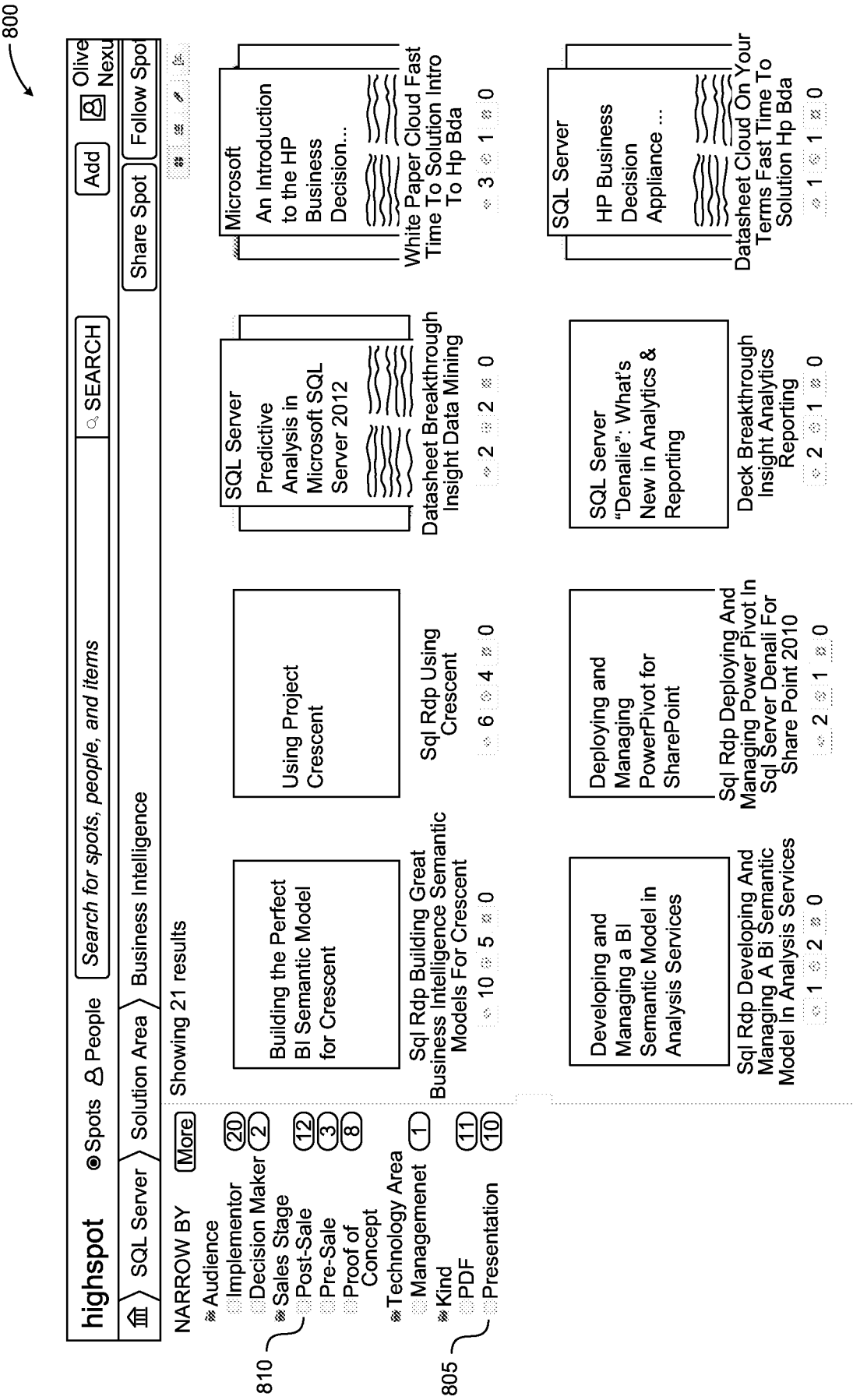

800
highspot
Spots
People
Search for spots, people, and items
SEARCH
Add
Olive
Nexu
SQL Server
Solution Area
Business Intelligence
Share Spot
Follow Spot
NARROW BY
More
Showing 21 results
Audience
Implementor 20
Decision Maker 2
Sales Stage
Post-Sale 12
Pre-Sale 3
Proof of Concept 8
Technology Area
Managemenet 1
Kind
PDF 11
Presentation 10
810
805
Building the Perfect BI Semantic Model for Crescent
Sql Rdp Building Great Business Intelligence Semantic Models For Crescent
10 5 0
Using Project Crescent
Sql Rdp Using Crescent
6 4 0
Developing and Managing a BI Semantic Model in Analysis Services
Sql Rdp Developing And Managing A Bi Semantic Model In Analysis Services
1 2 0
Deploying and Managing PowerPivot for SharePoint
Sql Rdp Deploying And Managing Power Pivot In Sql Server Denali For Share Point 2010
2 1 0
SQL Server
Predictive Analysis in Microsoft SQL Server 2012
Datasheet Breakthrough Insight Data Mining
2 2 0
SQL Server
“Denalie”: What’s New in Analytics & Reporting
Deck Breakthrough Insight Analytics Reporting
2 1 0
Microsoft
An Introduction to the HP Business Decision...
White Paper Cloud Fast Time To Solution Intro To Hp Bda
3 1 0
SQL Server
HP Business Decision Appliance ...
Datasheet Cloud On Your Terms Fast Time To Solution Hp Bda
1 1 0

FIG. 8

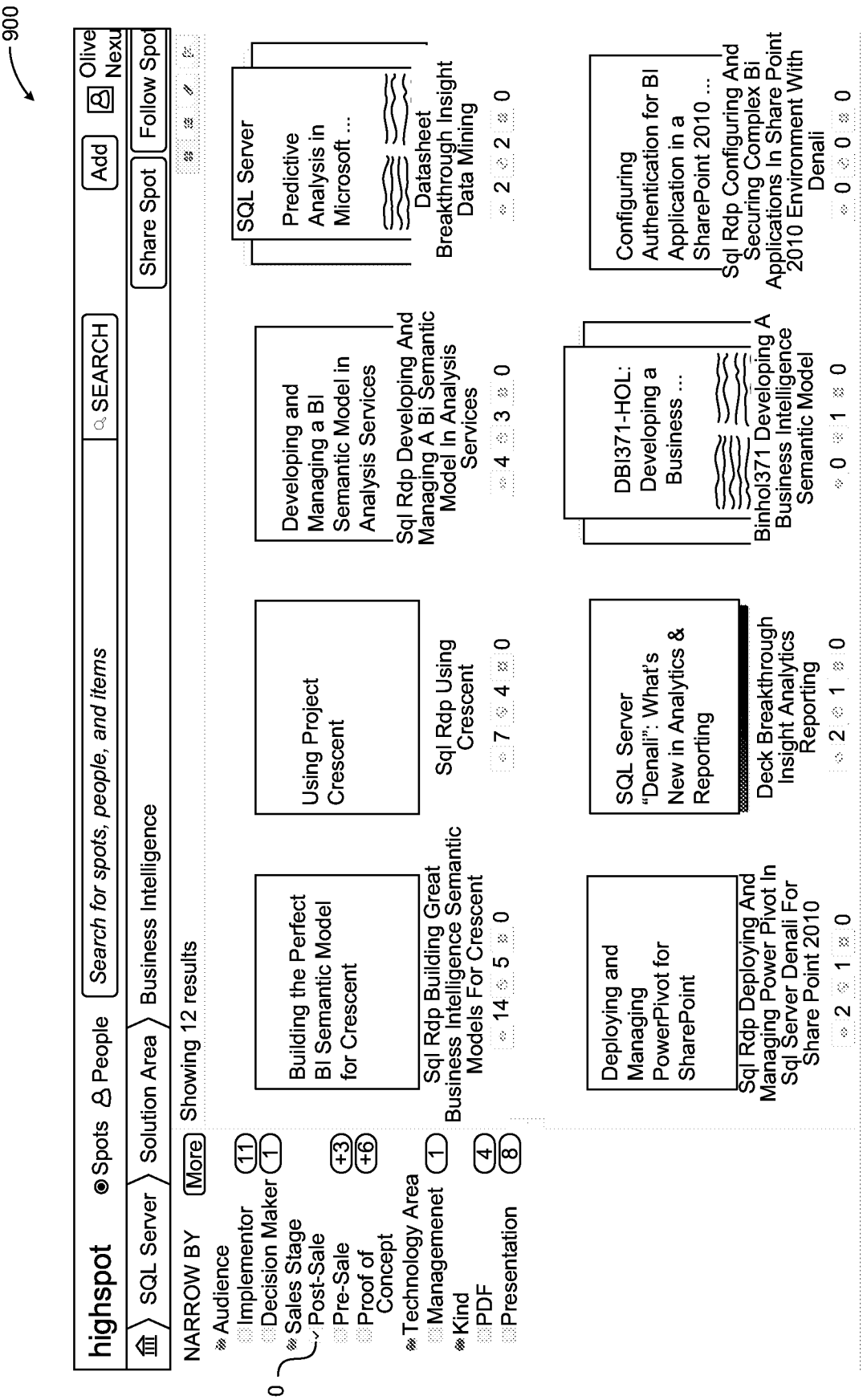

900
highspot
Spots
People
Search for spots, people, and items
SEARCH
Add
Olive
Nexu
SQL Server
Solution Area
Business Intelligence
Share Spot
Follow Spot
NARROW BY
More
Showing 12 results
Audience
Implementor 11
Decision Maker 1
Sales Stage
Post-Sale
Pre-Sale +3
Proof of Concept +6
Technology Area
Managemenet 1
Kind
PDF 4
Presentation 8
910
Sql Rdp Deploying And Managing Power Pivot In Sql Server Denali For Share Point 2010
2 1 0
Building the Perfect BI Semantic Model for Crescent
Sql Rdp Building Great Business Intelligence Semantic Models For Crescent
14 5 0
Using Project Crescent
Sql Rdp Using Crescent
7 4 0
Deploying and Managing PowerPivot for SharePoint
SQL Server "Denali": What's New in Analytics & Reporting
Deck Breakthrough Insight Analytics Reporting
2 1 0
Developing and Managing a BI Semantic Model in Analysis Services
Sql Rdp Developing And Managing A Bi Semantic Model In Analysis Services
4 3 0
SQL Server
Predictive Analysis in Microsoft ...
Datasheet Breakthrough Insight Data Mining
2 2 0
DBI371-HOL: Developing a Business ...
Binhol371 Developing A Business Intelligence Semantic Model
0 1 0
Configuring Authentication for BI Application in a SharePoint 2010 ...
Sql Rdp Configuring And Securing Complex Bi Applications In Share Point 2010 Environment With Denali
0 0 0

FIG. 9

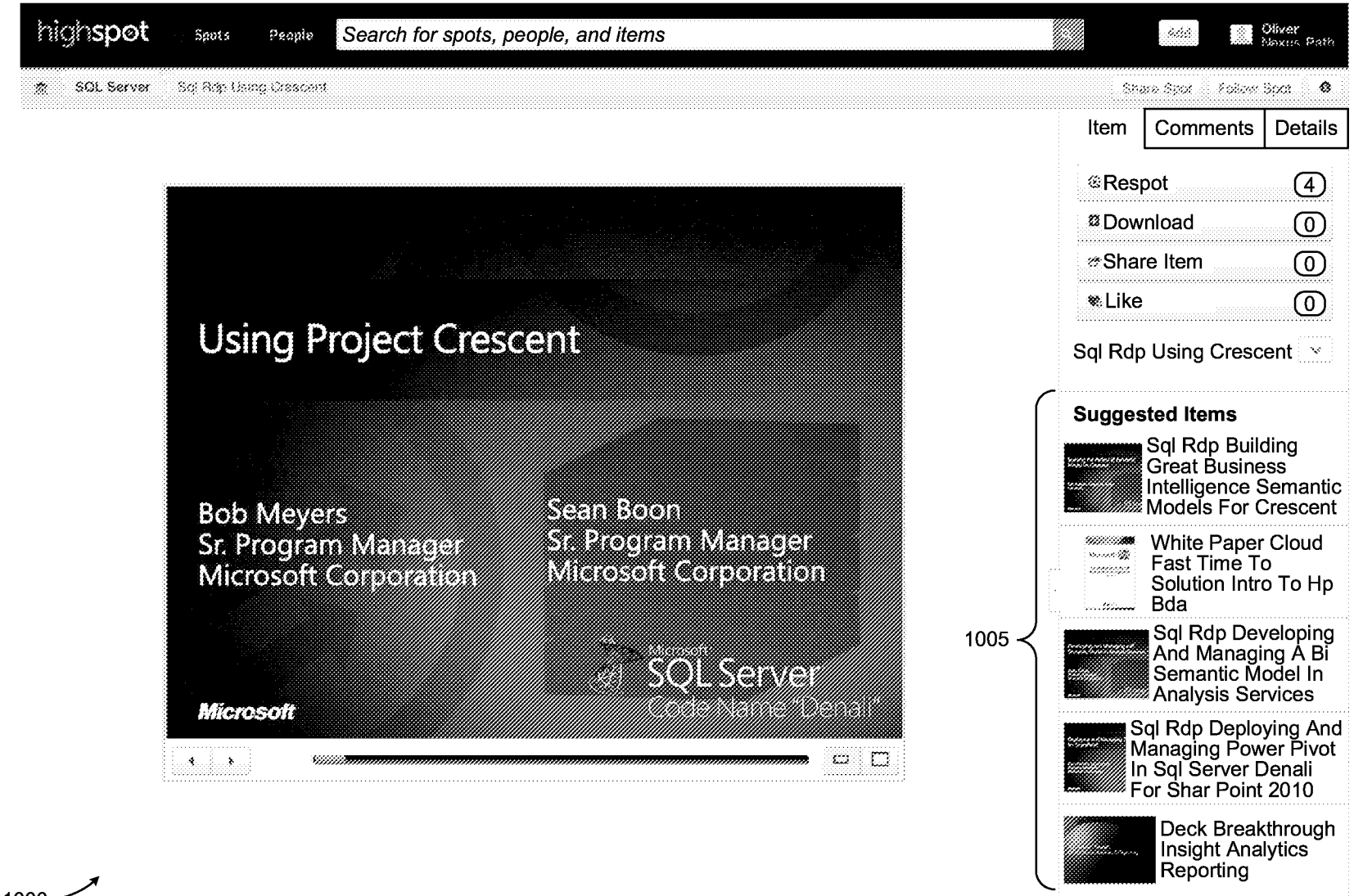

highspot
Spots
People
Search for spots, people, and items
Add
Oliver
SQL Server
Item
Comments
Details
Respot 4
Download 0
Share Item 0
Like 0
Sql Rdp Using Crescent
Using Project Crescent
Bob Meyers
Sr. Program Manager
Microsoft Corporation
Sean Boon
Sr. Program Manager
Microsoft Corporation
Microsoft SQL Server
Code Name "Denali"
Microsoft
Suggested Items
Sql Rdp Building Great Business Intelligence Semantic Models For Crescent
White Paper Cloud Fast Time To Solution Intro To Hp Bda
Sql Rdp Developing And Managing A Bi Semantic Model In Analysis Services
Sql Rdp Deploying And Managing Power Pivot In Sql Server Denali For Shar Point 2010
Deck Breakthrough Insight Analytics Reporting
1005
1000

FIG. 10

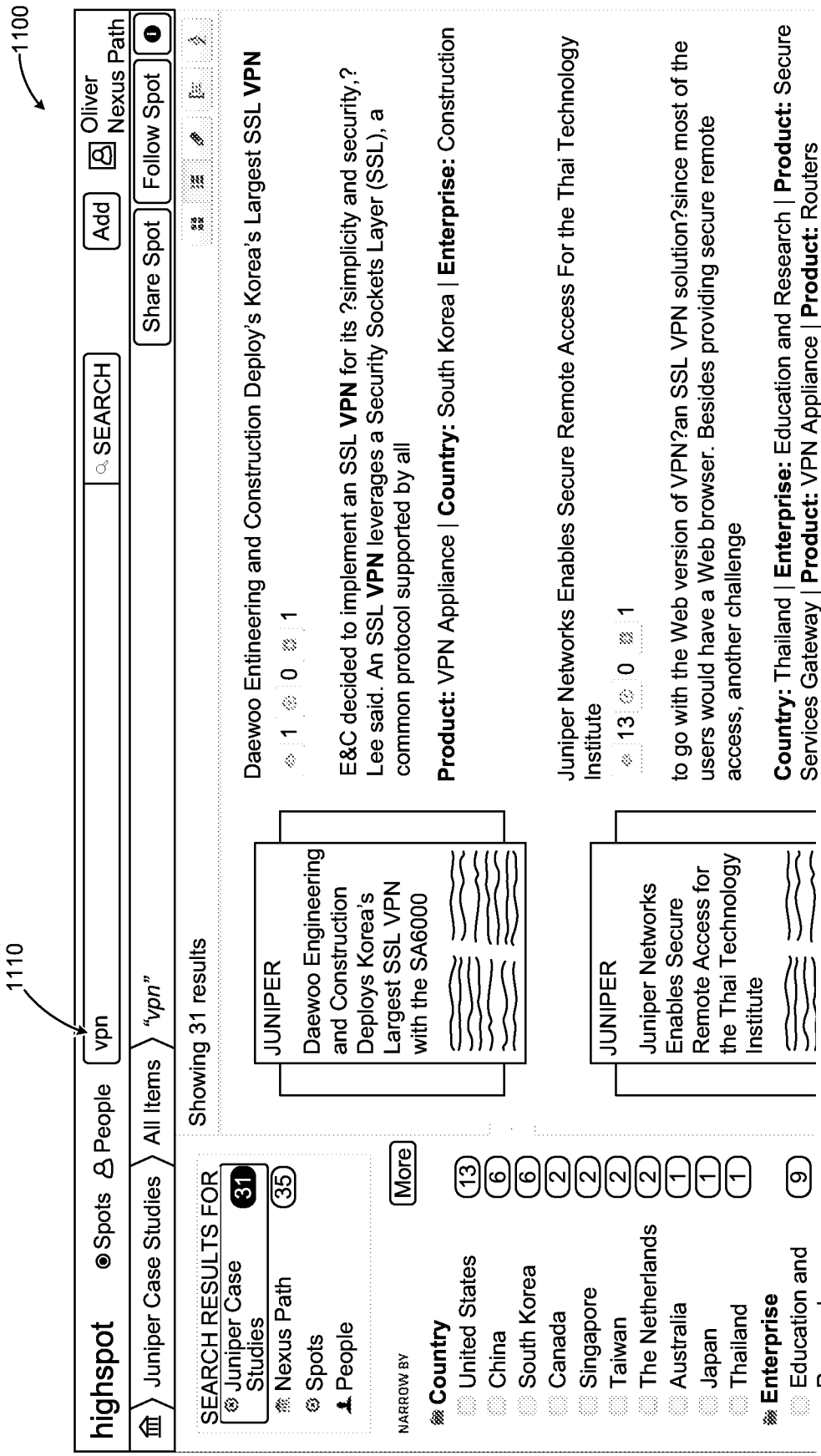


Figure 11: Screen Shot of Basic Search

FIG. 11

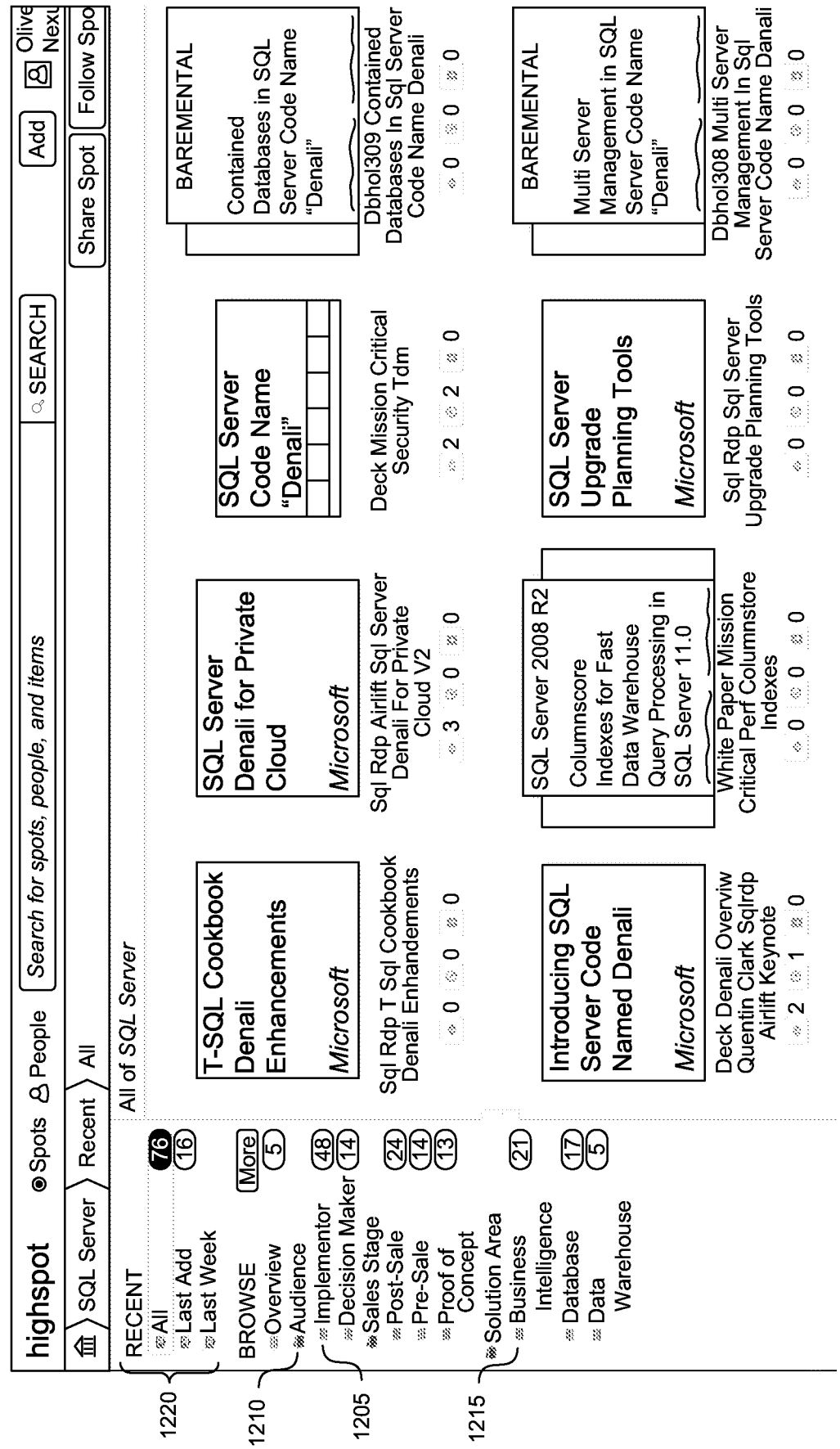

1200
highspot
Spots
People
Search for spots, people, and items
SEARCH
Add
Olive Nexu
SQL Server
Recent
All
Share Spot
Follow Spo
All of SQL Server
RECENT
All 76
Last Add 16
Last Week
BROWSE
Overview More 5
Audience
Implementor 48
Decision Maker 14
Sales Stage
Post-Sale 24
Pre-Sale 14
Proof of Concept 13
Solution Area
Business Intelligence 21
Database 17
Data Warehouse 5
1220
1210
1205
1215
T-SQL Cookbook Denali Enhancements
Microsoft
Sql Rdp T Sql Cookbook Denali Enhandements
0 0 0
SQL Server Denali for Private Cloud
Microsoft
Sql Rdp Airlift Sql Server Denali For Private Cloud V2
3 0 0
SQL Server Code Name "Denali"
Deck Mission Critical Security Tdm
2 2 0
BAREMENTAL
Contained Databases in SQL Server Code Name "Denali"
Dbhol309 Contained Databases In Sql Server Code Name Denali
0 0 0
Introducing SQL Server Code Named Denali
Microsoft
Deck Denali Overviw Quentin Clark Sqlrdp Airlift Keynote
2 1 0
SQL Server 2008 R2 Columnscore Indexes for Fast Data Warehouse Query Processing in SQL Server 11.0
White Paper Mission Critical Perf Columnstore Indexes
0 0 0
SQL Server Upgrade Planning Tools
Microsoft
Sql Rdp Sql Server Upgrade Planning Tools
0 0 0
BAREMENTAL
Multi Server Management in SQL Server Code Name "Denali"
Dbhol308 Multi Server Management In Sql Server Code Name Danali
0 0 0

FIG. 12

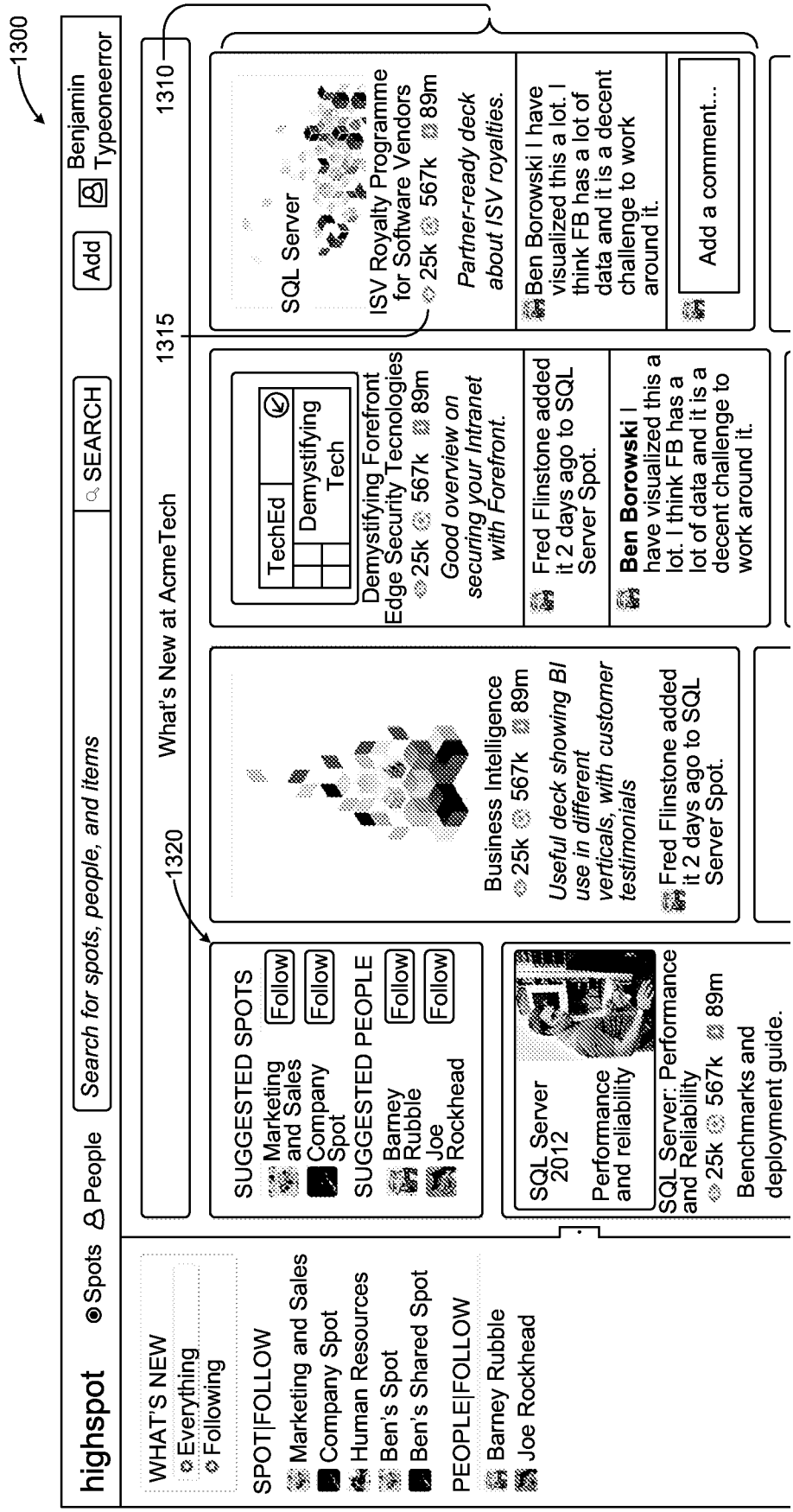

1300
highspot
Spots
People
Search for spots, people, and items
SEARCH
Add
Benjamin Typeoneerror
WHAT'S NEW
Everything
Following
SPOT|FOLLOW
Marketing and Sales
Company Spot
Human Resources
Ben's Spot
Ben's Shared Spot
PEOPLE|FOLLOW
Barney Rubble
Joe Rockhead
1320
What's New at AcmeTech
1315
1310
SUGGESTED SPOTS
Marketing and Sales
Follow
Company Spot
Follow
SUGGESTED PEOPLE
Barney Rubble
Follow
Joe Rockhead
Follow
SQL Server 2012
Performance and reliability
SQL Server: Performance and Reliability
25k 567k 89m
Benchmarks and deployment guide.
Business Intelligence
25k 567k 89m
Useful deck showing BI use in different verticals, with customer testimonials
Fred Flinstone added it 2 days ago to SQL Server Spot.
TechEd
Demystifying Tech
Demystifying Forefront Edge Security Tecnologies
25k 567k 89m
Good overview on securing your Intranet with Forefront.
Fred Flinstone added it 2 days ago to SQL Server Spot.
Ben Borowski I have visualized this a lot. I think FB has a lot of data and it is a decent challenge to work around it.
SQL Server
ISV Royalty Programme for Software Vendors
25k 567k 89m
Partner-ready deck about ISV royalties.
Ben Borowski I have visualized this a lot. I think FB has a lot of data and it is a decent challenge to work around it.
Add a comment...

FIG. 13

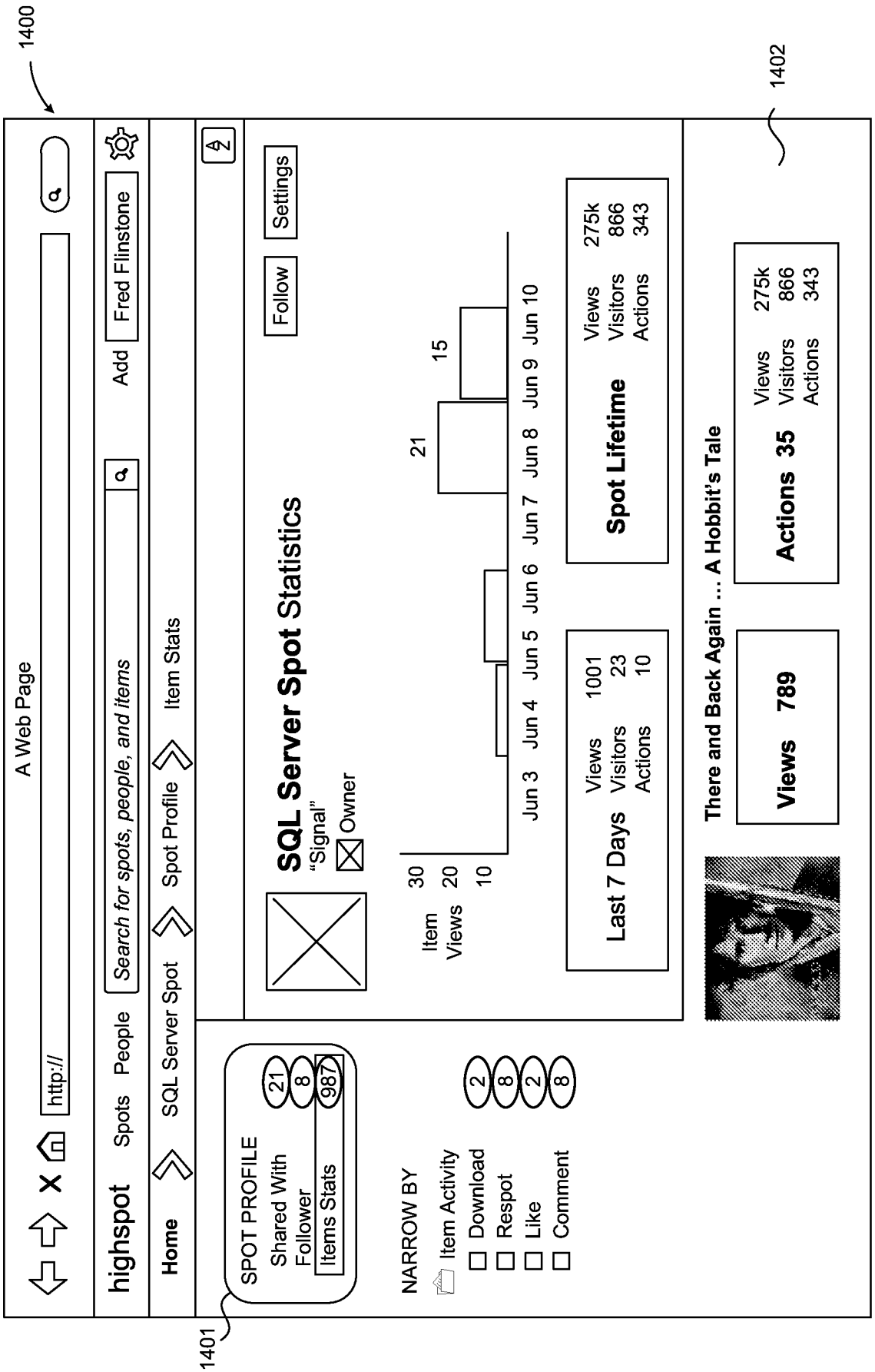

1400
A Web Page
http://
highspot
Spots
People
Search for spots, people, and items
Add
Fred Flinstone
Home
SQL Server Spot
Spot Profile
Item Stats
SPOT PROFILE
Shared With
Follower
Items Stats
21
8
987
NARROW BY
Item Activity
Download
Respot
Like
Comment
2
8
2
8
1401
SQL Server Spot Statistics
"Signal"
Owner
Follow
Settings
Item Views
30
20
10
Jun 3
Jun 4
Jun 5
Jun 6
Jun 7
Jun 8
Jun 9
Jun 10
21
15
Last 7 Days
Views 1001
Visitors 23
Actions 10
Spot Lifetime
Views 275k
Visitors 866
Actions 343
There and Back Again ... A Hobbit's Tale
Views 789
Actions 35
Views 275k
Visitors 866
Actions 343
1402

FIG. 14

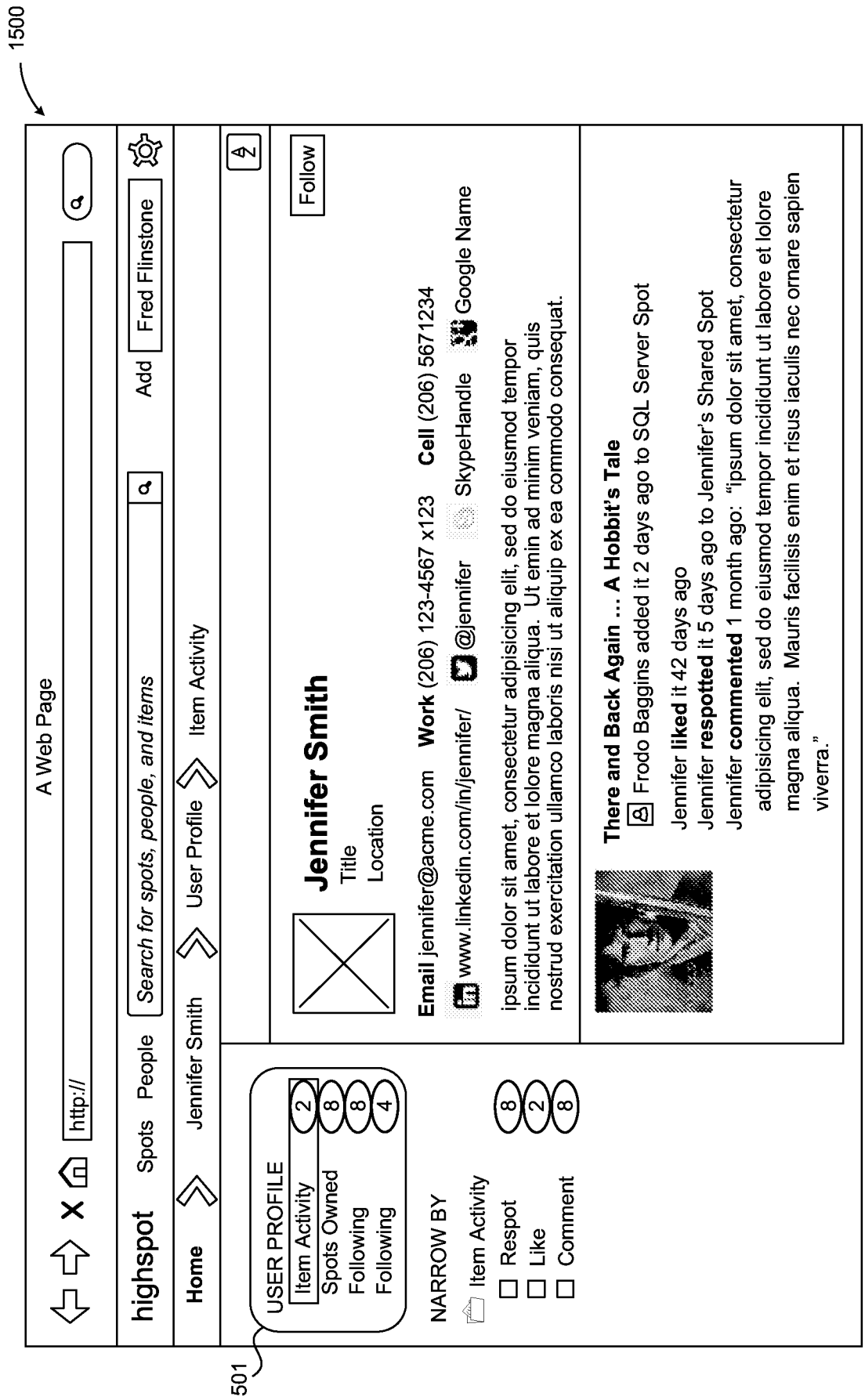

1500
A Web Page
http://
highspot
Spots
People
Search for spots, people, and items
Add
Fred Flinstone
Home
Jennifer Smith
User Profile
Item Activity
2
1501
USER PROFILE
Item Activity 2
Spots Owned 8
Following 8
Following 4
NARROW BY
Item Activity
Respot 8
Like 2
Comment 8
Jennifer Smith
Title
Location
Follow
Email jennifer@acme.com Work (206) 123-4567 x123 Cell (206) 5671234
www.linkedin.com/in/jennifer/ @jennifer SkypeHandle Google Name
ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et lolore magna aliqua. Ut emin ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.
There and Back Again ... A Hobbit's Tale
Frodo Baggins added it 2 days ago to SQL Server Spot
Jennifer liked it 42 days ago
Jennifer respotted it 5 days ago to Jennifer's Shared Spot
Jennifer commented 1 month ago: "ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et lolore magna aliqua. Mauris facilisis enim et risus iaculis nec ornare sapien viverra."

FIG. 15

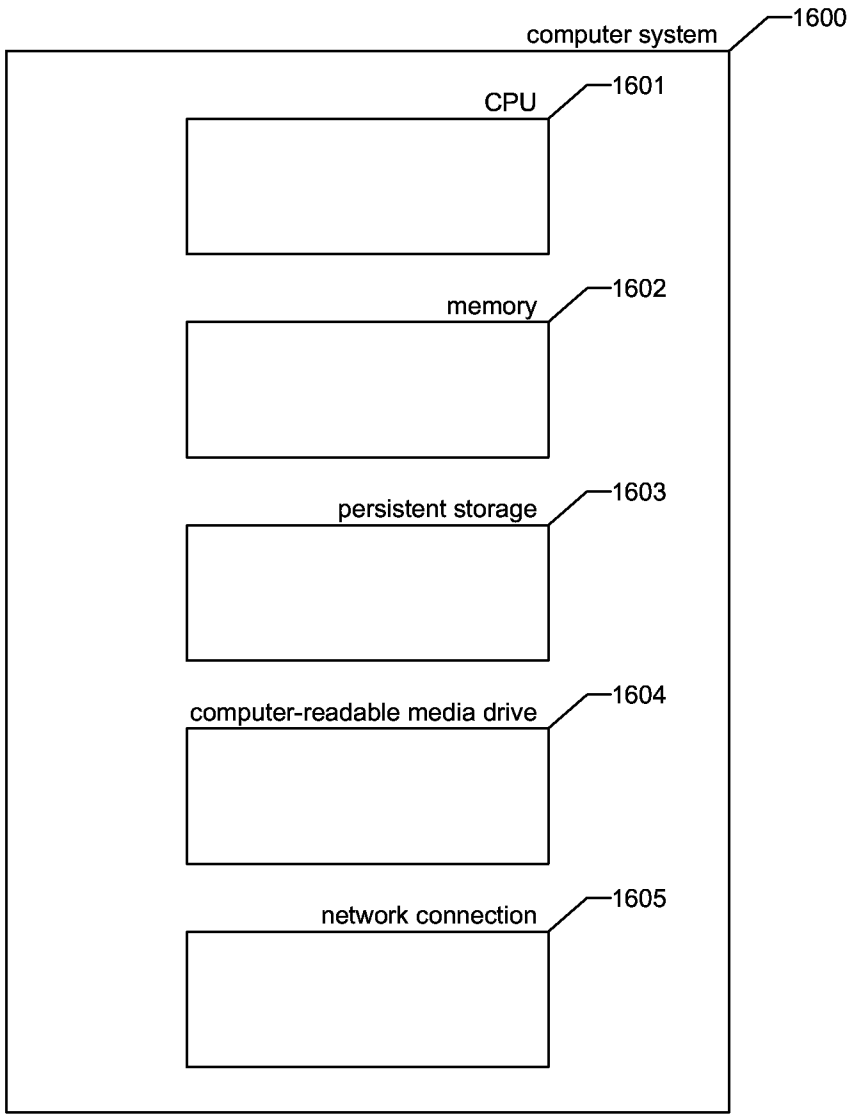

computer system
1600
CPU
1601
memory
1602
persistent storage
1603
computer-readable media drive
1604
network connection
1605

FIG. 16

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2014/029505

A. CLASSIFICATION OF SUBJECT MATTER

G06Q 50/10(2012.01)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)
G06Q 50/10; G06F 17/24; G06F 17/30; G06F 17/60

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
eKOMPASS(KIPO internal) & Keywords: interest, graph, organization, information, affinity, feed, request, profile

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2012-0278329 A1 (GEOFFREY BORGGAARD et al.) 01 November 2012 See paragraphs [0066]-[0072], claims 1-7 and figures 1-4.	1-20
Y	US 2005-0267799 A1 (WESLEY CHAN et al.) 01 December 2005 See paragraphs [0044]-[0060], claims 1,13-17 and figures 3-5.	1-20
A	US 2012-0143859 A1 (DIMITRIOS LYMPEROPOULOS et al.) 07 June 2012 See abstract, claims 1,10,16 and figures 1-3.	1-20
A	US 2010-0198869 A1 (CHRIS KALABOUKIS et al.) 05 August 2010 See paragraphs [0042]-[0061], claims 1,20,36 and figures 4-5.	1-20
A	US 2006-0085427 A1 (MARK S. D`URSO) 20 April 2006 See paragraphs [0035]-[0040], claims 18,31 and figures 1-2.	1-20

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

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

Date of the actual completion of the international search	Date of mailing of the international search report
04 July 2014 (04.07.2014)	**07 July 2014 (07.07.2014)**

Name and mailing address of the ISA/KR
International Application Division
Korean Intellectual Property Office
189 Cheongsa-ro, Seo-gu, Daejeon Metropolitan City, 302-701,
Republic of Korea
Facsimile No. +82-42-472-7140

Authorized officer
HONG, Kyoung Hee
Telephone No. +82-42-481-5781

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/US2014/029505

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2012-0278329 A1	01/11/2012	AU 2012-202764 A1	15/11/2012
		CA 2834339 A1	01/11/2012
		CN 103493089 A	01/01/2014
		EP 2710548 A2	26/03/2014
		KR 10-2013-0135977 A	11/12/2013
		WO 2012-148924 A2	01/11/2012
		WO 2012-148924 A3	31/01/2013
US 2005-0267799 A1	01/12/2005	AU 2005-242773 A1	24/11/2005
		CA 2566740 A1	24/11/2005
		CN 101076822 A0	21/11/2007
		EP 1774462 A2	18/04/2007
		EP 1774462 A4	20/08/2008
		KR 10-1384072 B1	09/04/2014
		KR 10-2007-0011591 A	24/01/2007
		KR 10-2012-0123148 A	07/11/2012
		WO 2005-111897 A2	24/11/2005
		WO 2005-111897 A3	21/12/2006
US 2012-0143859 A1	07/06/2012	CN 102591911 A	18/07/2012
		EP 2646973 A2	09/10/2013
		KR 10-2013-0130731 A	02/12/2013
		WO 2012-075273 A2	07/06/2012
		WO 2012-075273 A3	02/08/2012
US 2010-0198869 A1	05/08/2010	KR 10-1226403 B1	24/01/2013
		KR 10-2010-0089030 A	11/08/2010
		TW 201042476 A	01/12/2010
		TW 201042476 A	01/12/2010
US 2006-0085427 A1	20/04/2006	CN 1760867 A	19/04/2006
		CN 1760867 B	21/03/2012
		CN 1760867 C0	19/04/2006
		EP 1647904 A2	19/04/2006
		EP 1647904 A3	09/08/2006
		JP 2006-114021 A	27/04/2006
		JP 2012-027937 A	09/02/2012
		JP 2012-027938 A	09/02/2012
		JP 5068437 B2	07/11/2012
		KR 10-1159342 B1	25/06/2012
		US 2006-0085397 A1	20/04/2006
		US 2006-0085447 A1	20/04/2006
		US 2014-081947 A1	20/03/2014
		US 7779012 B2	17/08/2010
		US 8595223 B2	26/11/2013