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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
7 June 2012 (07.06.2012)

WIPO | PCT

(10) International Publication Number
WO 2012/075335 A2

(51) International Patent Classification: Not classified

(21) International Application Number:
PCT/US2011/062956

(22) International Filing Date:
1 December 2011 (01.12.2011)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/418,818 1 December 2010 (01.12.2010) US

(71) Applicant *(for all designated States except US)*: **GOOGLE INC.** [US/US]; 1600 Amphitheatre Parkway, Mountain View, CA 94043 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **LIEBALD, Benjamin** [DE/US]; c/o Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043 (US). **NANDY, Palash** [IN/FR]; c/o Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043 (US). **SAMPATH, Dasarathi** [IN/US]; c/o Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043 (US). **LIU, Junning** [CN/US]; c/o Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043 (US). **NIU, Ye** [CN/US]; c/o Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043 (US). **ILVENTO, Christina** [US/US]; c/o Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043 (US). **CHEN, Yu-To** [CN/US]; c/o Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043 (US). **DAVIDSON, Jamie** [US/US]; c/o Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043 (US).

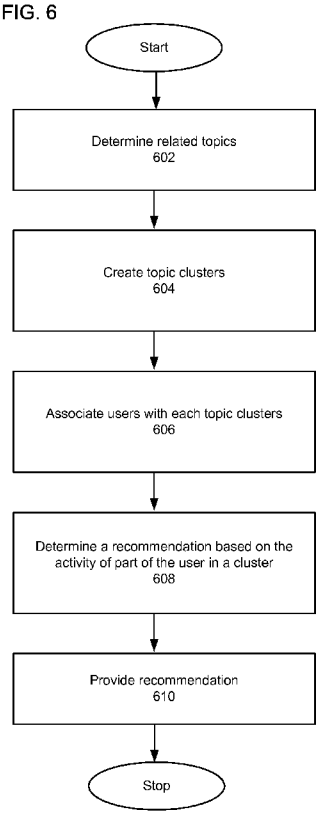
(74) Agents: **SACHS, Robert, R.** et al.; Fenwick & West LLP, Silicon Valley Center, 801 California Street, Mountain View, CA 94041 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN,

[Continued on next page]

(54) Title: RECOMMENDATIONS BASED ON TOPIC CLUSTERS

FIG. 6



(57) Abstract: A system and method for developing a user's profile based on the user's interaction with content items. A module on the client rendering the content items or the service including the content items tracks the user's interactions with the content items and transmits the tracked data to a user analysis module. The user analysis module determines the topics associated with the interacted upon content items. The user analysis module then selects the topics for the user's profiles based on the received tracked data and the associated topics. The selected topics are mapped to topic clusters and the topic clusters are stored in association with the user profile. Recommendations for a user are made based on the topic clusters associated with the user' profile.

WO 2012/075335 A2

HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, 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, MD, 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, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

RECOMMENDATIONS BASED ON TOPIC CLUSTERS

BACKGROUND

FIELD OF DISCLOSURE

[0001] The disclosure generally relates to creating and storing user profiles based on content consumption.

DESCRIPTION OF THE RELATED ART

[0002] Content hosting services generally attempt to present content that is generally of interest to its users. Some content hosting services allow users to create user profiles that indicate demographic information (e.g., gender, age), as well as areas of interests or content topics. The content hosting service then attempts to use such profiles to select content to provide to the users. However, the users may not be able to articulate all their interests while populating their profile. Additionally, users' interests typically change over time and the users may not update their profiles to reflect these changes.

SUMMARY

[0003] A user's profile is created based on the user's interaction with content items in a content hosting service. A user's interactions with the content items on the content hosting service are recorded. A user analysis module determines topics associated with the content items with which the user has interacted. The user analysis module then selects the topics for the user's profiles based recorded interactions and the associated topics. A user profile is created which represents the selected topics. In one embodiment, the topics associated with the content items have associated topic strengths and the user analysis module selects the topics for user's profiles based on the topic strengths. In another embodiment, the user's interactions with various content items have associated interaction strengths and the user analysis module selects the topics for user's profiles based on the associated interaction strengths, and stores the topic association strengths for the selected topics in the user profile. In one embodiment, topics in the user profile are mapped to clusters of topics, and the mapped cluster topics replace or accompany the user topics in the user profile. Various user-cluster communities are formed that include users whose profiles have a common topic cluster. Recommendations to these user-

cluster communities can be made based on the user interactions of some of the users in a community.

[0004] The features and advantages described in the specification are not all inclusive and, in particular, many additional features and advantages will be apparent to one of ordinary skill in the art in view of the drawings, specification, and claims. Moreover, it should be noted that the language used in the specification has been principally selected for readability and instructional purposes, and may not have been selected to delineate or circumscribe the disclosed subject matter.

BRIEF DESCRIPTION OF DRAWINGS

[0005] Fig. 1 illustrates a system for determining and storing the users' profile including their areas of interest according to one embodiment.

[0006] Fig. 2 is a flow diagram illustrating a method for determining and storing the users' profile including their areas of interest according to one embodiment.

[0007] Fig. 3 is a block diagram illustrating the user analysis module that determines and stores the user profiles according to one embodiment.

[0008] Fig. 4 is a screen illustrating an interface for receiving users' areas of interests for storage in their profiles according to one embodiment.

[0009] Fig. 5 illustrates a co-occurrence matrix that stores co-occurrence strengths indicating the measure of co-occurrence of a first topic with another topic according to one embodiment.

[0010] Fig. 6 illustrates a method for providing recommendations based on the users' interactions according to one embodiment.

DETAILED DESCRIPTION

[0011] The computing environment described herein enables determination and storage of user profiles that represent, for each user, a set of topics indicative of the user's interests, based on the user's interaction with content items. The figures and the following description describe certain embodiments by way of illustration only. One skilled in the art will readily recognize from the following description that alternative embodiments of the structures and methods illustrated herein may be employed without departing from the principles described herein. Reference will now be made in detail to several embodiments, examples of which are illustrated in the accompanying figures. It is noted that wherever practicable similar or like reference numbers may be used in the figures and may indicate similar or like functionality.

SYSTEM ENVIRONMENT

[0012] Fig. 1 illustrates a system for determining and storing user profiles. A video hosting service 100 includes a front end web server 140, a video serving module 110, a video database 155, a user analysis module 120, a user access log 160, a topic repository 164 and a profile repository 166. Video hosting service 100 is connected to a network 180. FIG. 1 also includes a client 170 and third-party service 175 having an embedded video 178.

[0013] Many conventional features, such as firewalls, load balancers, application servers, failover servers, network management tools and so forth are not shown so as not to obscure the features of the system. A suitable service for implementation of the system is the YOUTUBE™ service, found at www.youtube.com; other video hosting services are known as well, and can be adapted to operate according to the teaching disclosed here. The term "service" represents any computer system adapted to serve content using any internetworking protocols, and is not intended to be limited to content uploaded or downloaded via the Internet or the HTTP protocol. In general, functions described in one embodiment as being performed on the server side can also be performed on the client side in other embodiments if appropriate. In addition, the functionality attributed to a particular component can be performed by different or multiple components operating together.

[0014] The servers and modules described herein are implemented as computer programs executing on server-class computer comprising a CPU, memory, network interface, peripheral interfaces, and other well known components. The computers themselves in some embodiments run a conventional proprietary or open-source operating system such as Microsoft Windows, Mac OS, Linux, etc., have generally high performance CPUs, gigabytes or more of memory, and gigabytes, terabytes, or more of disk storage. Of course, other types of computers can be used, and it is expected that as more powerful computers are developed in the future, they can be configured in accordance with the teachings here. The functionality implemented by any of the elements can be provided from computer program products that are stored in tangible computer readable storage mediums (e.g., RAM, hard disk, or optical/magnetic media).

[0015] A client 170 connect to the front end server 140 via network 180, which is typically the internet, but can also be any network, including but not limited to any combination of a LAN, a MAN, a WAN, a mobile, wired or wireless network, a private network, or a virtual private network. While only a single client 170 is shown, it is understood that very large numbers (e.g., millions) of clients can be supported and can be in communication with the video hosting service 100 at any time. The client 170 may include a variety of different computing devices. Examples of client devices 170 are personal computers, digital assistants, personal digital assistants, cellular phones, mobile phones, smart phones or laptop computers. As will be clear to one of ordinary skill in the art, the present invention is not limited to the devices listed above.

[0016] The client includes a browser or a dedicated application that allows client 170 to present content provided on the video hosting service 100. Suitable applications include, for example, Microsoft Internet Explorer, Netscape Navigator, Mozilla Firefox, Apple Safari, and Google Chrome. The browser can also include or a support a plug-in for a video player (e.g., Flash™ from Adobe Systems, Inc.), or any other player adapted for the video file formats used in the video hosting service 100. Alternatively, videos can be accessed by a standalone program separate from the browser.

[0017] The digital content items can include, for example, video, audio or a combination of video and audio. Alternatively, a digital content item may be a still

image, such as a JPEG or GIF file or a text file. For purposes of convenience and the description of one embodiment, the digital content items will be referred to as a "video," "video files," or "video items," but no limitation on the type of digital content items are indented by this terminology. Other suitable types of digital content items include audio files (e.g. music, podcasts, audio books, and the like), documents, images, multimedia presentations, and so forth.

[0018] The video hosting service 100 provides videos that have been uploaded by other users of the video hosting service 100, or may have been provided by the video hosting service operator, or by third parties. Clients 170 can search for videos based on keywords or other metadata. These requests are received as queries by the front end server 140 and provided to the video serving module 110, which is responsible for searching the video database 155 for videos that satisfy the user queries and providing the videos to the users. The video serving module 110 supports searching on any fielded data for a video, including its title, description, metadata, author, category and so forth. Alternatively, users can browse a list of videos based on categories such as most viewed videos, sports, animals, or automobiles. For example, the user may browse a list of videos related to cars and select which videos from the list to view.

[0019] Video database 155 stores videos provided to clients 170. Each video in one embodiment has a video identifier (id). Each video file has associated metadata associated that includes video ID, author, title, description, and keywords, additional metadata can be included as available. The metadata also includes one or more topics that are associated with the video. The associated topics may include topics created by a community in a collaborative knowledge base like Freebase. Alternatively, the topics may be selected from the frequently occurring topics occurring in the titles, descriptions, and user comments of the videos, for example the 100,000 most frequently occurring term unigrams or bigrams.

[0020] In one embodiment, each topic is associated with a topic strength TS representing the topics' degree of association with the video. The topic strength for a particular topic and video is based on content analysis of the video, users' comments for the video, or other metadata associated with the video. Alternatively, instead of being stored with the metadata of each video, the topics and topic strength information can be stored in a separate database.

[0021] In one embodiment, the topic strength for a video is also adjusted based on the usefulness of a topic. The usefulness of a topic is a weight reflecting how useful is a topic to a system in representing the topic's association with the video. For example, the system operator may not prefer topics that represent racy or objectionable content and therefore the usefulness weight for such topics may be a low or a negative value. In another example, the usefulness of a topic is based on the frequency of topic in the corpus.

[0022] The user access log 160 stores access data describing the user's access and interactions with videos. The access data indicates whether a user watched an entire video, watched a video for a particular duration, skipped a video, scrolled up or down through a web page including a video, shared a video with other users, added a video to a playlist, flagged a video, blocked a video from a playlist or a collection of videos, favorited a video, gave a video a favorable rating (e.g. liked a video using a FACEBOOK™ account or +1'd a video using a GOOGLE+™ account), gave a video an unfavorable rating (e.g. "thumbs down"). In one embodiment, the user access log 160 or another entity associated with the user access log 160 provides the users with the opportunity to opt-out of having the users' access data collected and/or shared with other modules in the video hosting service 100 or other services.

[0023] The profile repository 164 stores the user profiles. A user profile includes a set of topics for a user. This set of topics represents the user's interest and the list may be partly populated by receiving a number of topics from the user. The user profile may include the topics as a list of topics (e.g., as terms or topic identifiers), or as vector (e.g., bit map, or vector of real valued weights). Additionally, the list is populated by the user analysis module 120. The topics stored in a user's profile can be used for various purposes. For example, the topics can be displayed as user's area of interest on the user's home page in a social network or a content hosting network. Additionally, the topics may be used to suggest to the user content, content channels, products, services, additional topics etc. that may be of interest to the user. The suggestions may be provided to the user on the user's home page or another web page like a "browse" page where a user may browse through various topics that may be of interest to the user.

[0024] In one embodiment, the topics displayed on the user's home page or browse page are selectable (for e.g. through a hyperlink). A user may select a topic and the selection leads the user to a web page partly or wholly dedicated to the selected topic. The selected topic's web page includes content related to the selected topic, like related multimedia content or textual content. Additionally, the topic's web page may include links to other related topics' web pages. These related topics may be displayed as topics related to the selected topic or recommended topics for a user visiting the selected topic's web page.

[0025] The user analysis module 120 determines and stores a user profile based on the videos accessed by the user, and is one means for performing this function. Fig. 2 illustrates method executed by the user analysis module 120 for determining and storing the topics for a user profile. To determine the topics, the user analysis module 120 queries the user access log 160 and determines 202 videos accessed by the user. This set of videos can be all videos accessed by the user, or just those accessed by the user within a certain time period, such as the previous thirty days.

[0026] The user analysis module 120 analyzes the user's access data stored in the user access log 160 and determines 204 the user's interactions with the accessed videos. The user analysis module 120 also determines 204 the user's interaction strength for each accessed video based on factors like the type of user's interaction with the accessed video. The user analysis module 120 also queries the video database 155 and determines 206, for each video accessed by the user, the topics associated with the accessed videos and the video's topic strengths indicating the video's degree of association with the topics. Based on the determined interaction strengths and topic strengths, the user analysis module 120 selects 208 and stores 210 topics in the user's profile.

[0027] The user analysis module 120 also determines and provides recommendations based on the users' interaction with the videos, and is also one means for performing this function. The recommendations can be recommended videos for the user or recommended topics for the videos. The operation of the user analysis module 120 to provide recommendation is further described below with respect to FIG. 6.

[0028] Fig. 3 is a block diagram illustrating the user analysis module 120 according to one embodiment. The user analysis module 120 comprises a user interaction module 302, an interaction strength module 304, a user profile module 306, a related topics module 308, a topic cluster module 301 and a cluster recommendation module.

[0029] The user interaction module 302 receives feedback regarding the users' interactions with videos and stores the received feedback as access data in the user access log 160. A module (not shown) in the client 170 (or the service 175) tracks data about the user's interactions (e.g. pause, rewind, fast forward). Additional user's interactions (e.g. the user requesting a video, rating a video, sharing a video) are tracked by a module (not shown) in the video hosting service 100 or at another service like a social networking service. Regardless of where the data is tracked, the data is transmitted to the user interaction module 302. The user interaction module 302 receives the transmitted data and stores the received data in the user access log 160 as access data. Examples of access data stored in access log 160 are described above. The user interaction module 302 repeatedly receives feedback regarding the user's interactions with various videos and updates the access data for the user based on the received feedback.

[0030] The interaction strength module 304 analyzes the access data for a user and determines an interaction strength IS_i indicating a user's degree of association with a particular video v_i. To determine the IS value, the interaction strength module 304 assigns different weights to different types of user's interactions with the video. For example, a user starting a video may be assigned a weight of 0.5, a user watching at least 80% of the video may be assigned a weight of 0.75, a user giving a favorable rating for the video may be assigned a weight of 1.5, a user favoriting a video may be assigned a weight of 2.0, and a user subscribing to a channel of videos associated with the watched video or with the user who uploaded the watched video may be assigned a weight of 5.0. The interaction strength module 304 assigns greater weight to the user's interactions indicating a greater involvement with a video. For example, the interaction strength module 304 assigns a greater weight to a user adding a video to a playlist, or sharing a video with others, than to the user watching the video. Additionally, the interaction strength module 304 adjusts the weight for a particular interaction based on the frequency or duration of

the interaction. For example, the interaction strength module 304 assign a greater weight to a user's view of a particular video if the user has viewed the video a number of times instead of just once or for a ten minute duration instead of thirty seconds. In one embodiment, the interaction strength module 304 normalizes the adjusted weights based on the total number of videos the user has interacted with, the total number of times the user has interacted with the videos, or the total amount of time the user has spent interacting with the videos.

[0031] The interaction strength module 304 assigns negative or relatively low weights to certain interactions indicating the user's lack of interest in a particular video. For example, skipping a presented video, flagging a video, or blocking a video from a playlist may be assigned a negative weight.

[0032] In one embodiment, the interaction strength module 304 discounts the weight based on their age. For example, the interaction strength module 304 exponentially decays the weight associated with a user interaction based on the amount of time elapsed since the user interaction occurred. Accordingly, a user interaction that occurred recently is assigned a higher weight than a user interaction that occurred at an earlier time.

[0033] After assigning and adjusting weights for the user's interactions with a particular video, the interaction strength module 304 determines and stores an interaction strength IS indicating the strength of the user's interactions or association with the video. The interaction strength is based on the assigned and adjusted weights. For example, the interaction strength is a sum or product of the assigned and adjusted weights.

[0034] As described above, the user analysis module 120 determines for a user, the videos v_i the user has interacted with (from the user access log 160) and the user's interaction strength IS_i for each of these videos (determined by the interaction strength module 304). Also, as described above, the user analysis module 120 determines for each of these videos v_i, topics t associated with the video (from the video database 155) and, for each of the associated topic t_k, a topic strength TS_k indicating the topic's degree of association with the video (from the video database 155).

[0035] Based on this information, the user profile module 306 determines a set T of topics for a user's profile. To determine the topics T for a user profile, the

user profile module 306 sorts the videos v_i the user interacted with based on the topics t_k associated with the videos. The sort results in sets $S=\{s_1, s_2, s_3 \ldots s_j\}$ of topics such that each set s_j includes a topic t_k and its associated user's videos $v_{i,k}$. The user profile module 306 selects a number of the topic sets s, where each selected set has a minimum number of videos, e.g., each selected topic set has at least 20 videos. The topics t_k of the selected sets s form the set T topics for the user's profile.

[0036] Alternatively, the user profile module 306 determines the set T of topics for a user profile based on a topic association strength TAS determined for each set s, where TAS_j indicates the degree of association between set s_j's topics t and the user. To determine the topic association strength TAS_j for a particular set s_j of topics t_k, the user profile module 306 combines the topic strengths TS_k of the set's topics t_k for each of the videos v_i in the set s_j. Combining the topic strengths TSs may occur by adding, averaging, or applying another arithmetic or statistical function to the topic strengths TSs. After determining the topic association strength TAS_j for each set s_j in S, the user profile module 306 selects a number of these sets based on the sets topic association strengths TAS_j. For example, the user association module 306 may select fifty sets s with fifty highest topic association strengths TAS. The topics t_k of the selected sets s form the set T topics for the user's profile.

[0037] The user profile module 306 also stores in the user's profile the topic association strengths TAS associated with the stored topics. The user profile module 306 can be configured to periodically updates the stored topics in a user's profile using the process described above, based on the videos that the user interacted with since a prior update.

[0038] Additionally, in one embodiment, the user profile module 306 receives topics that are related with the topics stored in a user profile and stores the related topics in the user profile. The user profile module 306 receives the related topics from the related topics module 308. Related topics module 308 accesses the topics in a user's profile and determines additional topics related to the profile's topics.

[0039] There are several different ways that the related topics module 308 can determine related topics. These include a demographic approach, a topic co-occurrence approach, and a combined demographic and topic co-occurrence approach. Additional approaches to determine related topics would be apparent to

one of ordinary skill in the art in light of the disclosure herein. For example, related topics may also be determined based on topics' relationships specified in a knowledgebase like Freebase.

[0040] Related Topics based on Demographics

[0041] In one embodiment, related topics module 308 determines related topics based on the popularity of various topics in each of a number of demographic groups. In this embodiment, the related topics module 308 organizes the user profiles in the profile corpus based on one or more demographic category, such as gender and age group. For example, the related topics module 308 can organize the user profiles into twelve demographic groups D_z of profiles based on the user's gender (male, female) and age group (e.g., 13-17, 18-24, 25-34, 35-44, 45-54; 55+). The related topics module 308 then determines, for each demographic group D_z of user profiles, a number of most frequently occurring topics t (e.g., the top 50 most frequently occurring topics); this forms the related topic set R_z for the demographic group D_z. Then for a given demographic group D_z, the related topics module 308 adds the related topics R_z to each user profile in D_z. If a topic t in R_z is already present in the user profile, then it can be handled either by skipping it, or by increasing its topic association strength TAS.

[0042] Related Topics based on Topic Co-Occurrence

[0043] In another embodiment, the related topics module 308 uses the co-occurrence of topics in the user profiles to determine which topics are related to each other. To determine the related topics, the related topics module 308 initially determines, across a collection of user profiles (e.g., all user profiles in the system), pairs of topics (t_i, t_j) that co-occur in at least some of the user profiles in the collection, and from there determines a measure of co-occurrence for each topic pair. The determination of these co-occurring topics is described in regards to Fig. 5 below. The related topics module 308 then determines for each topic t_k in the corpus, the most closely related topics t_l based on the co-occurrence measure. Next, given a user profile with topics t_j, the related topics module 308 adds to the user profile for each topic t_j the most closely related topics t_l.

[0044] Fig. 5 illustrates a co-occurrence matrix 500 that stores co-occurrence strengths $CS_{i,j}$ indicating the measure of co-occurrence of a topic t_i with another topic t_j. One of ordinary skill in the art will understand that the illustrated co-occurrence

matrix 500 is simply a graphical representation of co-occurrence strengths CSs used to aid the description of the related topics module 308, and that the matrix 500 may be stored in various data structures like arrays, lists etc. Given n topics t, the co-occurrence matrix 500 is an nXn matrix. Each row 502a-n represents a topic t_i and each column 504a-n represents a topic t_j. Each cell, like cell 508 represents the co-occurrence strength $CS_{i,j}$ of for the pair of topics t_i and t_j.

[0045] The co-occurrence strength $CS_{i,j}$ for the pair of topics (t_i, t_j) may be determined as follows. As noted above, each topic t_i in user profile has a topic association strength TAS_i. Thus, for a pair of topics t_i and t_j co-occurring in a given user profile, the related topics module 308 computes a profile co-occurrence strength $PCS_{i,j}$ based on the topic association strengths TAS_i and TAS_j. The profile co-occurrence strength $PCS_{i,j}$ may be a product, sum, average, or another arithmetic or statistical function of the pair's topic association strengths TAS_i and TAS_j. The co-occurrence strength $CS_{i,j}$ is then the combined $PCS_{i,j}$ summed across all user profiles in which topics t_i and t_j co-occur. Each $PCS_{i,j}$ is then normalized by the frequency of topic t_i in the profile corpus. In other embodiments, combining may include averaging, adding, or performing another arithmetic or statistical function on the profile co-occurrence strengths PCS.

[0046] An example illustrated in Fig. 5 assists in describing the method for computing the co-occurrence strengths (CSs). In Fig. 5, cell 508 includes the co-occurrence strength (CS) for topic t_i (topic for intersecting row 502i) co-occurring with topic t_j (topic for intersecting column 504j) in the profile corpus used to select topics for the co-occurrence matrix 500. This co-occurrence strength (CS) is a normalized sum of topic association strengths (TASs) of t_i and t_j for corpus' profiles that include both these topics. The sum of the topic association strengths (TASs) has been normalized by the frequency of t_i's appearance in corpus' profiles. Similarly, cell 506 includes the co-occurrence strength (CS) for topic t_j co-occurring with topic t_i. This co-occurrence strength (CS) is also a normalized sum of topic association strengths (TASs) of t_i and t_j, but this sum has been normalized by the frequency of t_j's, not t_i's, appearance in the corpus' profiles.

[0047] After populating the co-occurrence matrix 500, the related topics module 308 identifies for each topic t_i (by row) a number of cells with the highest co-occurrence strengths CSs (e.g., 50 highest values), or the cells with co-occurrence

strengths CS beyond a threshold value (e.g., $CS_{i,j} > 75\%$ of maximum $CS_{i,j}$). These cells represent the set of topics R_i that are determined to be related to topic t_i.

[0048] The example illustrated in Fig. 5 further illustrates the method employed by the related topics module 308 to select related topics for topic T_j. In Fig. 5, assume that cells 508, 510 include the highest co-occurrence strengths $CS_{i,j}$ for topic t_j (represented by row 502j). The related topics module 308 identifies these cells 506, 508 as the cells with the highest co-occurrence strengths $CS_{i,j}$ and thus identifies topics t_i and t_n (the topics of the intersecting columns 504i, 504n for cells 506, 508) as topics related to topic t_j.

[0049] Finally, given a user profile of topics t, for each topic t_i therein the related profile module 308 adds the related topics R_j to the user profile. If a topic t in R_i is already present in the user profile, then it can be handled either by skipping it, or by increasing its topic association strength TAS.

[0050] Related Topics based on Demographics and Co-Occurrence

[0051] In one embodiment, the related topics module 308 determines related topics for a selected user from a profile corpus of users that are in same demographic group as the selected user. To determine these related topics, the related topics module 308 constructs for each demographic group D_z a co-occurrence matrix 500 from a set of user profiles belonging to that group. Then for each demographic group D_z, the related topics module 308 determines the related topics $R_{z,i}$ for each $topic_i$ in that that group's co-occurrence matrix.

[0052] User Selected Topics

[0053] In the foregoing embodiments, the related topics module 308 automatically adds related topics to each user's profile. Alternatively, the related topics module 308 can be configured to enable users to selectively add related topics to their individual user profiles. In one embodiment, the users may add topics, including related topics, to their own profiles through an interface such as the one illustrated in Fig. 4. The interface in Fig. 4 includes a profile topics column 406 and a related topics column 410. The profile topics column 406 includes the topics 412 associated with a user's profile based on the analysis of the user's interactions with videos. In response to a user selecting one or more of the topics 412 in the profile topics column 406, the related topics column 410 is updated to include topics 422a-n related to the selected topics 412. The related topics 422a-n are determined by the

related topics module 308 and presented to the user in the related topics column 410. The user may select one or more related topics 422a-n, and in response to such selection, these topics are added to the user's profiles. In one embodiment, the user profile module 306 also determines and stores with the additional topics their topic association strengths TAS.

[0054] Recommendations Based on Topic Clusters

[0055] Fig. 6 illustrates a method executed by the user analysis module 120 for providing recommendations based on the user's interactions. The user analysis module 120 determines 602 related topics based on the user profiles, using any of the previously described methods (demographic, co-occurrence, or demographic co-occurrence).

[0056] After determining the related topics, the user analysis module 120 creates 604 topic clusters with related topics. The creation of topic clusters is further described in the next section, Creating Topic Clusters.

[0057] Next, the user analysis module 120 associates 606 various users to the created topic clusters based on the topics in the topic clusters and the user profiles. The association of users to topic clusters is further described below under the heading Associating Users with Clusters.

[0058] The user analysis module 120 then monitors the activity of users associated with a cluster and determines 608 a recommendation, like a video for the associated users, based on the monitored activity. The user analysis module 120 provides 610 the recommendation for display to the users. The manner of making the recommendations is further described below under the heading Recommendations Based on Clusters.

[0059] Creating Topic Clusters

[0060] In one embodiment, the user profiles module 306 stores topic clusters, in addition to, or instead of, the topics in the user profiles. The topic clusters include a set of related topics. The user profiles module 306 receives the topic clusters from the topic clusters module 310.

[0061] The topic cluster module 310 creates topics cluster TC_i including topics occurring in user profiles. The topic cluster module 310 may create topics clusters based on clustering algorithms like hierarchical agglomerative clustering (HAC), probabilistic models like Latent Dirichlet Allocation (LDA), or vector

models, such as k-means (using rows in the co-occurrence matrix as topic vectors). In one embodiment, the topic cluster module 310 clusters topics from the co-occurrence matrix 500 using HAC. Again, the co-occurrence matrix 500 stores co-occurrence strengths $CS_{i,j}$ for pairs of topics (t_i, t_j) that co-occur in at least some of the user profiles in a collection of user profiles. To create topics clusters from the co-occurrence matrix 500, the topic cluster module 310 identifies the cell in the co-occurrence matrix 500 with the highest co-occurrence strength $CS_{i,j}$. After identifying the cell with the highest co-occurrence strength $CS_{i,j}$, the topic cluster module 310 clusters the topics (t_i, t_j) associated with the identified strength $CS_{i,j}$.

[0062] To cluster the associated topics, the topic cluster module 310 determines the co-occurring topics (t_i, t_j) associated with the identified cell and the rows and columns associated with the determined co-occurring topics. Assume for illustration purposes that the topic cluster module 310 identifies cell 506 in co-occurrence matrix 500 as the cell with the highest co-occurrence strength $CS_{i,j}$. The topic cluster module 310 determines that co-occurring topics t_i and t_j are associated with the identified cell 506 and combines the two topics into a cluster.

[0063] To combine the two topics (t_i, t_j), the topic cluster module 310 combines cells in one row with the adjacent cells in the other to get a combined row. For example, to combine rows 502i and 502j, the topic cluster module 310 combines cell 506 with cell 507, cell 508 with cell 509, and so on to get a combined row 502i-j. Similarly, the topic cluster module 310 combines cells in one column with the adjacent cells in the other to get a combined column. To combine two cells, the topic cluster module 310 combines the co-occurrence strengths $CS_{i,j}$ of the two cells into cluster co-occurrence strengths $CCS_{i-j,k}$. The cluster co-occurrence strengths $CCS_{i-j,k}$ indicate the measure of co-occurrence of the cluster (including topics t_i and t_j) with another topic t_k. The topic cluster module 310 combines the co-occurrence strengths $CS_{i,j}$ into cluster co-occurrence strength $CCS_{i-j,k}$ by adding multiplying, averaging, or applying another arithmetic or statistical function on the co-occurrence strengths $CS_{i,j}$. In one embodiment, the topic cluster module 310 also normalizes the cluster co-occurrence strength $CCS_{i-j,k}$ based on a factor like the frequency of the combined topics in the user profile collection.

[0064] The combination of the cells in the identified rows and columns leads to a new co-occurrence matrix (not shown) that includes n-1 topics wherein

one of these topics is a cluster c including the combined topics t_i and t_j. The topic cluster module 310 then repeats the step of identifying the cell in the new co-occurrence matrix with the highest co-occurrence strength $CS_{i,j}$ and clustering the topics or clusters associated with the identified cell. The identified cell may be associated with two topics, a topic and an already formed cluster, or two clusters. The topic cluster module 310 keeps repeating this process of clustering until a termination condition is reached. The termination condition can be a threshold number of clusters of a threshold size (in number of topics or number of videos), or the resulting cluster co-occurrence strength $CCSs_{i-j,k}$ for the updated cluster does not fall below a threshold. After the termination condition is reached, the topic cluster module 310 stores the clusters with their combined topics and cluster co-occurrence strengths $CCSs_{i-j,k}$.

[0065] Associating Users with Clusters

[0066] As stated above, the user profile module 306 stores topic clusters instead of, or in addition to, topics in the user profiles. To determine topic clusters c for a user's profile, the user profile module 306 identifies, for each topic in the user profile, the cluster c to which that topic belongs, adds that cluster c to a list of clusters for the user profile. The result is a user-cluster profile C comprising a plurality of clusters c.

[0067] The user profile module 306 then determines a user cluster strength UCS_c for each identified cluster c in the user-cluster profile C indicating the degree of association of the identified cluster c with the user. The user cluster strength UCS_c for an identified cluster c is the sum of the topic association strengths TAS_i for those topics of the user profile that are in that cluster c. In one embodiment, the user cluster strength UCS_c for an identified cluster c is the weighted sum of the topic association strengths TAS_i for those topics of the user profile that are in that cluster c. The weight for each of the topic association strengths is the cluster co-occurrence strength CCS_i for the identified cluster. In other embodiments, the user profile module 306 performs other mathematical or statistical functions on the topic association strengths TAS_i, and cluster co-occurrence strength CCS_i to achieve the user cluster strength UCS_c. The result of this operation is the representation of the user cluster profile C comprising a list (or vector) of cluster strengths.

[0068] Optionally, the user profile module 306 then selects a threshold number of clusters with highest user cluster strengths, e.g., the 20 clusters c with the highest UCS values. In another embodiment, the user profile module 306 selects all clusters c with the user cluster strength UCS beyond a threshold value. The user profile module 306 stores the selected clusters as clusters for the user-cluster profile.

[0069] Recommendations Based on Clusters

[0070] The cluster recommendation module 312 determines a recommendation of a video or a topic for a user, based on the user-cluster profile, and the interactions of other users with the videos.

[0071] To determine the recommendation of a video to a given user, the cluster recommendation module 312 selects one or more of the clusters identified in the user-cluster profile of the user. Then the module 312 analyzes the access data for the other users who also have the identified cluster(s) in their respective user-cluster profiles. These users are called a user-cluster community. In this fashion, a user is a member of a plurality of user-cluster communities, each community corresponding to one of the clusters in the user's user-cluster profile.

[0072] Given the user –cluster community, the module 312 determines which video(s) are currently popular with this set of users, for example in terms of user interactions, such as number of views, user ratings, user comments, or other measures of popularity. The module 312 then selects one of more of the most popular videos as recommended videos for the given user. For example, if a threshold number of these users have interacted with a particular video, the cluster recommendation module 312 selects the video as a recommendation. If the given user has already viewed a recommended video, it can be removed from the recommendations, or demoted in the recommendation list.

[0073] The module 312 can also make recommendations for topics, following a similar process. Here, the module 312 determines which topics are currently popular with the user-cluster community, rather than which specific videos are popular. Here, popularity of a topic is based on the aggregated interaction measure for the videos associated with the topic and viewed by users in the particular user-cluster community. The module 312 can then select one or more topics based on their aggregated interaction measures. For example, if a threshold number of

selected users have interacted with videos that have a particular topic, the cluster recommendation module 312 selects the particular topic.

[0074] From the selected topics, the module 312 then selects videos therein, based on factors such as popularity in the user-cluster community, recency, topic strength, and so forth. For example, if a threshold number of users in the user-cluster community have interacted with a particular video, the cluster recommendation module 312 recommends that video to the user.

[0075] In another embodiment, user communities are created by clustering the user-cluster profiles of a population of users. Here, the user-cluster profiles are form a sparse vector ("user cluster vector"), including for each cluster c the UCS_c as described above, and a zero value for each cluster that does not have one of its topics in the user profile. These user-cluster vectors can then be again clustered using k-means or other vector clustering methods, to form a number of user communities. Each user is then associated with the user communities containing the user's user-cluster vector. From there, recommendations may be made as above, using the user community as the population for identifying popular videos or topics.

[0076] The present invention has been described in particular detail with respect to a limited number of embodiments. Those of skill in the art will appreciate that the invention may additionally be practiced in other embodiments.

[0077] Within this written description, the particular naming of the components, capitalization of terms, the attributes, data structures, or any other programming or structural aspect is not mandatory or significant, and the mechanisms that implement the invention or its features may have different names, formats, or protocols. Further, the system may be implemented via a combination of hardware and software, as described, or entirely in hardware elements. Also, the particular division of functionality between the various system components described herein is merely exemplary, and not mandatory; functions performed by a single system component may instead be performed by multiple components, and functions performed by multiple components may instead be performed by a single component.

[0078] Some portions of the above description present the feature of the present invention in terms of algorithms and symbolic representations of operations on information. These algorithmic descriptions and representations are the means

used by those skilled in the art to most effectively convey the substance of their work to others skilled in the art. These operations, while described functionally or logically, are understood to be implemented by computer programs. Furthermore, it has also proven convenient at times, to refer to these arrangements of operations as modules or code devices, without loss of generality.

[0079] It should be borne in mind, however, that all of these and similar terms are to be associated with the appropriate physical quantities and are merely convenient labels applied to these quantities. Unless specifically stated otherwise as apparent from the present discussion, it is appreciated that throughout the description, discussions utilizing terms such as "selecting" or "computing" or "determining" or the like, refer to the action and processes of a computer system, or similar electronic computing device, that manipulates and transforms data represented as physical (electronic) quantities within the computer system memories or registers or other such information storage, transmission or display devices.

[0080] Certain aspects of the present invention include process steps and instructions described herein in the form of an algorithm. It should be noted that the process steps and instructions of the present invention could be embodied in software, firmware or hardware, and when embodied in software, could be downloaded to reside on and be operated from different platforms used by real time network operating systems.

[0081] The present invention also relates to an apparatus for performing the operations herein. This apparatus may be specially constructed for the required purposes, or it may comprise a general-purpose computer selectively activated or reconfigured by a computer program stored in the computer. Such a computer program may be stored in a computer readable storage medium, such as, but is not limited to, any type of disk including floppy disks, optical disks, DVDs, CD-ROMs, magnetic-optical disks, read-only memories (ROMs), random access memories (RAMs), EPROMs, EEPROMs, magnetic or optical cards, application specific integrated circuits (ASICs), or any type of media suitable for storing electronic instructions, and each coupled to a computer system bus. Furthermore, the computers referred to in the specification may include a single processor or may be architectures employing multiple processor designs for increased computing capability.

[0082] The algorithms and displays presented herein are not inherently related to any particular computer or other apparatus. Various general-purpose systems may also be used with programs in accordance with the teachings herein, or it may prove convenient to construct more specialized apparatus to perform the required method steps. The required structure for a variety of these systems will appear from the description above. In addition, the present invention is not described with reference to any particular programming language. It is appreciated that a variety of programming languages may be used to implement the teachings of the present invention as described herein, and any references to specific languages are provided for disclosure of enablement and best mode of the present invention.

[0083] Finally, it should be noted that the language used in the specification has been principally selected for readability and instructional purposes, and may not have been selected to delineate or circumscribe the inventive subject matter. Accordingly, the disclosure of the present invention is intended to be illustrative, but not limiting, of the scope of the invention.

WHAT IS CLAIMED IS:

1. A computer-implemented method for developing a profile of a user including topic clusters, the method comprising:

retrieving the profile including a plurality of topics, the plurality of topics indicating interests of the user;

determining a plurality of topic clusters including related topics;

identifying, from the plurality of topic clusters, a topic cluster including a topic indicating an interest of the user; and

storing the identified topic cluster in association with the profile of the user.

2. A computer-implemented method for developing a profile of a user, the method comprising:

determining digital content items interacted upon by the user, each of the digital content items associated with a plurality of topics;

determining the plurality of associated topics for each of the digital content items;

retrieving access data indicating the user's interactions with the digital content items;

selecting profile topics based on the topics associated with the digital content items and the retrieved access data;

storing the selected profile topics in association with the user's profile;

determining additional topics related to the stored profile topics, the additional topics co-occurring with profile topics in additional user profiles; and

storing the additional topics in association with the user's profile.

3. The method of claim 2, wherein the additional topics are determined based on a strength of co-occurrence of the additional topics with at least one of the profile topics.

4. The computer-implemented method of claim 2, wherein

the topics associated with each digital content item have topic strengths indicating the topics' degree of association with their content item, and

the profile topics are further selected based on the topic strengths.

5. The computer-implemented method of claim 2, wherein

each of the topics associated with each digital content item have a usefulness weight indicating how useful a topic is in representing its association with the digital content item, and

the profile topics are further selected based on the usefulness weights associated with the digital content items' topics.

6. The computer-implemented method of claim 5, wherein the usefulness weight of a topic is based on whether digital content items associated with the topic have objectionable content.

7. The computer-implemented method of claim 5, wherein the usefulness weight of a topic is based on a frequency with which the topic appears in a video corpus.

8. The computer-implemented method of claim 2, wherein

each of the user's interactions is associated with an interaction strength, and

the profile topics are further selected based on the interaction strengths associated with the user's interactions.

9. The computer-implemented method of claim 8, wherein the interaction strength of at least one user interaction is based on a frequency of the at least one user interaction.

10. The computer-implemented method of claim 8, wherein the interaction strength of at least one user interaction is based on a duration of the at least one user interaction.

11. The computer-implemented method of claim 8, wherein the interaction strength of at least one user interaction is reduced based on an amount of time elapsed since the at least one user interaction occurred.

12. The computer implemented method of claim 2, further comprising:

determining additional topics that co-occur in user profiles with the selected profile topics; and

storing the additional topics in association with the user's profile.

13. A computer system for developing a profile of a user, the system comprising a non-transitory computer readable medium storing instructions for:

determining digital content items interacted upon by the user, each of the digital content items associated with a plurality of topics;

determining the plurality of associated topics for each of the digital content items;

retrieving access data indicating the user's interactions with the digital content items;

selecting profile topics based on the topics associated with the digital content items and the retrieved access data;

storing the selected profile topics in association with the user's profile;

determining additional topics related to the stored profile topics, the additional topics co-occurring with profile topics in additional user profiles; and

storing the additional topics in association with the user's profile.

14. The computer system of claim 13, wherein the additional topics are determined based on a strength of co-occurrence of the additional topics with at least one of the profile topics.

15. The computer system of claim 13, wherein

the topics associated with each digital content item have topic strengths indicating the topics' degree of association with their content item, and

the profile topics are further selected based on the topic strengths.

16. The computer system of claim 13, wherein

each of the topics associated with each digital content item have a usefulness weight indicating how useful a topic is in representing its association with the digital content item, and

the profile topics are further selected based on the usefulness weights associated with the digital content items' topics.

17. The computer system of claim 16, wherein the usefulness weight of a topic is based on whether digital content items associated with the topic have objectionable content.

18. The computer system of claim 16, wherein the usefulness weight of a topic is based on a frequency with which the topic appears in a video corpus.

19. The computer system of claim 13, wherein

each of the user's interactions is associated with an interaction strength, and

the profile topics are further selected based on the interaction strengths associated with the user's interactions.

20. The computer system of claim 19, wherein the interaction strength of at least one user interaction is based on a frequency of the at least one user interaction.

21. The computer system of claim 19, wherein the interaction strength of at least one user interaction is based on a duration of the at least one user interaction.

22. The computer system of claim 19, wherein the interaction strength of at least one user interaction is reduced based on an amount of time elapsed since the at least one user interaction occurred.

23. The computer system of claim 13, further comprising instructions for:

determining additional topics that co-occur in user profiles with the selected profile topics; and

storing the additional topics in association with the user's profile.

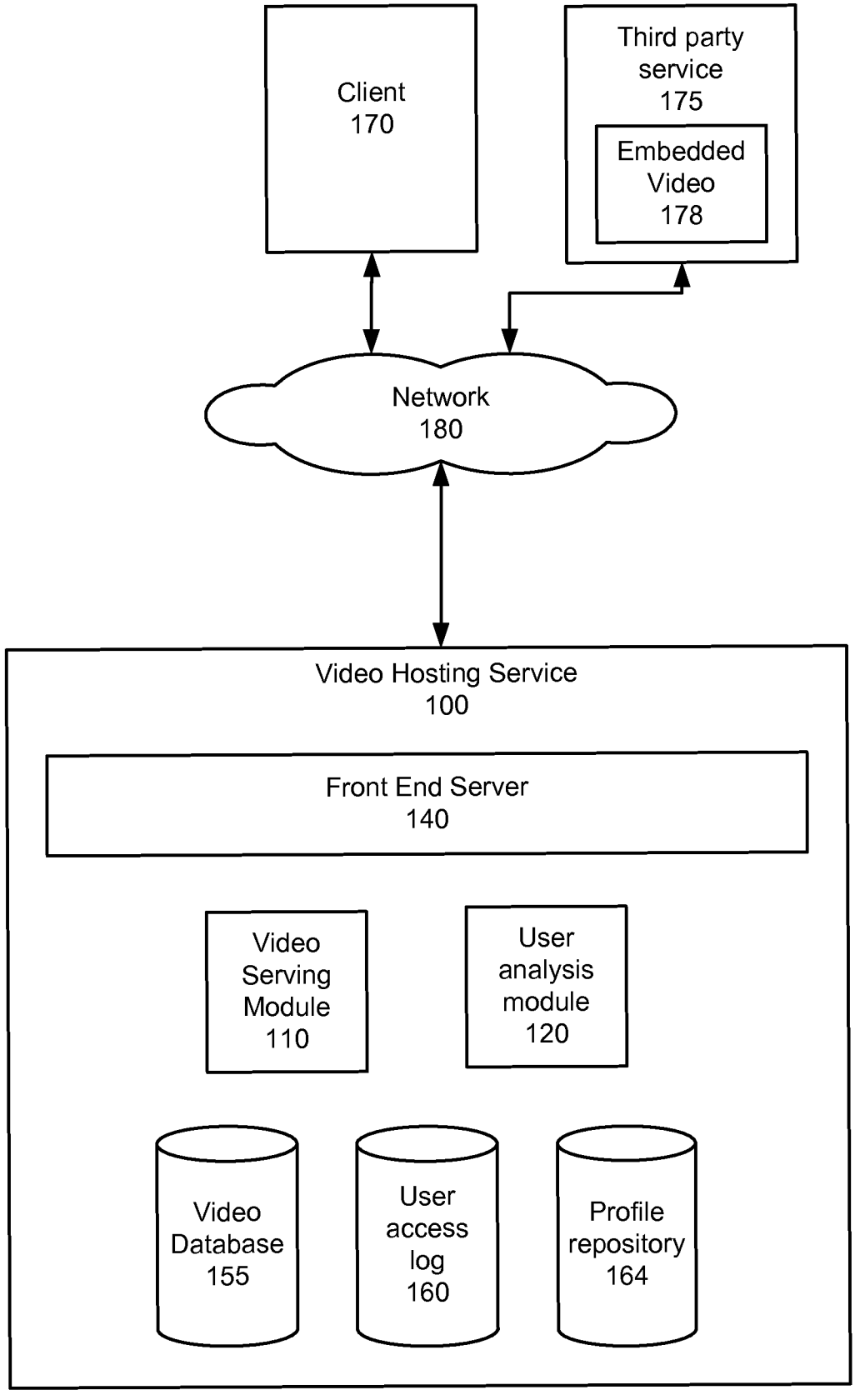

Client
170
Third party
service
175
Embedded
Video
178
Network
180
Video Hosting Service
100
Front End Server
140
Video
Serving
Module
110
User
analysis
module
120
Video
Database
155
User
access
log
160
Profile
repository
164

FIG. 1

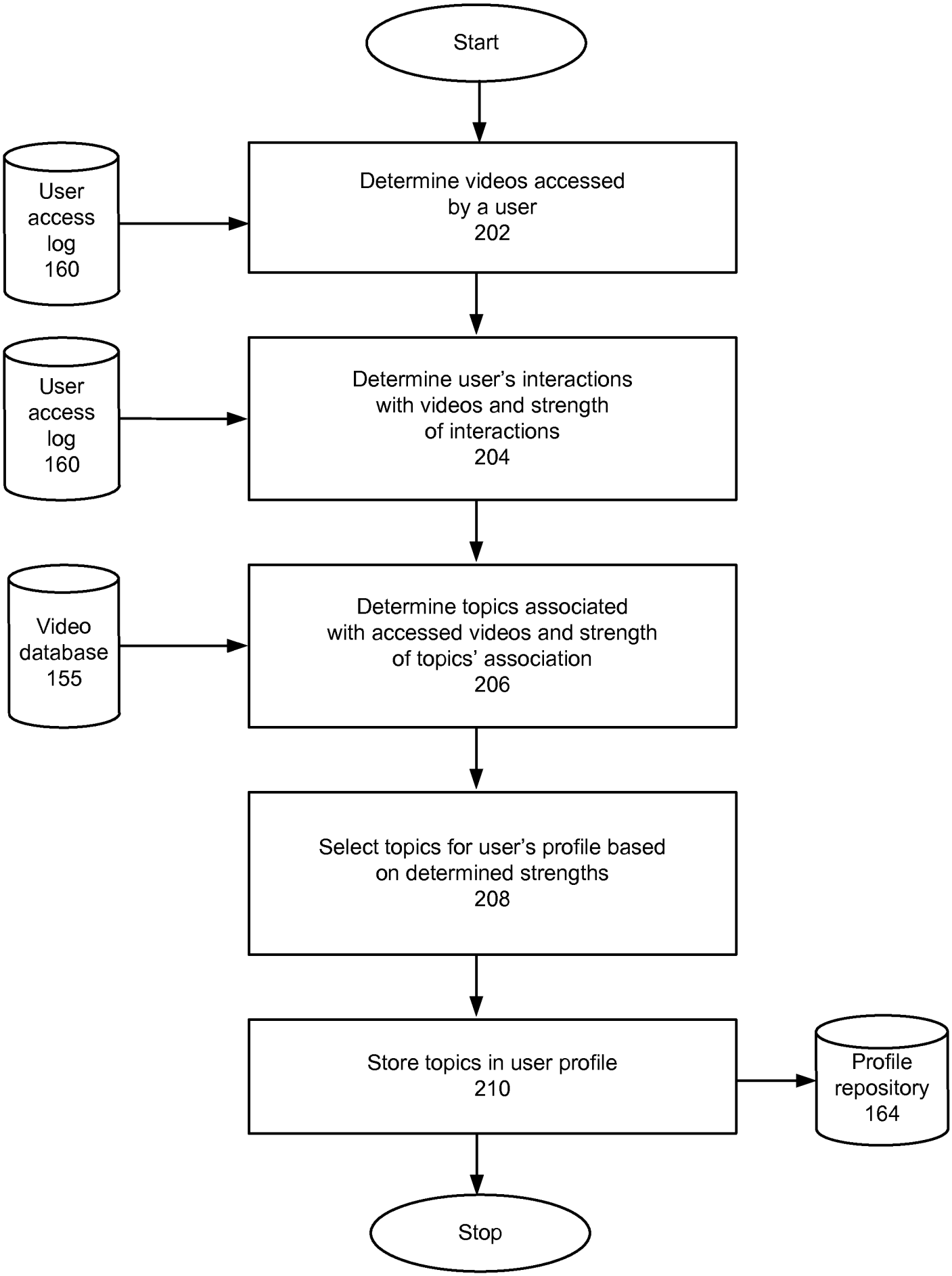

Start
User access log 160
Determine videos accessed by a user 202
User access log 160
Determine user's interactions with videos and strength of interactions 204
Video database 155
Determine topics associated with accessed videos and strength of topics' association 206
Select topics for user's profile based on determined strengths 208
Store topics in user profile 210
Profile repository 164
Stop

FIG. 2

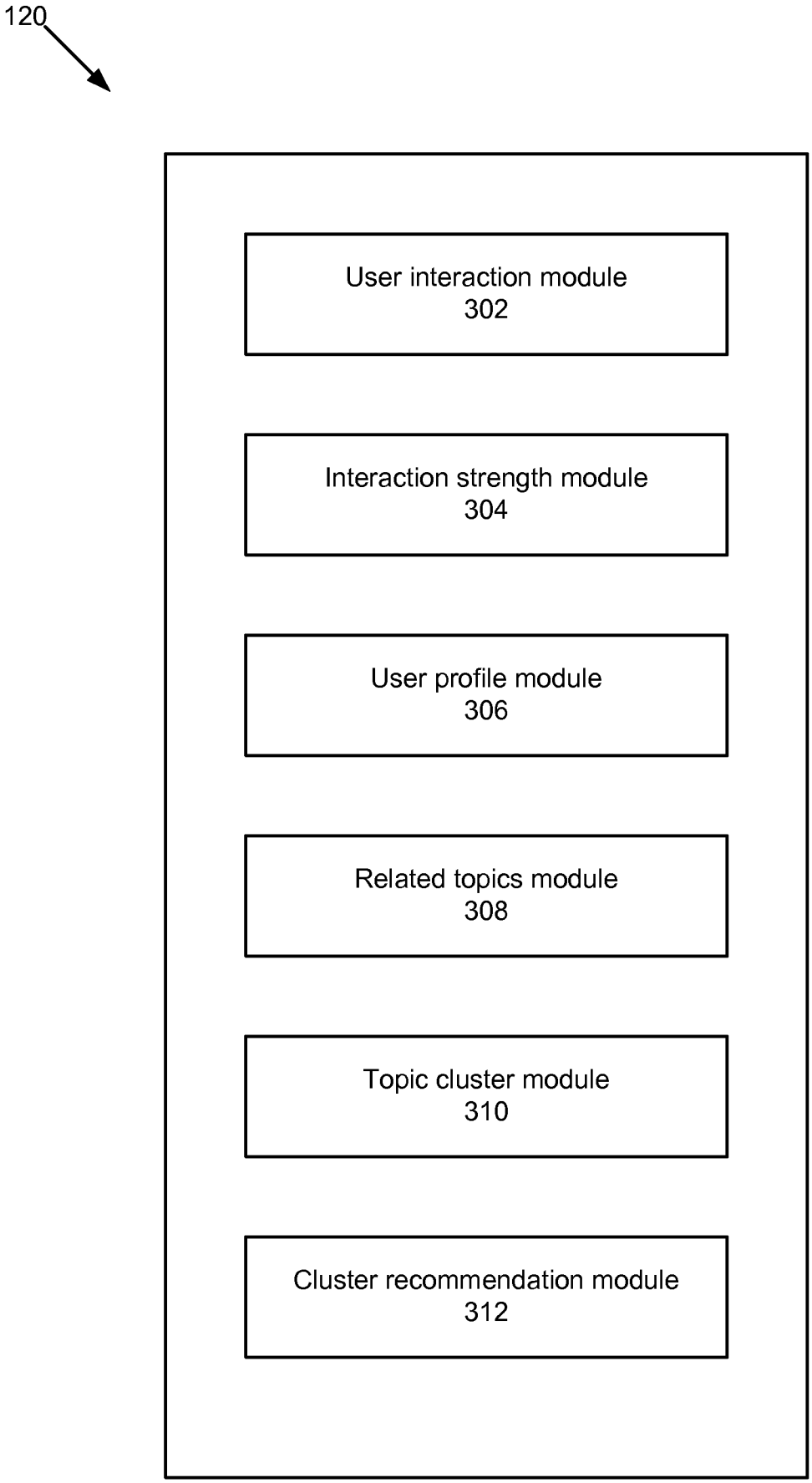

120
User interaction module
302
Interaction strength module
304
User profile module
306
Related topics module
308
Topic cluster module
310
Cluster recommendation module
312

FIG. 3

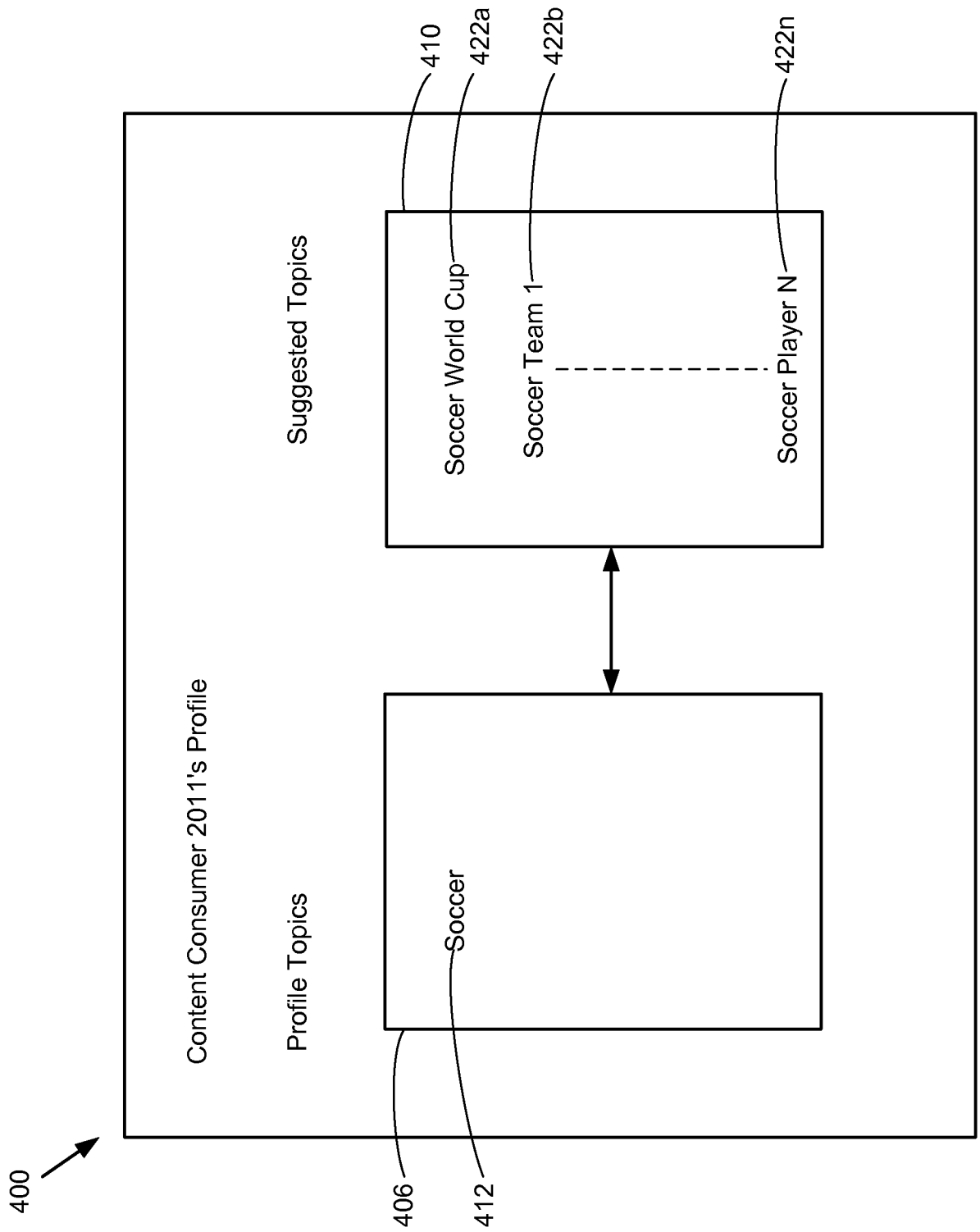

400
Content Consumer 2011's Profile
Profile Topics
Soccer
406
412
Suggested Topics
Soccer World Cup
Soccer Team 1
Soccer Player N
410
422a
422b
422n

FIG. 4

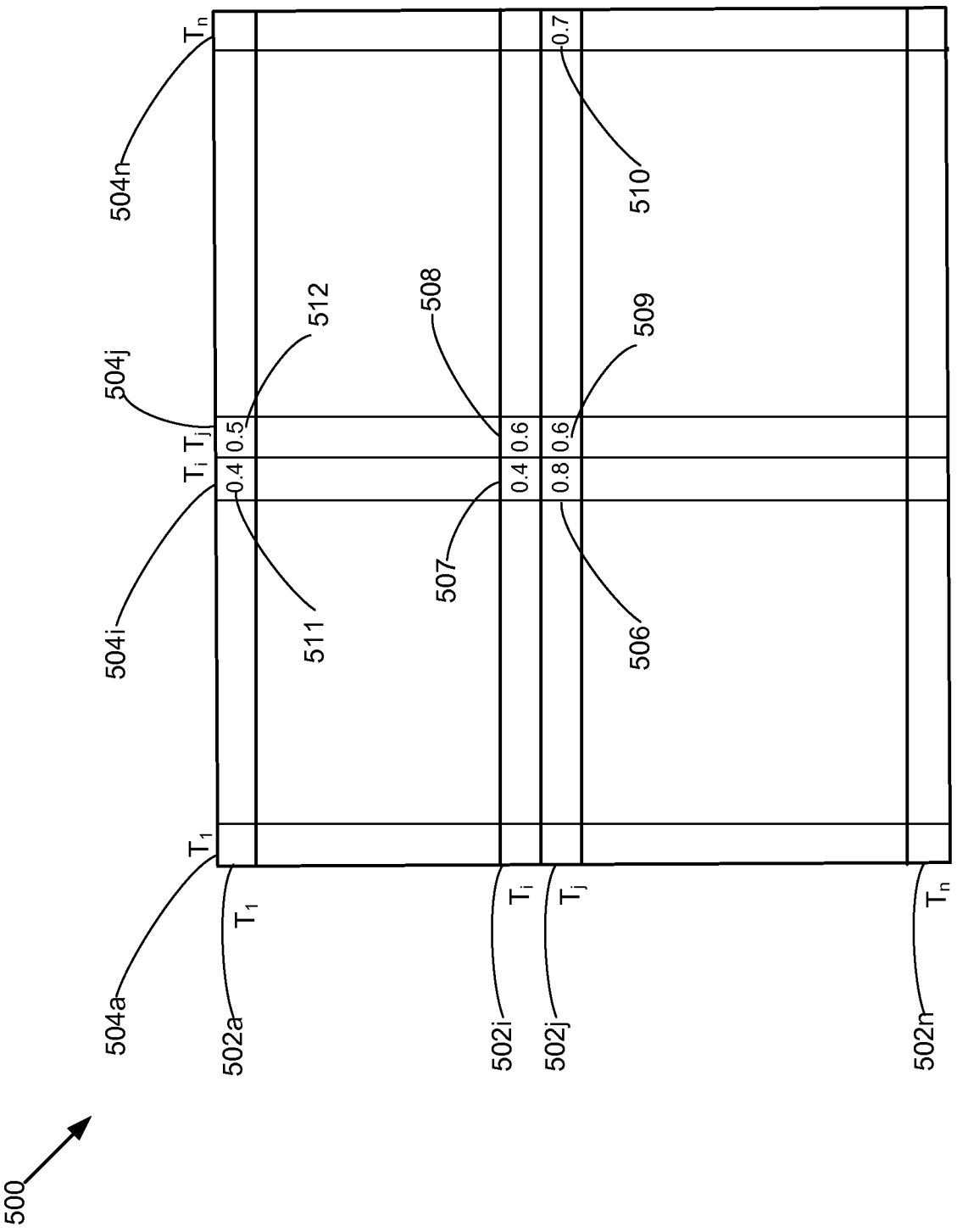

500
504a
502a
T_1
504i
T_i
504j
T_j
504n
T_n
T_1
T_i
T_j
T_n
502i
502j
502n
0.4
0.5
0.4
0.6
0.8
0.6
0.7
507
508
509
510
511
512
506

FIG. 5

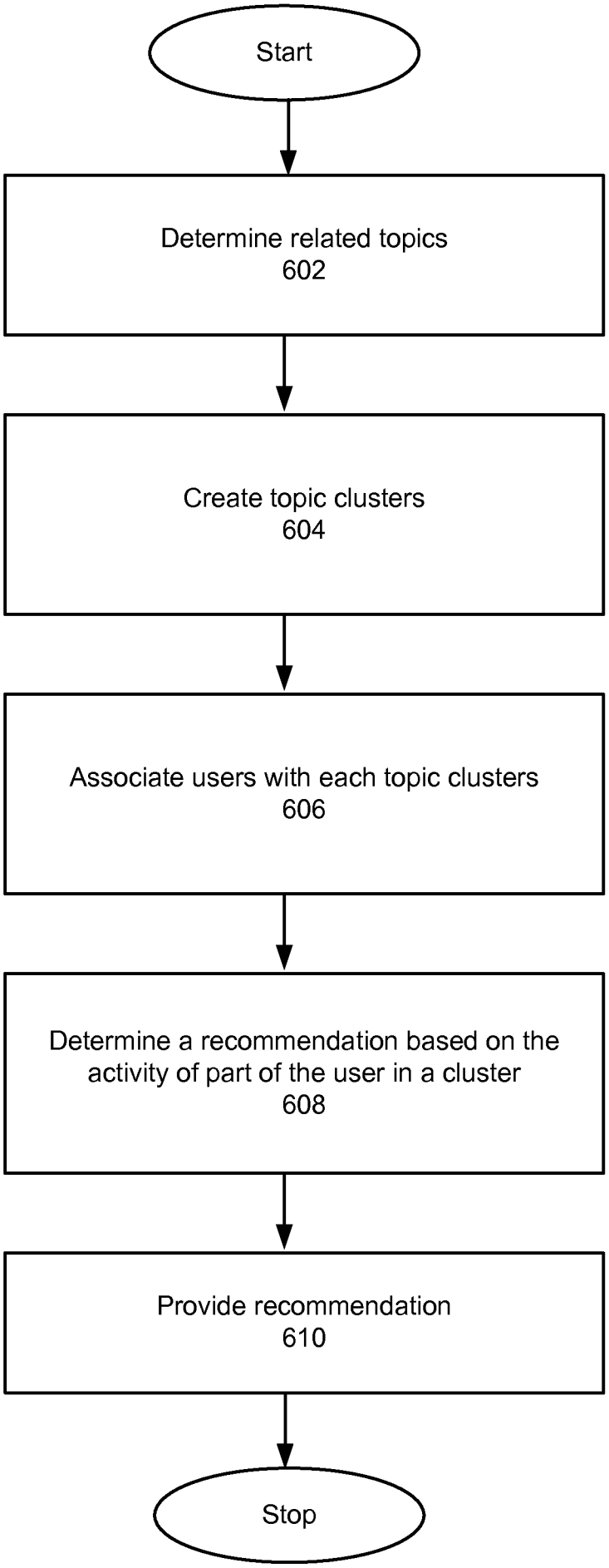

Start
Determine related topics
602
Create topic clusters
604
Associate users with each topic clusters
606
Determine a recommendation based on the activity of part of the user in a cluster
608
Provide recommendation
610
Stop

FIG. 6